Meet the Developers

Spectrum Scale SMB

Tips & Tricks for SMB

Ingo Meents & Ralph Wuerthner
2019-03-21, 11:00am
IBM Spectrum Scale Strategy Days 2019

IBM **Spectrum Scale**

IBM

Agenda

CES Overview

SYNC ACL

Contention / Hop Counts

Ingo Meents

Log File Troubleshooting

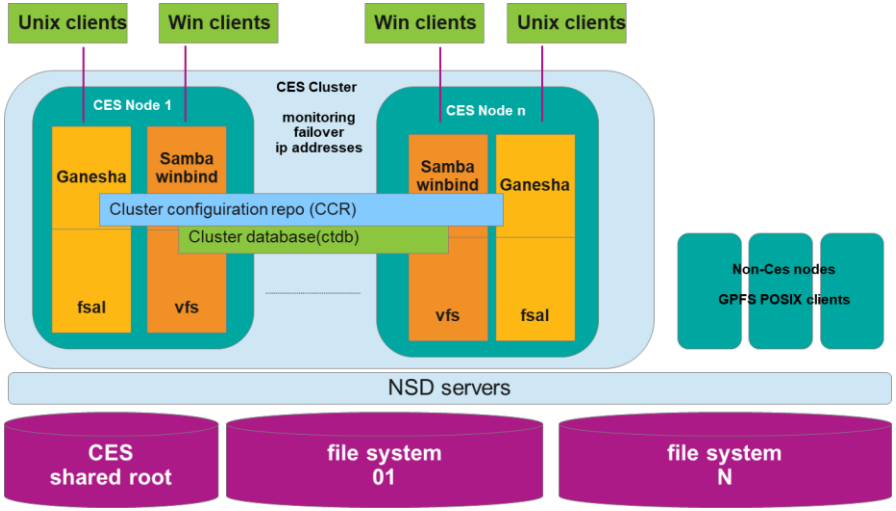
(CTDB / Samba)

Active Directory Troubleshooting

Further Reading on Developerworks

Ralph Wuerthner

Review of CES High Level Architecture




```
[root@node003 bin]# mmlscluster

GPFS cluster information
========================
  GPFS cluster name:         openstack-cluster.node001gpfs
  GPFS cluster id:           7079645339935612107
  GPFS UID domain:           openstack-cluster.node001gpfs
  Remote shell command:      /usr/bin/ssh
  Remote file copy command:  /usr/bin/scp
  Repository type:           CCR

 Node  Daemon node name  IP address  Admin node name  Designation
-----------------------------------------------------------------
   1   node001gpfs       172.31.0.3  node001gpfs      quorum-perfmon
   2   node002gpfs       172.31.0.4  node002gpfs      quorum-perfmon
   3   node003gpfs       172.31.0.5  node003gpfs      quorum-manager-perfm
   4   node004gpfs       172.31.0.6  node004gpfs      manager-perfmon
```

```
[root@node003 bin]# mmlscluster --ces

GPFS cluster information
========================
  GPFS cluster name:         openstack-cluster.node001gpfs
  GPFS cluster id:           7079645339935612107

Cluster Export Services global parameters
-----------------------------------------
  Shared root directory:                /ibm/gpfs0/ces
  Enabled Services:                     OBJ SMB NFS
  Log level:                            0
  Address distribution policy:          even-coverage

 Node  Daemon node name            IP address       CES IP address list
-----------------------------------------------------------------------
   3   node003gpfs                 172.31.0.5       192.168.1.13
   4   node004gpfs                 172.31.0.6       192.168.1.14
```

Spectrum Scale and Samba Release Overview

Spectrum Scale Release	General Availability	Samba Version	Platform Support (accu.)
4.1.1	2Q15	4.2	x86_64/RHEL7
4.2.0	4Q15	4.3	ppc64/RHEL7
4.2.1	2Q16	4.3	x86_64/SLES12
4.2.2	4Q16	4.4	ppc64le/RHEL7
4.2.3	2Q17	4.5	RHEL 7
5.0.0	4Q17	4.6	x86_64/Ubuntu
5.0.1	1Q18	4.6	No new platforms
5.0.2	3Q18	4.6	No new platforms
		Plan 4.9 (current stable Samba release)	No new platforms

