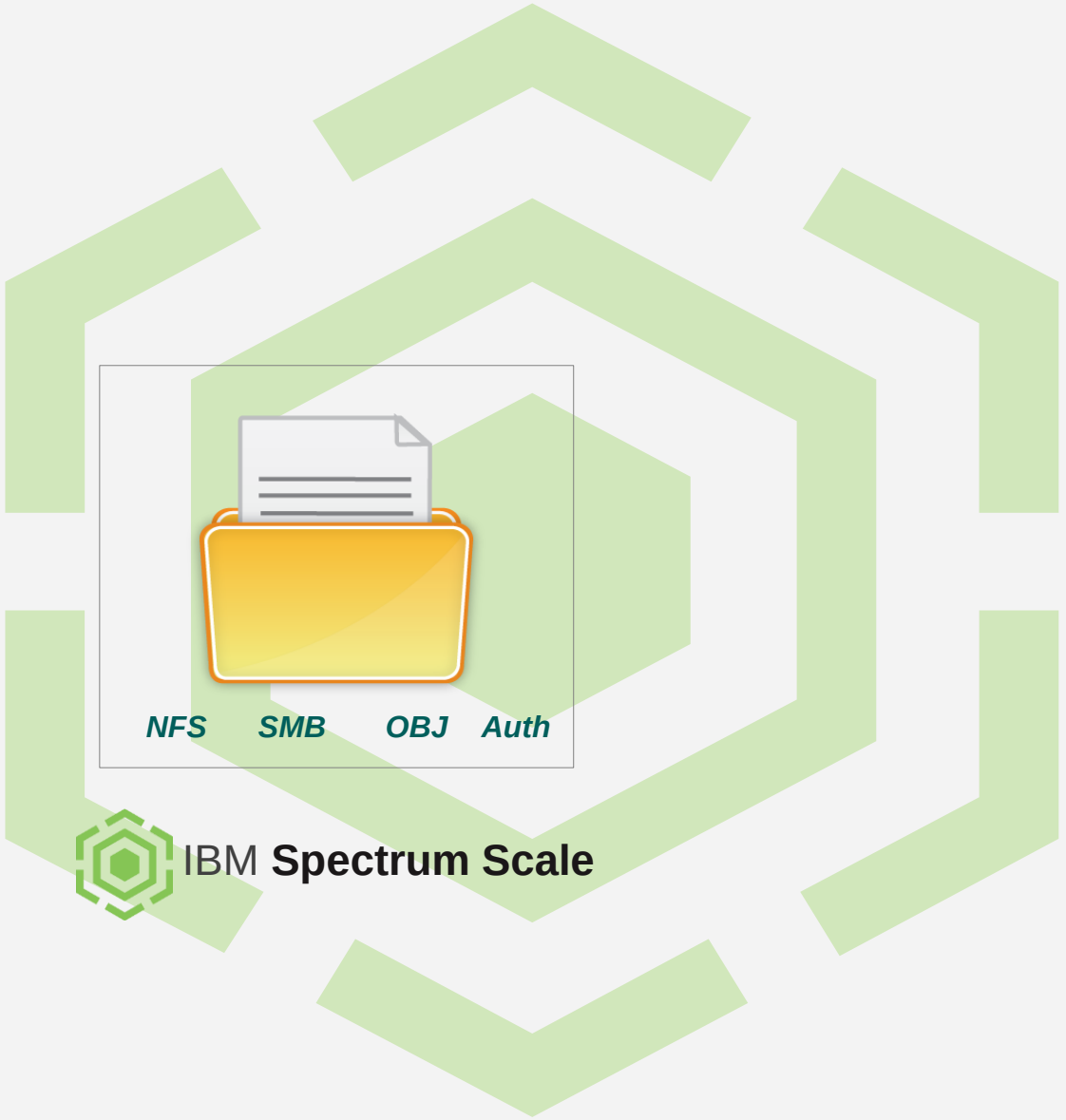
IBM Spectrum Scale File Protocols

What is new in Spectrum Scale Cluster Export Services (CES)?



