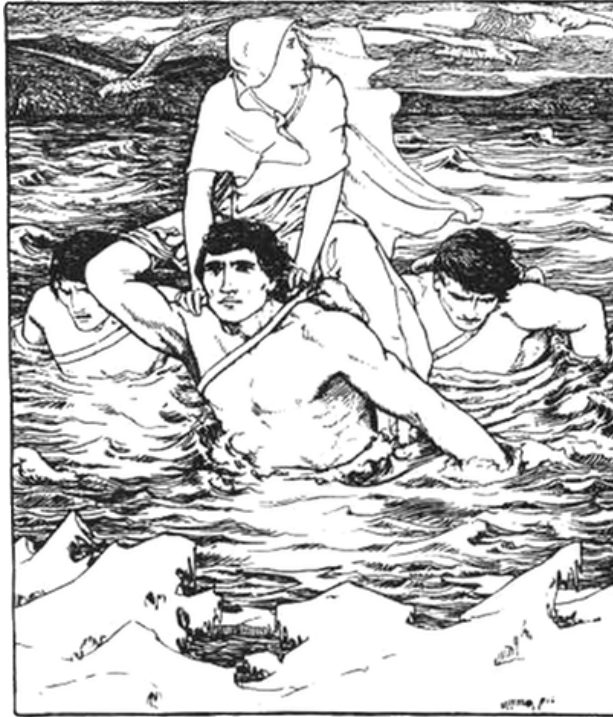


This page intentionally
includes a pointless message

SAMBA TEAM

Samba on Gluster



The Things We Do For Love

Christopher R. Hertel
Samba Team

Jose A. Rivera
Just Enjoy the Ride



Introductions





Introductions

Me

- ▶ Author
- ▶ Storage Architect
- ▶ Network Engineer
- ▶ SMB Know-It-All
- ▶ Samba Team Member
(since '98)
- ▶ Incurable Idealist



A ruminant mammal (Geekus geekus) with long legs, humped shoulders, and broadly palmated antlers.



Introductions

The Other Guy



Swimming in the deep end
of the SMB cesspool since
2008.



Introductions

Now
With



redhat®

The opinions expressed are my own
and not necessarily those of my
employer, my spouse, my childrenz,
the dog, or "the Voices".





Introductions

You





Introductions

Where are we going?

- 🌍 “Metadata” and “Semantics”
- 🌍 Semantic Translation
- 🌍 Samba/VFS and CTDB
- 🌍 Gluster Basics
- 🌍 Gluster ♥ Samba
- 🌍 Questions
(and answers, if I have any)







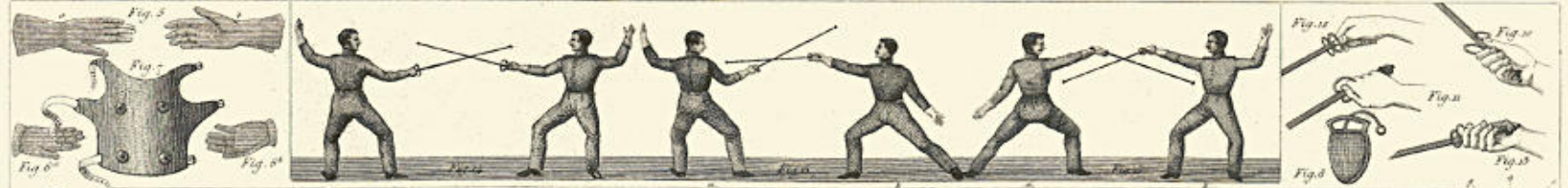
What's in a Name?

A Rose by Any Other Name
Would Wither and Die

-- Alan Swann (Peter O'Toole),
My Favorite Year



What's in a Name?



Metadata: Data about Data

Things like:

- 👤 Path Names and inode numbers
- 👤 Timestamps
- 👤 Permissions/Access Controls
- 👤 File size and Quota

Metadata identifies files, provides current state, sets limits, etc.



What's in a Name?

Semantics: The Meaning

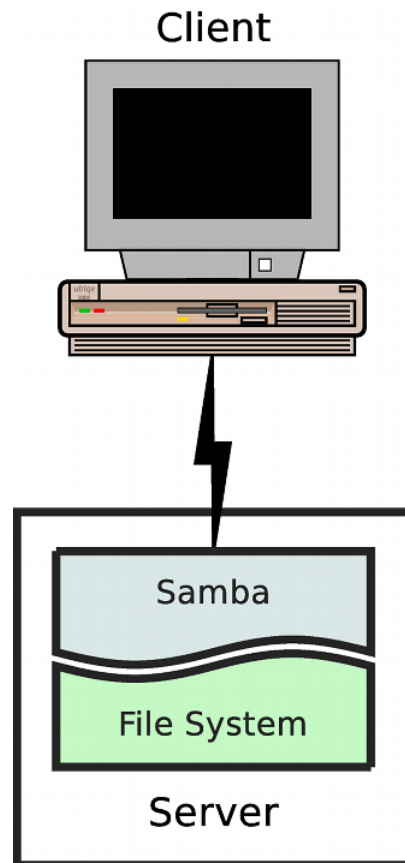
- Context and rules for interpreting metadata
- Enforcement of set limits
- Reasonable assurance of correctness

Semantic rules, such as access controls and quota limits, must be enforced by the **Operating System and File System**.



