redhat.

SAMBA AND SMB3: ARE WE THERE YET?

Ira Cooper
Principal Software Engineer – Red Hat
Samba Team

September 22, 2015

NO FLASH PHOTOGRAPHY
SLIDES ARE ONLINE

GLOSSARY

What does that acronym mean?

- SMB – Server Message Block
- RDMA – Remote Direct Memory Access
- TDB – Trivial DataBase
- WIP – Work In Progress
- CTDB – Clustered TDB (Does much more)
  - Full cluster manager
  - VIP (Virtual IP) management etc
- VFS – Virtual File System
  - Pluggable and Stackable Filesystem Layer
  - Can have NO mount on the server – ex: vfs_glusterfs
  - Can just implement features – ex: vfs_shadow_copy2
  - Two types of modules: stackable and non-stackable

WHAT IS SAMBA?

No not the dance

- For Linux, *BSD, *nix, Solaris, Illumos, and maybe, VMS:
  - An implementation of :
    - NT Lan Manager v1.0 to SMB 3.1.1, for file serving
    - Active Directory Server and Client
    - DCE/RPC Services
    - WINS/Name Services
- Often considered the most “compatible” SMB server
  - Except for Microsoft
    - Maybe?
- GPLv3, and openly developed
  - Free for all to use

ARE WE?

Let us find out.

- SMB 3.1 Support
- Leases
- Multichannel
- SMB Direct
- Persistent Handles
- Witness Protocol
- Workload Support
- Future Topics

SMB 3.1

Not totally done, but...

- Negotiate Contexts
- Pre-authentication

- Encrypted by default
  - Matching MS
  - Encryption performance issues
    - Improvements still needed here
      - Use of hardware primitives
      - Use of standard libraries

LEASES

Client Side Caching Support

- Leases replace the old “oplock” model

- Come with 3 basic flags:
  - READ: Allows caching for reading (R)
  - WRITE: Allows caching for writing (W)
  - HANDLE: Allows caching of the handle for future use (H)
- Conceptually simpler to understand than the old oplocks

- In SMB3 directories can have RH leases
  - Better metadata caching
  - No more “5 second rule”

LEASES

SMB 2.1 Leases are go

- Standard SMB 2.1 Lease Support
- Interoperates with oplocks
- CTDB support
- VFS layer support is a WIP (Work In Progress)
  - Allow passing of leases to the underlying filesystem
- No directory leases
- No support for dynamic shares (like home directories)
  - Windows client limitation

MULTICHANNEL

Multiple Data Paths, One Connection

- Multichannel allows better performance in many situations
  - The client can decide how many TCP connections to open
  - The client can decide how many NICs to use
  - The client can decide how all of the above gets mapped out

- In addition, it enables SMB Direct support to be developed (RDMA)
  - Which is also, multichannel

MULTICHANNEL

Not quite there yet...

- Working with Samba's mutiprocess architecture
  - Channels for the same session will be passed to the same process
  - Async I/O across all channels handled by threads
- Client GUID and Session ID tracked to enable multichannel
  - Still clarifying with MS for the exact details
- Network interface list code is incomplete at this time
  - Will be non-portable

- Please goto Michael Adam's talk on multichannel for more details
  - Wednesday, 1:00PM – San Tomas / Lawerence

SMB DIRECT

SMB3 RDMA Support

- Multichannel is a pre-requisite to doing SMB Direct
- SMB Direct is a thin wrapper around SMB3
  - Just enough to let us send SMB3 over RDMA
  - Multiple RDMA controllers can be used
    - Just like multiple nics in normal multichannel
- SMB Direct support on the client side is quite fast
  - MS has some really amazing benchmarks here

MULTICHANNEL

Not quite there yet...

- Working with Samba's mutiprocess architecture
  - Channels for the same session will be passed to the same process
  - Async I/O across all channels handled by threads
- Client GUID and Session ID tracked to enable multichannel
  - Still clarifying with MS for the exact details
- Network interface list code is incomplete at this time
  - Will be non-portable

- Please goto Michael Adam's talk on multichannel for more details
  - Wednesday, 1:00PM – San Tomas / Lawerence

SMB DIRECT

How will we get there?