User Group Meeting / Expert Days
March 4-5, 2020
Ehningen

Ingo Meents
Protocol Tribe Lead
IBM Germany Research & Development GmbH
Kelsterbach

Spectrum Scale File Protocols

- What has happened since last year?
- Current state
  - some recent customer performance numbers
  - Some thoughts on Power tuning / SMB best practices
- Discussion of some Authentication related RFEs
  - Sssd
  - Server side group lookups in AD environments
  - Group membership
- Protocols and containers

What's New – Besides Currency/Quality

Scale Version	NFS	SMB	OBJ
5.0.3	• **Ganesha 2.5.3-ibm036.00** • Auth / Performance / NFS config GUI • Increased send queue (performance) • More performance stats • Mem / readdir improvements • PTF-N: many fixes • PTF-3: migrated large internal NFS server to	• **Samba 4.9** • Auth / Performance GUI • SMB 3.1.1 • SYNCHRONIZE bit in NFSv4 ACLS is now honored • FSCTL_SET_ZERO_DATA is now supported on sparse files	• OS Swift Pike • mmobj with password file
5.0.4	• **Ganesha 2.7.5-ibm053.00** • mmnfs command includes an additional NFS configuration parameter RPC_IOQ_THRDMAX. • CES NFS grace period is changed from 60 seconds to 90 seconds • ganesha is upgraded to version 2.7 • trim subcommand is added to the ganesha_mgr tool to periodically call malloc_trim() to return the freed memory to the kernel • PTF-2 enhanced performance stats	• **Samba 4.9** • Mmuserauth service create **--enable-overlapping-unixmap-ranges** • Support for vfs_fruit module • Immutability over SMB: set immutable by setting read only, retention time = last access time, RO can be cleared and file deleted after expiry • Sync behaviour changed: syncops:onclose is disabled by default	• OS Swift Pike
Planned for next release	• **Ganesha 2.7.5-ibm054.06**	• **Samba 4.11** • Recovery lock	• OS Swift Pike

Power (8) Tunings Revisited

- On developer works, will be sunset end of March
- https://www.ibm.com/developerworks/community/wikis/home?lang=en#!/wiki/General%20Parallel%20File%20System%20(GPFS)/page/Spectrum%20Scale%20Tuning%20on%20Power
- Contents will be reviewed and updated into info center
- Firmware and O/S level recommendations, needs update, in general: use current Firmware
- LPAR hardware allocations for NUMA based POWER Servers, under review for post Power8 platforms
- Keep SMT at 2, again needs revalidation for post Power8 platforms
- GPFS settings
  - mmchconfig numaMemoryInterleave=yes (requires numactl)
  - mmchconfig workerThreads=512 (various recommendations, good starting point)
  - mmchconfig envVar="MLX4_USE_MUTEX=1 MLX5_SHUT_UP_BF=1 MLX5_USE_MUTEX=1“ (→ Flash)
  - mmchfs <fs_name> -L 32M , mmchfs <fs_name> -S relatime (have become default in later releases)

SMB Best Practices Revisited

- Cross-protocol options – can be switched off if you do not need them
  - gpfs:leases = yes/no
  - gpfs:sharemodes = yes/no
  - posix locking = yes/no
- Lock Coherency option: fileid:algorithm
  - fsname → filesystem name, cluster-wide (default)
  - fsname_norootdir → if share root does not get modified
  - fsname_nodirs → on cohereny on directories, but on files
  - fsname_hostname → per node
- Hide unreadable= yes/no
  - expensive, especially when ACLs are big and many files around

Try to keep concurrent access to files low and number of files in a directory at a reasonable level.

Authentication & ID Mapping Requests Seen From the Field

- sssd support
  - either already existing
  - Corporate policy
  - Red Hat recommendation
- Server side group lookup
  - NFS only cluster with AD auth → s4u2self
- Unix users - Enterprise-wide AD, departmental LDAP
  - Group membership not defined in AD

"We have to use sssd!"

"We need more groups for NFS users!"

We need our own Unix identities!