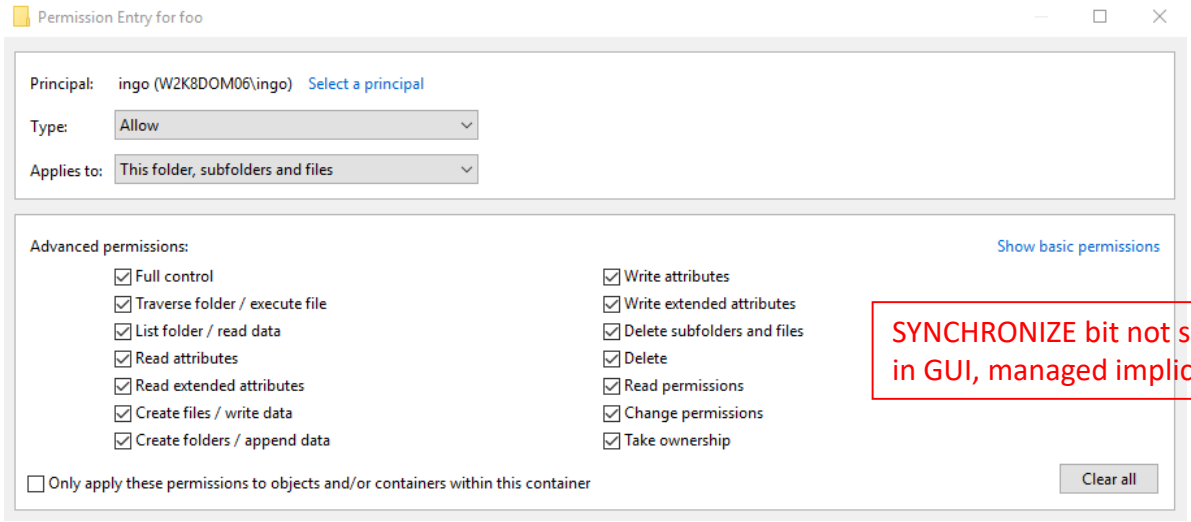
Current community work, next upcoming release 4.10

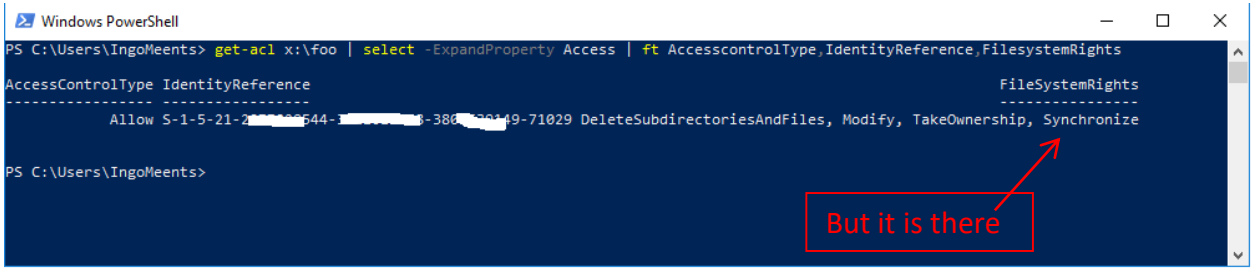
SYNCHRONIZE ACL Bit – The Windows View

- Microsoft definition
  "The right to use the object for synchronization. This enables a thread to wait until the object is in the signaled state. Some object types do not support this access right."