- Two different approaches proposed:
  - Usermode, single RDMA daemon
  - Kernel mode
- Both have real problems

SMB DIRECT

User Space Only

- First likely implementation
- metze has already began prototyping this approach

- Advantages:
  - Should be fairly portable
  - All in user space, much simpler to debug and fix
  - No kernel community “politics”

- Disadvantages:
  - More context switching.
  - Less potential for deep offloading of ops, etc.

SMB DIRECT

A Kernel of Truth?

- Kernel driver, assisting userspace.

- Advantages:
  - Less context switches
  - Some ops may be able to be short circuited totally in kernel
    - SMB2_READ
    - SMB2_WRITE
  - Review by the kernel community may help find issues
- Disadvantages:
  - Non portable
  - Harder to debug
  - Parts of Samba in the kernel?

PERSISTENT HANDLES

Allowing Client Recovery From Disconnects

- Attempts to get this “right”:
  - Durable Handles
  - Resilient Handles
  - Now, Persistent Handles

- Persistent Handles allow for hard guarantees across client disconnects
- Persistent Handles are really a full “system” feature
  - Software is but one part

- Major part of Hyper-V and MSSQL support
  - If you choose not to lie

PERSISTENT HANDLES

How?

- Two approaches:
  - Use CTDB
  - Push information into the filesystem

PERSISTENT HANDLES

CTDB Based

- To be done:
  - Develop fast enough persistence in CTDB
- Advantages:
  - Cross filesystem consistency
  - Samba controlled semantics
  - Databases are managed by CTDB
- Disadvantages:
  - Cross protocol compatibility will be difficult
  - Added compexity in Samba
  - May not take full advantages of what filesystems provide

PERSISTENT HANDLES

Filesystem Based

- To be done:
  - VFS Improvements
- Advantages:
  - Mutliprotocol consistency
    - Semantic enforcement by the filesystem
  - Databases are managed by the filesystem
- Disadvantages:
  - Added complexity in the filesystem
  - Not filesystem independent
  - Samba semantics can become filesystem dependent
  - We will still be using CTDB for other things – can't get rid of it

WITNESS PROTOCOL

Failure and Client Connection Management

- Advises clients as to the state of the share
  - Like when it became available again after failure
  - What other servers are serving it

- Allows admins to control client movement
  - Emptying a node before upgrade
  - Emptying a node before decommission
  - Load balancing, insertion of a new node
  - Load balancing, avoid hot spots

WITNESS PROTOCOL

I've witnessed its progress.

- Wireshark dissector – Complete
- Torture tests – Complete
- CTDB integration – On going
- CLI for management – WIP
- Demo – Go see their presentation

- Expected to land for Samba 4.4 or 4.5
- Please goto Jose and Guenther's talk on witness for more details
  - Wednesday, 3:05PM - San Tomas / Lawerence

WORKLOAD SUPPORT

HYPER-V only, so far

- Improved support for hole punching
- Improved support for hole finding
  - This call used to be faked, it is now real
- vfs_hyperv can be used for initial support

- Highlights the critical need for persistent handles, multichannel and SMB Direct

- Nothing done with MSSQL so far

OTHER TOPICS

Interesting stuff in Samba!

- POSIX Extensions for SMB2+
  - Allows for real expansion of SMB3 into Linux workloads
- New improved messaging infrastructure
  - Much more scalable and easier to work with
  - Please goto Volker Lendeke's talk for more details
    - Wednesday, 2:00PM – San Tomas / Lawerence
- Samba interfacing with cloud storage
  - Work in progress.
  - Please goto Jeremy Allison's talk for more details
    - Wednesday, 5:05PM – San Tomas / Lawerence

DADDY, ARE WE THERE YET

Not QUITE yet.

- SMB 3.1 Support – Done
- Leases – Initial implementation in place
- Multichannel – In progress
- SMB Direct – Being researched
- Persistent Handles – In progress
- Witness Protocol – In progress
- Workload Support – Being researched
- Other Topics – Plenty of exciting stuff going on

MY THOUGHTS

Progress

- People are dedicated to working on SMB3 now
- Progress is being made
- Code is being committed
- Most importantly, we are going in the right direction


QUESTIONS?

redhat.

THANK YOU

plus.google.com/+RedHat

facebook.com/redhatinc

linkedin.com/company/red-hat

twitter.com/RedHatNews

youtube.com/user/RedHatVideos