How can we break this down into things we can improve ? ... let's try

Authentication Matrix Revisited

Authentication method	ID Mapping method	SMB	SMB With Kerberos	NFSv3	NFSv3 With Kerberos	NFSv4	NFSv4 With Kerberos	Object
USER DEFINED	USER DEFINED	NA	NA	NA	NA	NA	NA	NA
LDAP with TLS	LDAP	✓	NA	✓	NA	✓	NA	✓
LDAP with Kerberos	LDAP	✓	✓	✓	✓	✓	✓	NA
LDAP with Kerberos and TLS	LDAP	✓	✓	✓	✓	✓	✓	NA
LDAP without TLS and without Kerberos	LDAP	✓	NA	✓	NA	✓	NA	✓
AD	Automatic	✓	✓	X	X	X	X	✓
AD	RFC2307	✓	✓	✓	✓	✓	✓	✓
AD	LDAP	✓	✓	✓	X	X	X	✓
NIS	NIS	NA	NA	✓	NA	✓	NA	NA
Local	None	NA	NA	NA	NA	NA	NA	✓

The Complication of ID Mapping

Unix / LDAP

Unix / LDAP /POSIX user
Name: ingo (or even imee)
Uid: 32 bit number for uid and gid
12071029

AD Domain W2K8DOM06

Windows user
Name: w2k8dom06\ingo
SID: theoretical size 454 bits, actual 128bit
S-1-5-21-2977929544-3112025818-3809539149-71029

Some trust relationship
SID filtering, SID history,
nested groups, group types,
Policies

AD Domain DSUB01

AD Domain DSUB02

Domain join

Scale CES cluster

There is no standard way but many Mapping assumptions
- We can throw away the domain assuming we only have one
- Win user name and Unix user name are the same
- Win group name and Unix group name are the same

This can work in special environments but is not the general case product needs to support.

```
[root@fscc-x36m3-32 ~]# id w2k8dom06\\ingo
uid=12071029(W2K8DOM06\ingo)
gid=12087532(W2K8DOM06\gr_ingo_1)
groups=12087532(W2K8DOM06\gr_ingo_1),
    12071029(W2K8DOM06\ingo),
    12087534(DSUB01\gr_ingo_3),
    12000513(W2K8DOM06\domain users),
    12087535(DSUB02\gr_ingo_4),
    11000545(BUILTIN\users)
```

Note that id only works after successful authentication of the user!

The General Mapping Questions

Windows user
Name: Domain\user
SID User
SID Primary group
SID Secondary group 1 (from Domain 1)
SID Secondary group 2 (from domain 2)
SID Secondary group 3 (from domain 3)

Unix / POSIX user
Name: unixuser
Uid
GID primary group
GID secondary group 1
GID secondary group 2
GID secondary group 3
GID secondary group 4

Scale User

- Name
  - Unix or windows user name ?
  - With domain prefix or without ?
  - What about clashes ? (domain1\john vs domain2\john vs john)
  - Currently:
    - windows name
    - (or winbind used default domain → no domain part)
    - Mapping MUST exists (otherwise access denied)
- Primary group
  - Windows or Unix? Can be configured in scale today
  - How to map?
  - Currently:
    - map by name, full domain required
    - Mapping MUST exists (otherwise access denied)
- Secondary groups
  - Windows or Unix? Or a union of both, or selected ones
  - Currently:
    - map by name, full domain required
    - Mapping MUST exists (otherwise group missing in Scale token)
- Home share ; login shell
  - Login shell does not matter for Samba
  - Home share does [homes]
    - Share for each user

Only at login can groups be reliably determined.

This currently prevents unix users that just authenticate against AD but get there groups elsewhere.

ID Mapping Done Today (for AD authentication)

Windows user
Name: Domain\user
SID User
SID Primary group
SID Secondary group 1 (from Domain 1)
SID Secondary group 2 (from domain 2)
SID Secondary group 3 (from domain 3)

Unix / POSIX user
Name: unixuser
Uid
GID primary group
GID secondary group 1
GID secondary group 2
GID secondary group 3
GID secondary group 4