- **Important for file and directory access**

- Managed implicitly by Windows





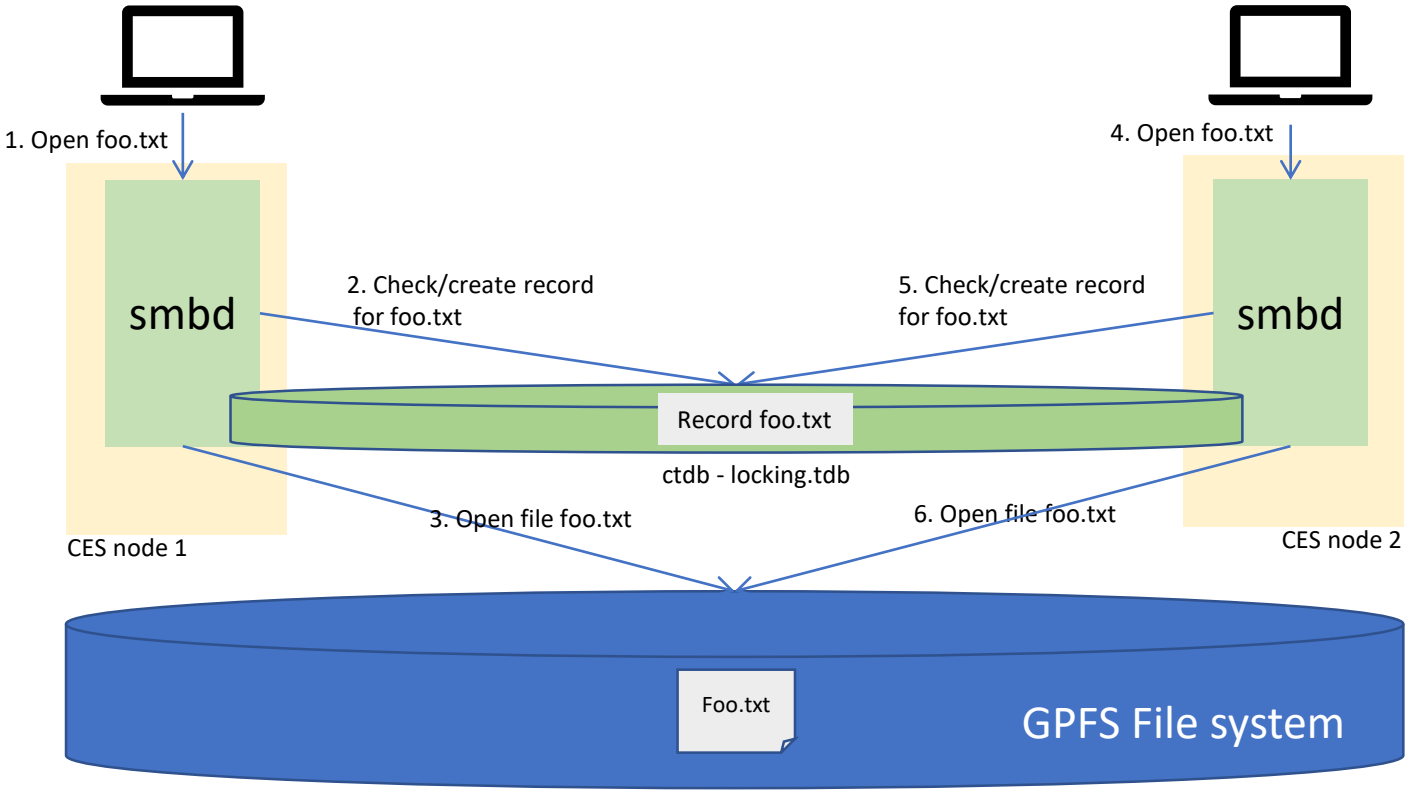
NFSv4 ACL bit SYNCHRONIZE – Scale / Samba

```
[root@fscc-x36m3-32 testingo]# mmgetacl foo
#NFSv4 ACL
#owner:W2K8DOM06\ingo
#group:W2K8DOM06\domain users
user:W2K8DOM06\ingo:rwxc:allow:FileInherit:DirInherit
 (X)READ/LIST (X)WRITE/CREATE (X)APPEND/MKDIR (X)SYNCHRONIZE (X)READ_ACL  (X)READ_ATTR  (X)READ_NAMED
 (X)DELETE    (X)DELETE_CHILD (X)CHOWN        (X)EXEC/SEARCH (X)WRITE_ACL (X)WRITE_ATTR (X)WRITE_NAMED
```

- Upstream Samba 4.6.9 fixed a bug/feature that had set this implicitly for SMB clients
  - https://bugzilla.samba.org/show_bug.cgi?id=7909
- We will finally include that fix with Samba 4.9 / 5.0.3
- It's normally set by Windows, but might be missed by NFS clients creating files/directories
- If this is causing issues and re-acl'ing is not possible the Samba option **nfs4:set synchronize** can be used to restore the old behaviour
  - yes SYNCHRONIZE is always set on ALLOW ACLS
  - No SYNCHRONIZE is passed through unmodified (default)

Files & Directory Contention Can Have Performance Impact

Simplified flow of concurrent file opens.