What's in a Name?

Samba is a Semantic Translator



There and back again...



- The client expects Windows semantics from the server.
- Samba expects POSIX semantics from the file system.

Samba must translate from Windows to POSIX and back again.



What's in a Name?

The Pieces Must Fit

-  If Samba does not properly handle SMB protocol, we call it a bug and expect trouble.
-  If the file system does not properly handle the POSIX calls made by Samba, we also expect trouble.



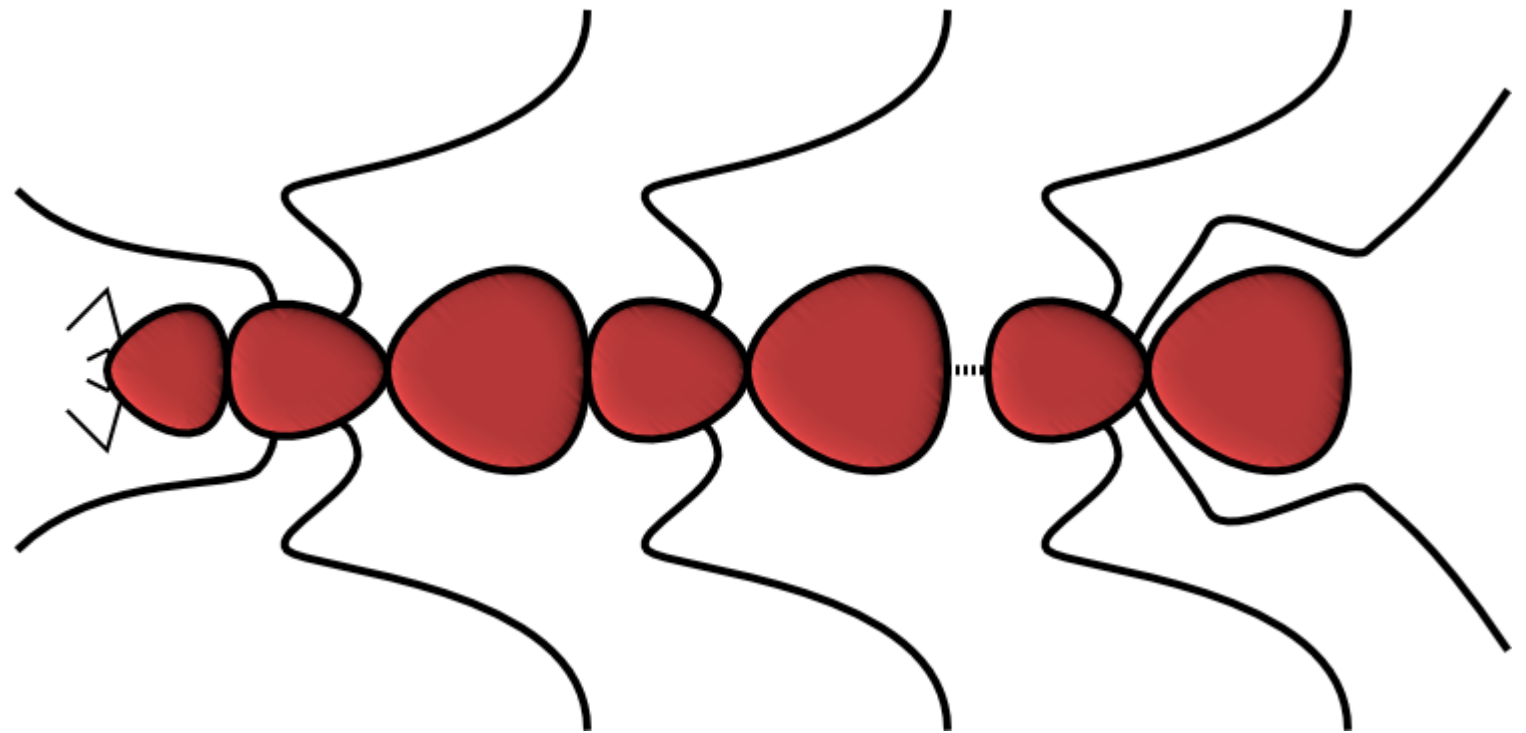
What's in a Name?