Scale User	Idmap autorid	Idmap AD (SFU)	Idmap LDAP	Flexibility required
Name	Windows name	Windows name (exception use winbind default domain, not officially supported)	LDAP name = Windows name (with domain)	Requirement: Chose include domain or not
Primary group	Win primary group as there is no Unix group here, auto mapping	Win or Unix primary group, win group needs mapping	Win or Unix primary group, win group needs mapping	Selection possible
Secondary groups	All windows groups, mapping created internally	All windows groups, need mapping, ignored otherwise	Windows groups, need mapping, ignored otherwise	Requirement: allow groups provided thru nsswitch, ignore windows groups
Trusted domains	possible	Possible, SFU access directly to trusted domains, overlapping ranges possible	Different LDAP domain possible Tr doms do not matter here	Requires domain prefix

+ Centrify support via idmap_script as a frequent solution

Server Side Group Resolution for NFS Users

- This means scale has to query the group membership for users that have not necessarily gone through and SMB authentication.
- Why required ?
  - NFS v3 supports only 10 Groups
  - Ganesha option MANAGE_GIDS = true
- What is happening ?
  - Server queries nsswitch for groups
    - Winbind (or sssd) query AD server, heuristic
      - LDAP attribute memberOF (non recursive)
      - Computed attribute tokenGroup
      - Not supported because it can fail
      - Some customers are using it nevertheless
- Right way: Kerberos delegation → s4u2self

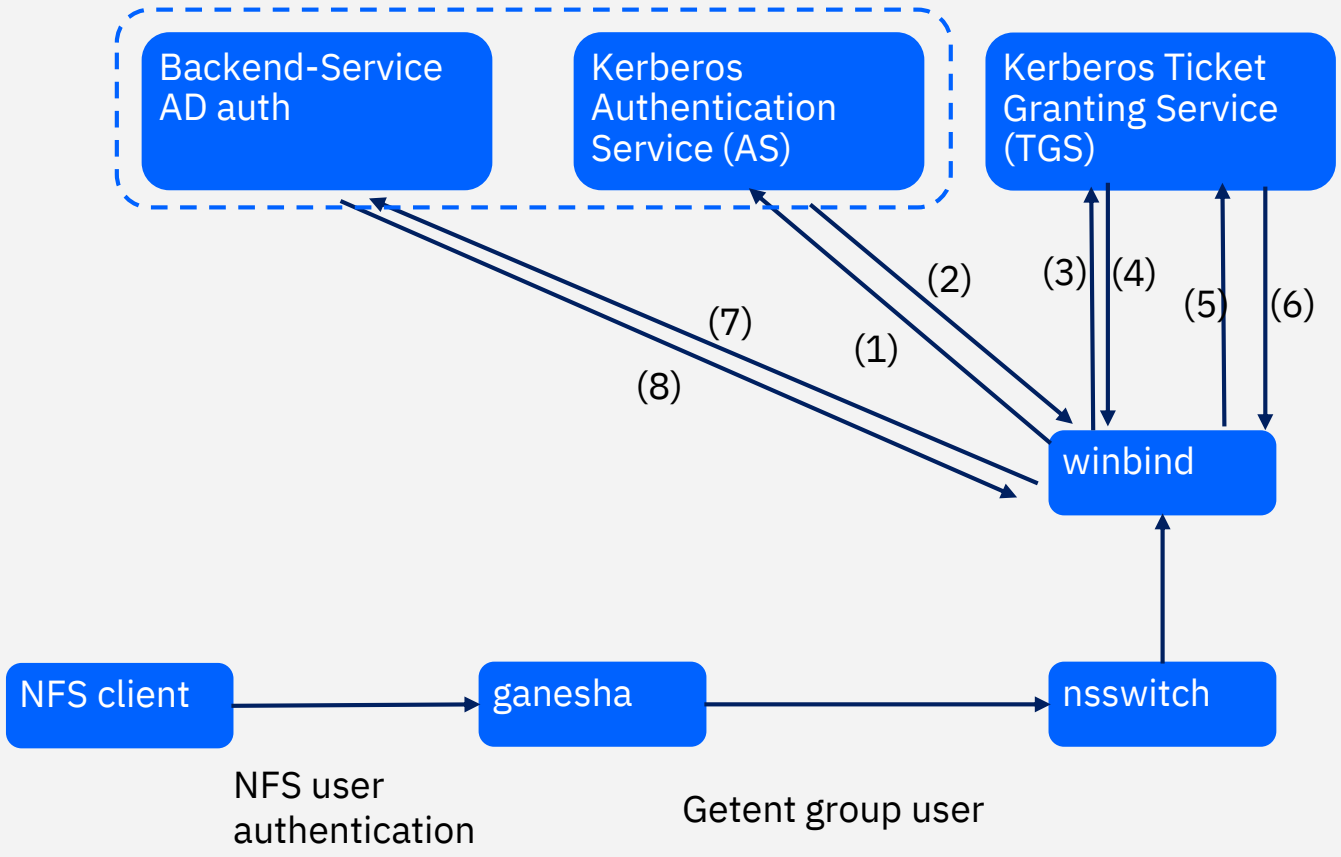
Remember: Only at login can groups be reliably determined.