Scaling
Number of clients
Number of CES nodes

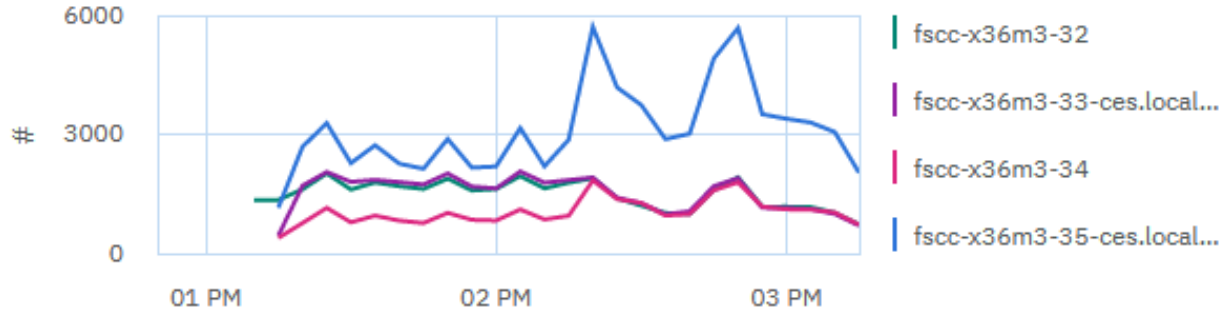
Use cases
SW patch distribution
Shared libraries
Shared binaries
Project directories
…

Analysis of this Situation

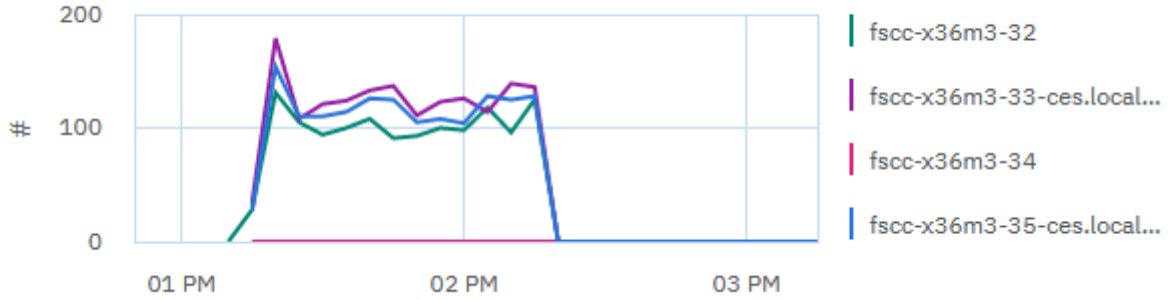
- CPU utilization of ctdb
  - ps, top, atop
  - Syslog future: Mar 7 15:55:20 fscc-x36m3-32 ctdbd[9561]: WARNING: CPU utilisation 97% >= threshold (90%)
- Hop count histogram can be looked at
  - Histogram buckets
    - Samba < 4.10: power of 4: 0, 4, 16, 64, 256,…
    - Samba >= 4.10: power of 2: 0, 2, 4, 8, 16, 32, … (changed for better granularity)
  - Messages in syslog (high hopcounts)
  - Ctdb statistics
  - CLI performance metrics: %> mmperfmon query ctdbHopCountDetails ← Predefined query
  - GUI thresholds
- Hot keys
  - Reported by ctdb statistics, syslog
  - Examined by net tdb locking – as long as file is open

GUI Thresholds: Hop Counts of Sample Workload

Locking tdb
Hop Count bucket 0
(hop counts == 0)



Locking tdb
Hop Count bucket 1
(hop counts < 4



Currently, GUI allows hop counts only with thresholds, if important, that could be changed.