If you're looking for dragons,
you'll find them in the semantics.



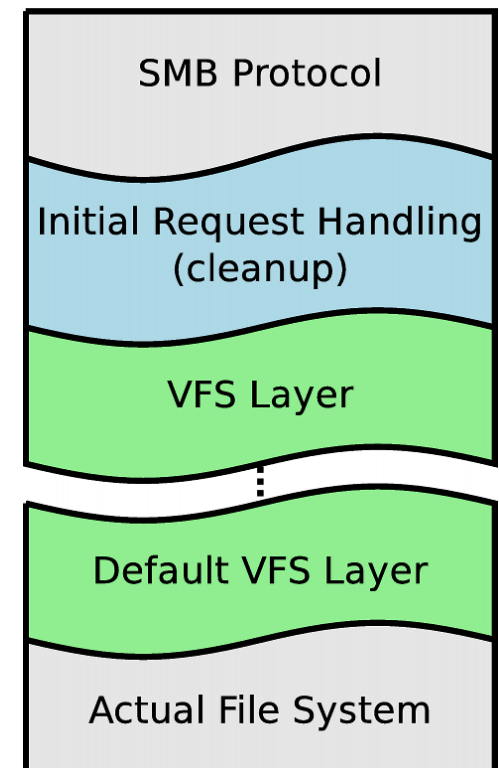
The Samba VFS layer



The Samba VFS Layer

Samba is Built in Layers (conceptually)

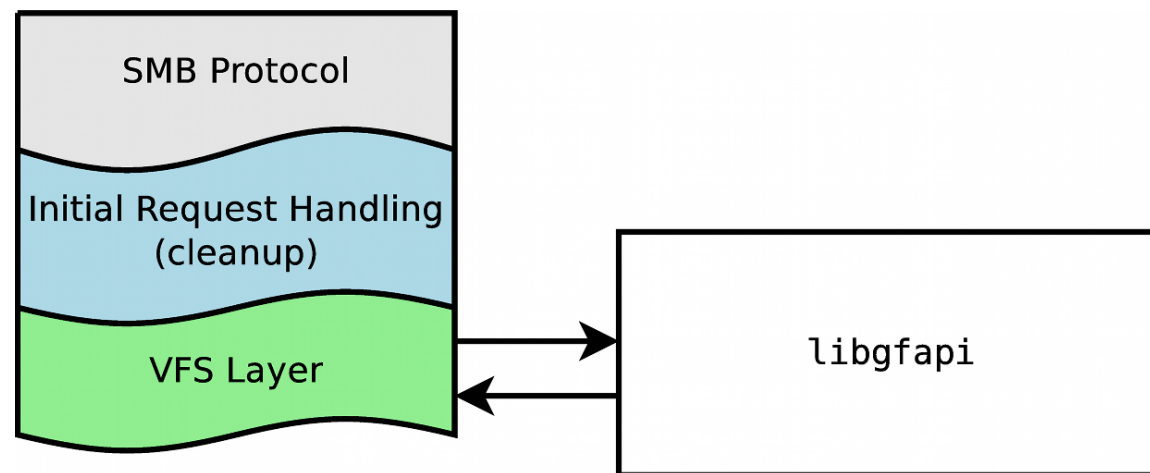
- ④ SMB messages are received and parsed
- ④ (Non-FS commands are handled elsewhere)
- ④ Call the VFS layer
- ④ The final VFS module talks to the File System
 - ☀ Higher VFS layers may bypass lower layers





The Samba VFS Layer

If there's no real File System, we can bypass the lower VFS layers



All VFS calls must be implemented
(possibly returning `ENOTSUP`)
to avoid errors.