tokenGroup

computed attribute that contains the list of SIDs due to a transitive group membership expansion operation on a given user or computer. Token Groups cannot be retrieved if no Global Catalog is present to retrieve the transitive reverse memberships.

https://docs.microsoft.com/en-us/windows/win32/adschema/a-tokengroups

Server Side Group Resolution – service4u2self



(1) AS-REQ: winbind authenticates
(2) AS-REP: winbind gets TGT

(3) TGS-REQ: service ticket for winbind but for NFS user's identify (field PA-S4U2self for client principal name)
(4) S4u2self ticket

(5) TGS-REQ: ticket request for backend on behalf of user

KDC checks
1. Is service trusted for constrained delegation
2. Is service trusted delegation for backend-service
3. Is client allowed to delegate ?

(6) TGS-REP: backend ticket for client identity
(7) Access to backend-service (AD auth)
(8) Response of backend service

→ Kerberos constrained delegation, protocol transition

sssd

Origin

- Started as part of FreeIPA, RedHat's attempt to simplify Unix/win integration wrt. services and identities
- Linux counterpart when AD started to deploy complex domains
- FreeIPA bundles DNS, LDAP, Kerberos, can serve as Linux DC
- Sssd was forked and is still used in FreeIPA
- Sssd was invented with enhanced offline capabilities in mind

Features

- Authentication (local, LDAP, Kerberos, AD, iPA, iDM, Proxy)
- Identity lookup / LDAP
- Name resolution and service discovery
- Policy management (sudo, hbac, seLinux, automount)

Selected feature gaps

- NTLM authentication (→ fallback, access by IP address)
- Clustering (→ organized password changes, multi-host auto idmapping, failover)
- MS RPCS (winbind is required for some operations)
- Trusted domains (each domain needs to be accessible by LDAP)

The Scale View: Winbind vs sssd - Where are the Conflicts ?

Usernames	• fully qualified or just username ?
nsswitch	• who is first for passwd and groups ?
System keytab	• System Keytab – who can (over-)write it ? • SMB + kerberized NFS → secrets tdb + system keytab (with NFS principals)
Computer account	• Winbind: single machine account (netbiosname), clustered password changes • Sssd: not clustered, each node separately or same account ?
Auto idmapping	• Sssd uses different algorithm than winbinds autorid • Sssd's auto mapping is nondeterministic (order of ranges) – bad in cluster without further assumptions
Group Resolution	• would we get the same results ? • MemberOf, TokenGroups fallback used with different/similar fallback logic, different accounts, different caching

Customer Use Cases Analysed

- AD users **login locally** on CES SMB node
  - Default scale config does not mess with PAM→ you can do that (conflict with system keytab file?)
- Sssd can manage **sudo** users
  - Scale / winbind do not care → you can do that (conflict with system keytab file?)
- **Automount** rules
  - Should not apply on a Scale SMB server node
- Server side **group** lookup
  - Currently sssd and winbind both have tokenGroups fallback (which can cause troubles)
- Clients are using sssd **idmapping** already
  - We do not have a solution here yet
  - idmap_sss – part of sssd, not winbind, requires configured sssd (clustering, etc.)
  - idmap_nss with sssd in nsswitch – single domain, winbind tries to create a windows-mapped token