CTDB Statistics

```
[root@fscc-x36m3-32 ~]# ctdb dbstatistics locking.tdb
DB Statistics locking.tdb
 db_ro_delegations              0
 db_ro_revokes                  0
 locks
    num_calls                 184
    num_current                 0
    num_pending                 0
    num_failed                  0
 hop_count_buckets: 1254702 273 0 0 0 0 0 0 0 0 0 0 0 0 0
 lock_buckets: 0 145 39 0 0 0 0 0 0 0 0 0 0 0 0
 locks_latency      MIN/AVG/MAX     0.003863/0.007882/0.036822 sec out of 184
 vacuum_latency     MIN/AVG/MAX     0.000461/0.001013/0.016799 sec out of 26539
 Num Hot Keys:     10
    Count:2 Key:ff5bd7cb3ee3822e023b011800000000000000000000000000
    Count:3 Key:ff5bd7cb3ee3822e8243001c00000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e8143001c00000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e015a000400000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e035a000400000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e8343001c00000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e001d011000000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e021d011400000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e8352000c00000000000000000000000000
    Count:2 Key:ff5bd7cb3ee3822e806f010800000000000000000000000000
```

Locking.tdb – all open files

Hop count buckets: 0, 4, 16, 64, …

Represent records, i.e. for locking.tdb open files, can be checked with net tdb locking <key> dump

Looking at File Record: Net dump locking

/usr/lpp/mmfs/bin/net tdb locking ff5bd7cb3ee3822e023b011800000000000000000000000000 dump

```
SHARE_MODE_DATA: struct share_mode_data
    sequence_number        : 0x399d763e5bb4765c (4151604441704068700)
    servicepath            : *
      servicepath          : '/ibm/gpfs0/fsroot1_ldx'
    base_name              : *
      base_name            : 'Brown/junior'
    stream_name            : NULL
    num_share_modes        : 0x00000014 (20)
    share_modes: ARRAY(20)
      share_modes: struct share_mode_entry
        pid: struct server_id
        …
        access_mask        : 0x001f01ff (2032127)
        share_access       : 0x00000007 (7)
        private_options    : 0x00000000 (0)
        time               : Thu Mar 14 01:23:53  CET 2019 CET.18635
        share_file_id      : 0x0000000018730853 (410191955)
        uid                : 0x00b90bde (12127198)
        flags              : 0x0000 (0)
        …..
```

Exported path

File / Folder

Number of concurrent opens

Node and smbd pid of opener

Agenda

CES Overview

SYNC ACL

Contention / Hop Counts

Ingo Meents

Log File Troubleshooting

(CTDB / Samba)

Active Directory Troubleshooting

Further Reading on Developerworks

Ralph Wuerthner

SMB Log Files

IBM Spectrum Scale

- Samba and Winbind log to syslog and files in /var/adm/ras
  - Samba errors and warnings show up in syslog and files in /var/adm/ras
  - Samba warnings show up in files in /var/adm/ras only
- Samba and Winbind log files in /var/adm/ras

```
log.smbd                  ← Samba main & client daemons (smbd)
log.wb-BUILTIN            ← Winbind domain child process handling BUILTIN domain
log.wb-CLUSTER3           ← Winbind domain child process handling CLUSTER3 domain
log.wb-VIRTUAL1           ← Winbind domain child process handling VIRTUAL1 domain
log.winbindd              ← Winbind main daemon (winbindd)
log.winbindd-dc-connect   ← Winbind domain controller connection manager
log.winbindd-idmap        ← Winbind component providing ID mapping
log.winbindd-locator      ← Winbind component locating domain controllers
```

- CTDB logs to syslog only

```
ctdbd[6120]: CTDB starting on node
ctdbd[6120]: Recovery lock not set
ctdbd[6121]: Starting CTDBD (Version 4.9.4.gpfs.15) as PID: 6121
...
```

Samba Log Messages

- ‘`WARNING: VFS call "xxxxx" took unexpectedly long`’
  - Duration of internal operations is monitored and compared with a 5 second threshold

- ‘`VFS call “create_file” took unexpectedly long`’
  - High level operation to open a file took too long
  - Overall operation requires open(), stat() and other system calls
  - Usually indicates contention on the locking record for this file or high load on CTDB