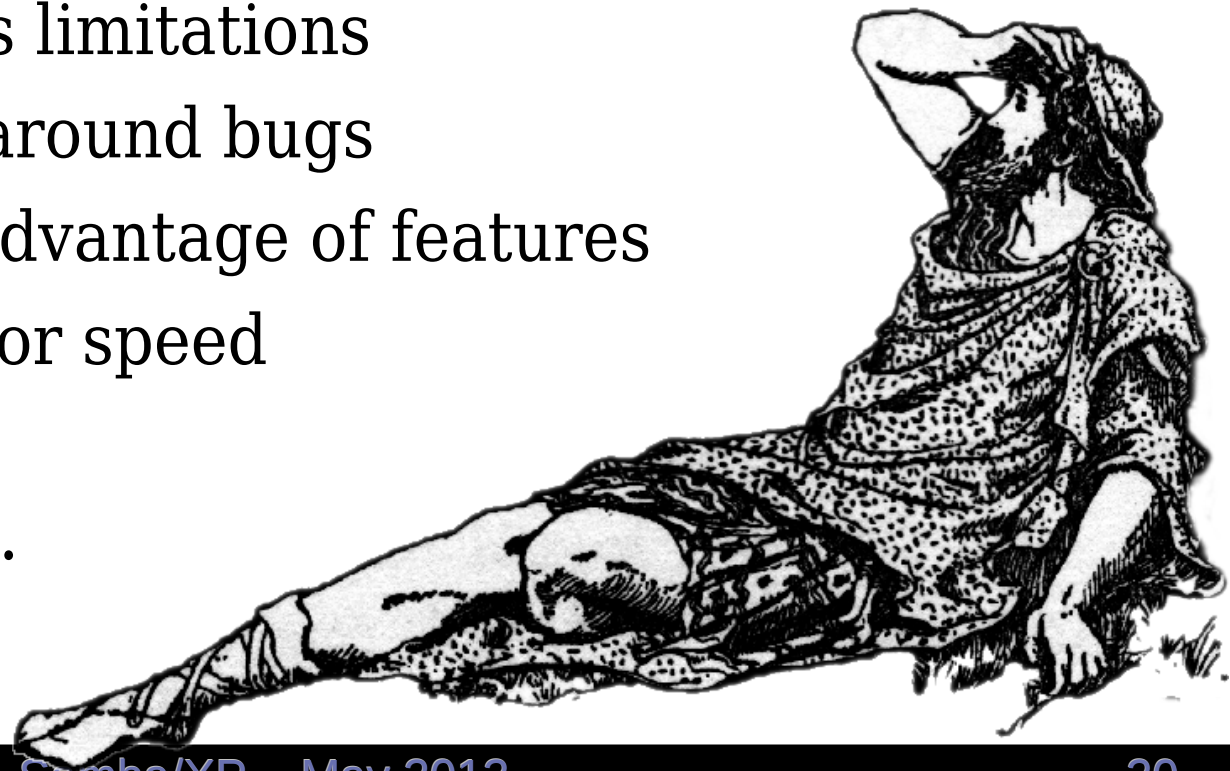
The Samba VFS Layer

Samba is Flexible

The VFS layer allows us to adapt Samba to the behaviors of the underlying file system.

- Bypass limitations
- Work around bugs
- Take advantage of features
- Tune for speed

...and more.





Getting to Know Gluster



Gluster Basics

Gluster is a Distributed File System



- Each node provides storage “Bricks”
- Each Brick is actually a directory
- Bricks are bound together as “Volumes”
- Volumes are distributed and/or replicated
- Made available via “Access Methods”
- SMB (Samba) is one such Access Method



Gluster Basics

Gluster can be FUSE Mounted

- Just another access method
- Client may be local or remote
 - Local to the Gluster server node
 - Remote on any networked client node
- Samba runs on the FUSE mount (until recently)

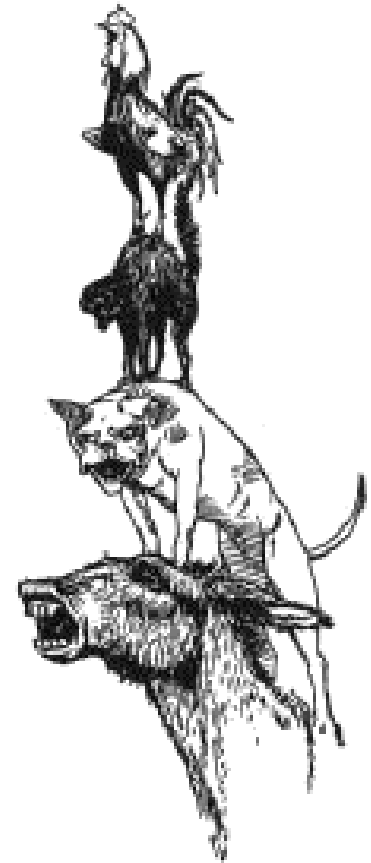


The Gluster server daemon provides a FUSE interface, used for local or remote access to Gluster volumes.