Conclusion

- If you need clustered SMB with the highest possible Windows compatibility (NTLM, cross-forest trusts) use winbind.
- We need to make sure that we have proper interfaces to allow for sssd without interferences
- Possible approaches (all with challenges to resolve conflicts and issues)
  - Get group memberships from sssd
  - Get idmappings from sssd
  - Native sssd integration – no winbind → need to live with limitations

- … so there is lot's of work to do
- … looking for use case discussion

RHEL 76 documentation

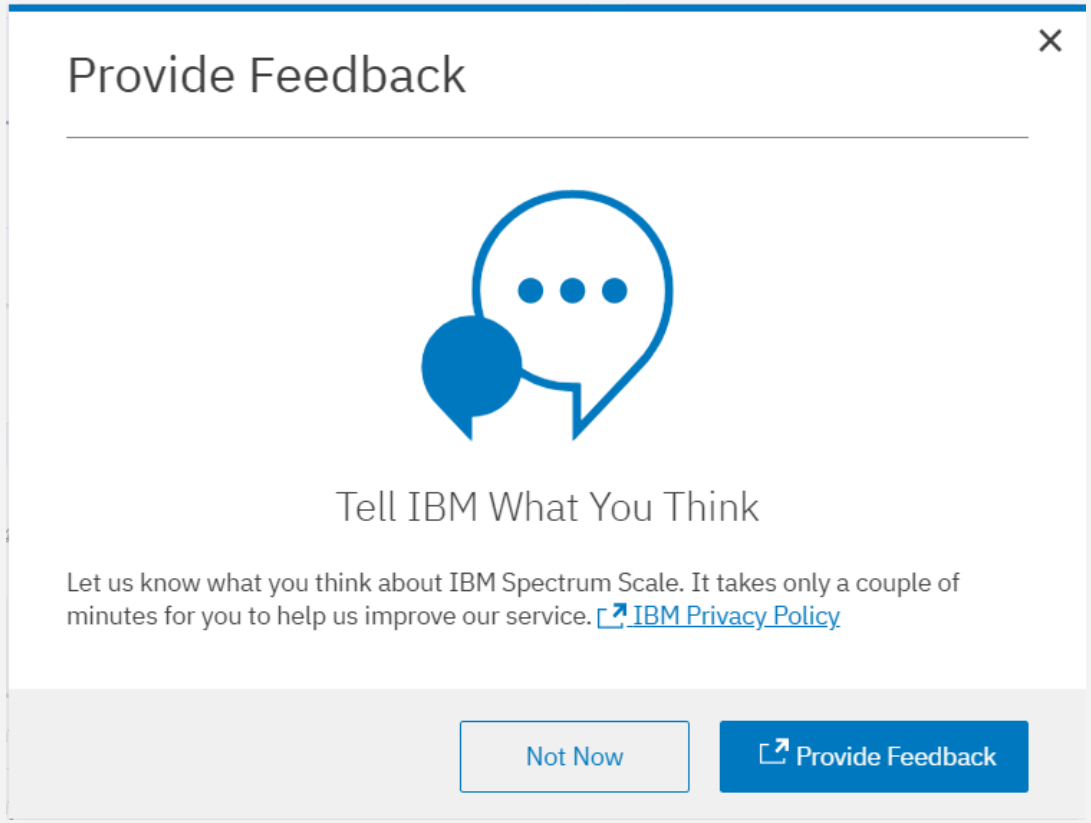
https://access.redhat.com/documentation/en-us/red_hat_enterprise_linux/7/html/windows_integration_guide/smb-sssd

Important

Using SSSD as a client in IdM or Active Directory domains has certain limitations, and Red Hat does not recommend using SSSD as ID mapping plug-in for Winbind. For further details, see the "What is the support status for Samba file server running on IdM clients or directly enrolled AD clients where SSSD is used as the client daemon" article.

SSSD does not support all the services that Winbind provides. For example, SSSD does not support authentication using the NT LAN Manager (NTLM) or NetBIOS name lookup. If you need these services, use Winbind. Note that in Identity Management domains, Kerberos authentication and DNS name lookup are available for the same purposes.

Thank you !



Please help us to improve Spectrum Scale with your feedback

- If you get a survey in email or a popup from the GUI, please respond
- We read every single reply

Trademarks