- ‘`VFS call “pread|pwrite|open|close|unlink|rename|readdir|opendir|closedir|stat” took unexpectedly long`’
  - A system call going directly into GPFS took long
  - Check GPFS, backend storage and network
  - Snapshot deletion can also negatively impact GPFS

CTDB related Log Messages (1)

IBM Spectrum Scale

```
db_ctdb_fetch_locked for /var/run/ctdb/CTDB_DBDIR/locking.tdb.1 key
B33680292738C54A9E386E0100000000000000000000000000, chain 8139 needed 1
attempts, 5203 milliseconds, chainlock: 0.003000 ms, CTDB 5203.146000 ms
```

- Fetching a data record from CTDB took long
  - `/var/run/ctdb/CTDB_DBDIR/locking.tdb.1`: database holding record
  - `key B336802927...`: identification of requested record
  - `chain 8139`: location (hash chain) of record in database
  - `needed 1 attempts`: number of requests to CTDB for getting the requested record
  - `5203 milliseconds`: total duration of fetch operation
  - `chainlock: 0.003000 ms`: total duration required to lock record in database
  - `CTDB 5203.14600 ms`: total duration of fetch operation within CTDB
- Typical cause:
  - record contention (watch for identical record keys)
  - high system load and/or high load on CTDB

CTDB related Log Messages (2)

```
Held tdb lock on db /var/run/ctdb/CTDB_DBDIR/locking.tdb.1, key
B33680292738C54A9E386E0100000000000000000000000000 7570.559000 milliseconds
```

- Samba held a record lock for too long
- Samba should immediately release records locks
- Typical cause:
  - Long running system calls or high system load

```
tdb_chainunlock on db /var/run/ctdb/CTDB_DBDIR/locking.tdb.1, key
B33680292738C54A9E386E0100000000000000000000000000 took 8.190000
milliseconds
```

- Unlocking a record took too long
- Kernel operation which should be completed immediately
- Typical cause:
  - Contention within kernel or system overload

CTDB related Log Messages (3)

IBM Spectrum Scale

Examples:

```
[2019/02/28 18:57:09.008990,  0] ../source3/modules/vfs_time_audit.c:46(smb_time_audit_log_msg)
  WARNING: VFS call "open" took unexpectedly long (7.92 seconds) filename = "/gpfs/share/file.dat" --
Validate that file and storage subsystems are operating normally
[2019/02/28 18:57:09.010365,  0] ../source3/modules/vfs_time_audit.c:46(smb_time_audit_log_msg)
  WARNING: VFS call "create_file" took unexpectedly long (7.92 seconds) cwd = "/gpfs/share", filename =
"file.dat" -- Validate that file and storage subsystems are operating normally


[2019/03/02 17:30:20.810469,  0] ../source3/lib/dbwrap/dbwrap_ctdb.c:1208(fetch_locked_internal)
  db_ctdb_fetch_locked for /var/run/ctdb/CTDB_DBDIR/locking.tdb.1 key
B33680292738C54AD02BD900000000000000000000000000, chain 42137 needed 1 attempts, 7649 milliseconds,
chainlock: 0.005000 ms, CTDB 7649.536000 ms
[2019/03/02 17:30:20.810594,  0] ../source3/modules/vfs_time_audit.c:46(smb_time_audit_log_msg)
  WARNING: VFS call "create_file" took unexpectedly long (7.65 seconds) cwd = "/gpfs/share", filename =
"notes.txt" -- Validate that file and storage subsystems are operating normally


[2019/02/28 18:56:42.008671,  0] ../source3/modules/vfs_time_audit.c:46(smb_time_audit_log_msg)
  WARNING: VFS call "rename" took unexpectedly long (62.68 seconds) cwd = "/gpfs/share", filename =
"data.dat" -- Validate that file and storage subsystems are operating normally
[2019/02/28 18:56:42.009032,  0] ../source3/lib/dbwrap/dbwrap_ctdb.c:1018(db_ctdb_record_destr)
  Held tdb lock on db /var/run/ctdb/CTDB_DBDIR/locking.tdb.1, key
B33680292738C54AFF4154010000000000000000000000000 62684.464000 milliseconds
```