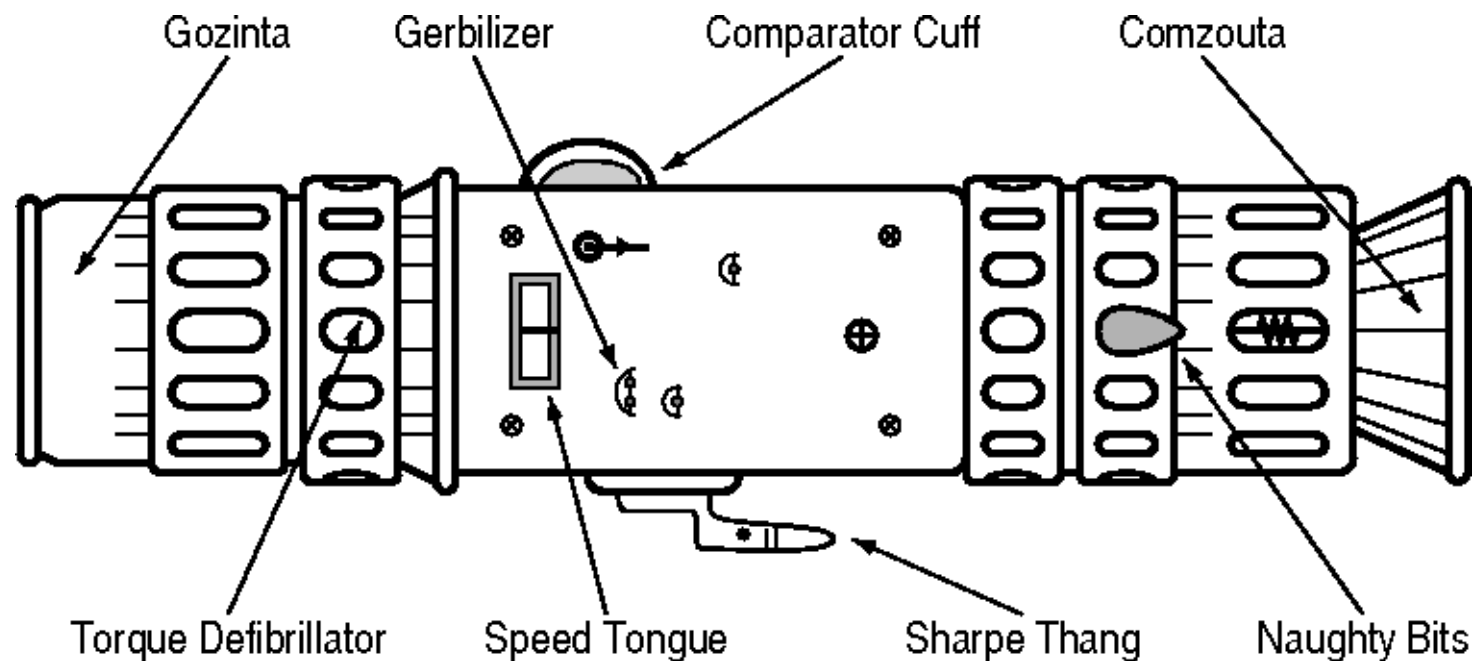
Samba and Gluster and FUSE

(oh my)





Samba/Gluster/FUSE



**Now we put the pieces together
and see how well they fit.**



Samba/Gluster/FUSE

“Samba is a semantic translation machine”

—We said that already

- 🚲 Samba translates from Windows to POSIX and back again
- 🚲 Samba expects POSIX behaviors:
 - 🔵 Read/Write Coherency
 - 🔵 POSIX byte-range locking support
- 🚲 Samba also wants extra FS features:
 - 🔵 Extended Attributes
 - 🔵 “POSIX” ACLs
 - 🔵 RichACLs (would be nice)





Samba/Gluster/FUSE

Gluster

...is adaptable, and cool.

- ▶ Can add support for SMB-specific features:
 - Windows ACLs?
 - OpLocks and Leases?
 - Windows timestamps?
- ▶ “Translators” stack like Samba VFS modules



There are possibilities to explore here.



Samba/Gluster/FUSE

FUSE

...presents a problem.

- 🔑 A generic mount point
- 🔑 No way to enable/disable per-access-method features

FUSE provides a single standard interface, but effectively locks away Gluster internals.

What's a coder to do?



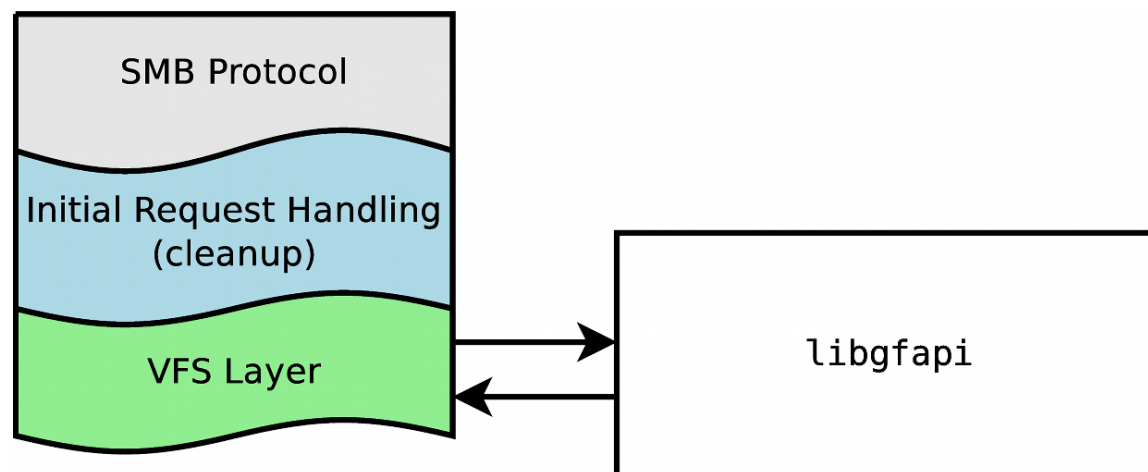


Samba/Gluster/FUSE

vfs_glusterfs

- New code, recently submitted
- Similar to the Ceph VFS

Writing the first draft VFS module took less than a week. (Credit not mine.)





Enter CTDB





CTDB Basics

CTDB offers three basic services:

- 👁 Distributed Metadata Database
- 👁 Node Failure Detection/Recovery
- 👁 IP Address (Service) Failover

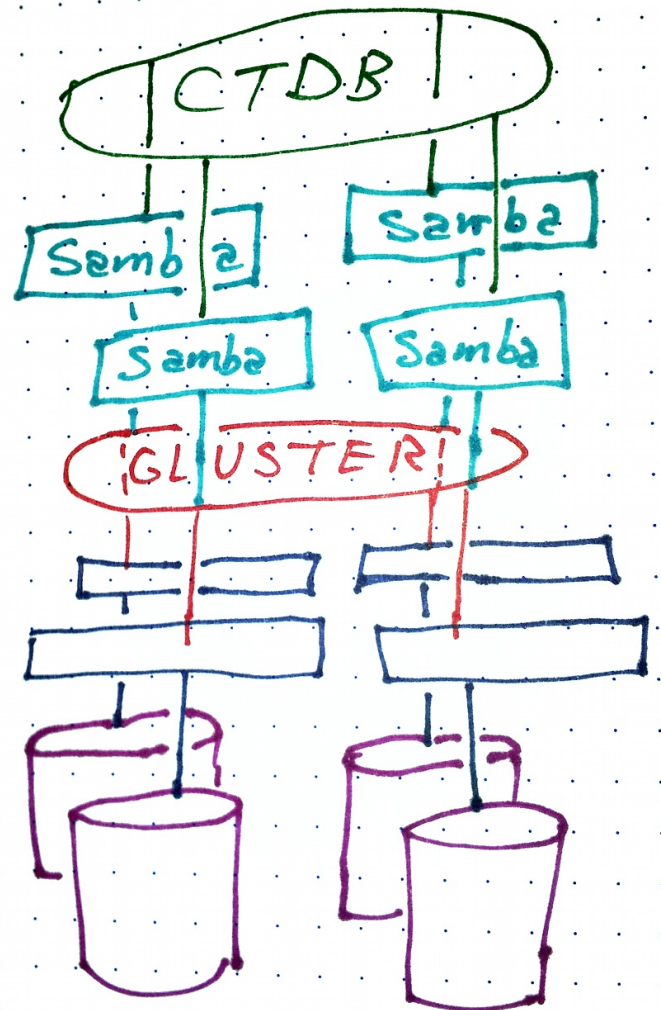
CTDB allows multiple Samba servers to operate synchronously, over the same file system.



CTDB Basics

CTDB Forms a Samba Cluster

- ✿ Separate from the underlying cluster
- ✿ May duplicate some activities
 - Messaging
 - Heartbeat
- ✿ Flexible configuration





What We Found





What we Found

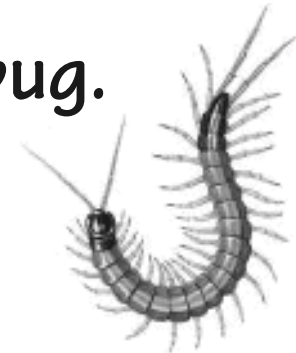
POSIX Byte Range Locking



"To be blunt, unless your cluster filesystem is called GPFS, locking is probably completely broken and should be avoided."

— From Samba-Technical, 29-Mar-2013




FIXED: F_GETLK return value bug.





What we Found

Cache Coherency

-  Stock config fails ping_pong.
-  Caching occurs in multiple locations in Gluster.
-  Turning off caching seems to solve the problem, but performance suffers.