Active Directory Troubleshooting (1)

IBM Spectrum Scale

- Verify Active Directory DNS setup

```
dig -t SRV _kerberos._tcp.$(/usr/lpp/mmfs/bin/net conf getparm global realm)
```

- Verify DNS server; check for expected Domain Controller entries

```
/usr/lpp/mmfs/bin/net ads info
```

- Basic information on LDAP, KDC, etc.
  - Connect to Domain Controller and retrieve machine account info
  - `/usr/lpp/mmfs/bin/net ads status -P`
  - Use existing or new connection to Domain Controller, issue request and expect response
  - `/usr/lpp/mmfs/bin/wbinfo --ping-dc`
  - Check all nodes - each node has its own connection!
  - `mmdsh -N CesNodes ...`

Active Directory Troubleshooting (2)

IBM Spectrum Scale

- Valid id mapping required for user and primary group for user to authenticate
- Goal should be having all id mapping valid
  - Test authentication
  - Verify id mapping for at least user id and primary group
  - `/usr/lpp/mmfs/bin/wbinfo --name-to-sid=<username>` and `--sid-to-name`
  - `/usr/lpp/mmfs/bin/wbinfo --sid-to-uid=<SID>` and `--uid-to-sid`
  - `/usr/lpp/mmfs/bin/wbinfo --sid-to-gid=<SID>` and `–gid-to-sid`
  - In case of using RFC2307/SFU id mapping (--unixmap-domains), attempt to query LDAP on Domain Controller
  - `/usr/lpp/mmfs/bin/net ads search -P sAMAccountName=<username> uidNumber primaryGroupID objectSID`
  - → primaryGroupID is RID, not GID
  - `/usr/lpp/mmfs/bin/net ads search -P objectSID=<SID> gidNumber`

Further Reading on Developerworks

IBM Spectrum Scale

- Best practices

  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/General%20Parallel%20File%20System%20(GPFS)/page/SMB%20Best%20Practices

- Authentication Planning

  https://developer.ibm.com/storage/2017/07/17/authentication-file-access-planning-smb-access/

- Protocol Node Tuning and Analysis

  https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/General%20Parallel%20File%20System%20(GPFS)/page/Protocol%20Node%20-%20Tuning%20and%20Analysis

- Authorization

  https://developer.ibm.com/storage/2016/07/06/ibm-spectrum-scale-security-blog-series-authorization/

Thank you. Time for questions.

Trademarks

The following terms are trademarks of International Business Machines Corporation in the United States and/or other countries:
alphaWorks, BladeCenter, Blue Gene, ClusterProven, developerWorks, e business(logo), e(logo)business, e(logo)server, IBM, IBM(logo), ibm.com, IBM Business Partner (logo), IntelliStation, MediaStreamer, Micro Channel, NUMA-Q, PartnerWorld, PowerPC, PowerPC(logo), pSeries, TotalStorage, xSeries; Advanced Micro-Partitioning, eServer, Micro-Partitioning, NUMACenter, On Demand Business logo, OpenPower, POWER, Power Architecture, Power Everywhere, Power Family, Power PC, PowerPC Architecture, POWER5, POWER5+, POWER6, POWER6+, Redbooks, System p, System p5, System Storage, VideoCharger, Virtualization Engine, GPFS.

A full list of U.S. trademarks owned by IBM may be found at: http://www.ibm.com/legal/copytrade.shtml.

Wireshark and the "fin" logo are registered trademarks of the Wireshark Foundation

UNIX is a registered trademark of The Open Group in the United States, other countries or both.
Linux is a trademark of Linus Torvalds in the United States, other countries or both.
Fedora is a trademark of Redhat, Inc.
Microsoft, Windows, Windows NT and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries or both.
Sun, the Sun logo, Sun Microsystems, Sun Microsystems Computer Corporation, SunSoft, the SunSoft logo, Solaris, SunOS, OpenWindows, DeskSet, ONC, ONC+, and NFS are trademarks or registered trademarks of Sun Microsystems, Inc. in the U.S. and certain other countries.
SLES is a registered trademark of SUSE LLC in the United States and other countries:

Other company, product and service names may be trademarks or service marks of others.