SMB has strict locking and consistency requirements.







What we Found

CTDB Node Banning

Under Heavy Load, CTDB permanently bans a running node.

-  Recently discovered.
-  May be related to a known (and fixed) bug in the version of CTDB we are testing.

Will re-test with a newer CTDB version.





What we Found

Slow Directory Lookups

Samba must do extra work to detect and avoid name collisions.

 WiNdoWs IS cAse-INsensitive

 POSIX is case-sensitive

FUSE calls to `getdents(2)` were taking a very long time, and there were a lot of calls to `getdents(2)`.

FIXED: ...by using `vfs_gluster`.







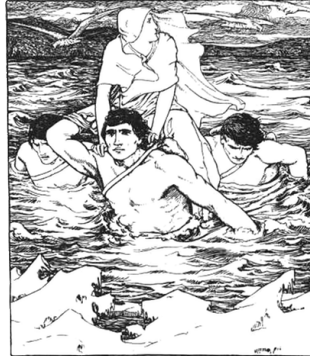
The End



This page intentionally
includes a pointless message



Samba on Gluster



The Things We Do For Love

Christopher R. Hertel
Samba Team

Jose A. Rivera
Just Enjoy the Ride

Samba/XP • May, 2013
Copyright © 2013 Christopher R. Hertel



Introductions

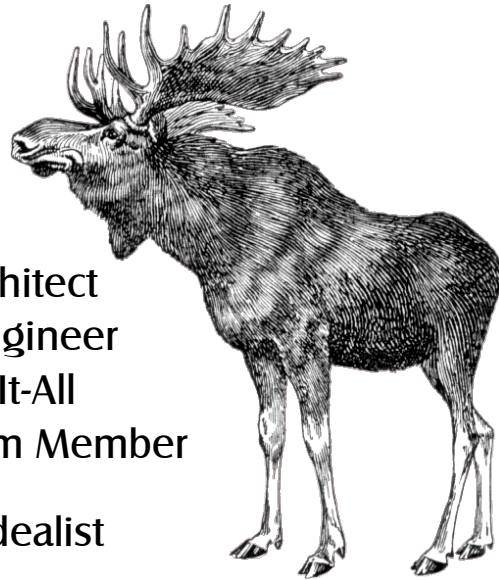




Introductions

Me

- ▶ Author
- ▶ Storage Architect
- ▶ Network Engineer
- ▶ SMB Know-It-All
- ▶ Samba Team Member
(since '98)
- ▶ Incurable Idealist



A ruminant mammal (Geekus geekus) with long legs, humped shoulders, and broadly palmated antlers.



Introductions

The Other Guy



Swimming in the deep end
of the SMB cesspool since
2008.



Introductions

Now
With



redhat®

The opinions expressed are my own
and not necessarily those of my
employer, my spouse, my childrenz,
the dog, or "the Voices".





Introductions

You





Introductions

Where are we going?

- 🌐 “Metadata” and “Semantics”
- 🌐 Semantic Translation
- 🌐 Samba/VFS and CTDB
- 🌐 Gluster Basics
- 🌐 Gluster ♥ Samba
- 🌐 Questions
(and answers, if I have any)







What's in a Name?

A Rose by Any Other Name
Would Wither and Die

-- Alan Swann (Peter O'Toole),
My Favorite Year





What's in a Name?

Metadata: Data about Data

Things like:

- ✂ Path Names and inode numbers
- ✂ Timestamps
- ✂ Permissions/Access Controls
- ✂ File size and Quota

Metadata identifies files, provides current state, sets limits, etc.



What's in a Name?

Semantics: The Meaning

- Context and rules for interpreting metadata
- Enforcement of set limits
- Reasonable assurance of correctness

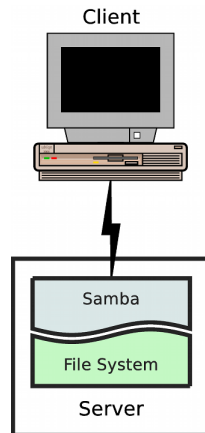
Semantic rules, such as access controls and quota limits, must be enforced by the **Operating System and File System**.





What's in a Name?

Samba is a Semantic Translator



There and back again...


- The client expects Windows semantics from the server.
- Samba expects POSIX semantics from the file system.


Samba must translate from Windows to POSIX and back again.



What's in a Name?

The Pieces Must Fit

 If Samba does not properly handle SMB protocol, we call it a bug and expect trouble.

 If the file system does not properly handle the POSIX calls made by Samba, we also expect trouble.





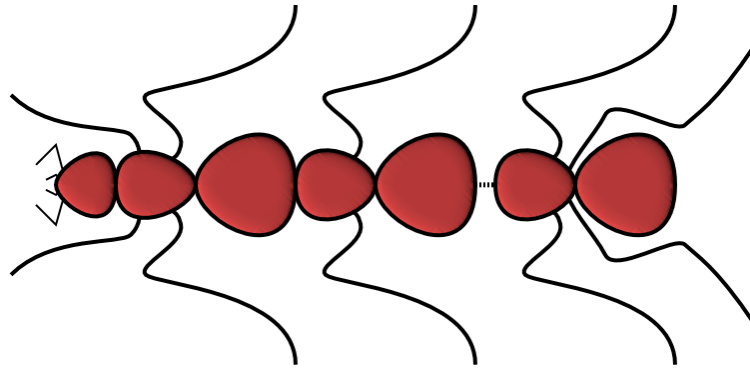
What's in a Name?



If you're looking for dragons,
you'll find them in the semantics.



The Samba VFS layer

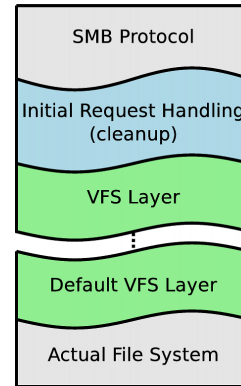




The Samba VFS Layer

Samba is Built in Layers (conceptually)

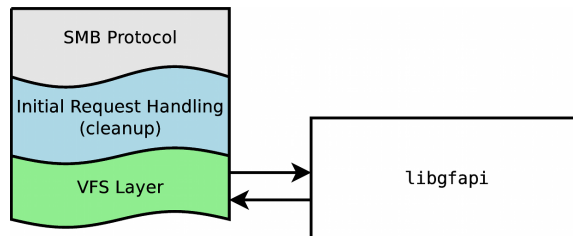
- ④ SMB messages are received and parsed
- ④ (Non-FS commands are handled elsewhere)
- ④ Call the VFS layer
- ④ The final VFS module talks to the File System
 - ⚙ Higher VFS layers may bypass lower layers





The Samba VFS Layer

If there's no real File System, we can bypass the lower VFS layers



All VFS calls must be implemented
(possibly returning ENOTSUP)
to avoid errors.



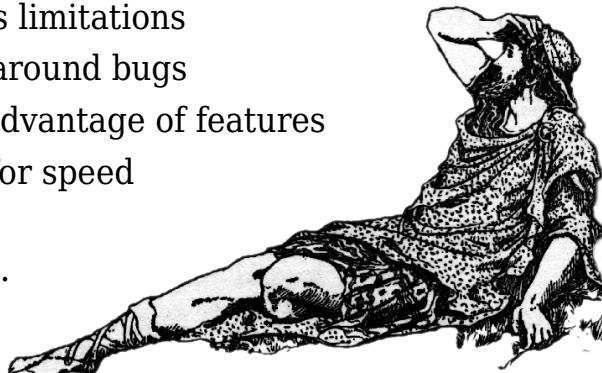
The Samba VFS Layer

Samba is Flexible

The VFS layer allows us to adapt Samba to the behaviors of the underlying file system.

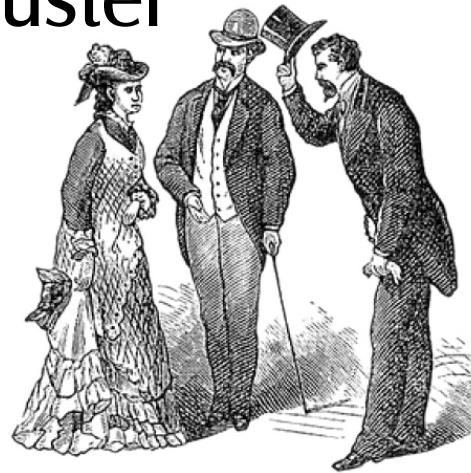
- Bypass limitations
- Work around bugs
- Take advantage of features
- Tune for speed

...and more.





Getting to Know Gluster





Gluster Basics

Gluster is a Distributed File System



- 🗄️ Each node provides storage “Bricks”
 - 📁 Each Brick is actually a directory
- 🗄️ Bricks are bound together as “Volumes”
 - 📁 Volumes are distributed and/or replicated
- 🗄️ Made available via “Access Methods”
 - 📁 SMB (Samba) is one such Access Method



Gluster Basics

Gluster can be FUSE Mounted

- Just another access method
- Client may be local or remote
 - Local to the Gluster server node
 - Remote on any networked client node
- Samba runs on the FUSE mount (until recently)



The Gluster server daemon provides a FUSE interface, used for local or remote access to Gluster volumes.



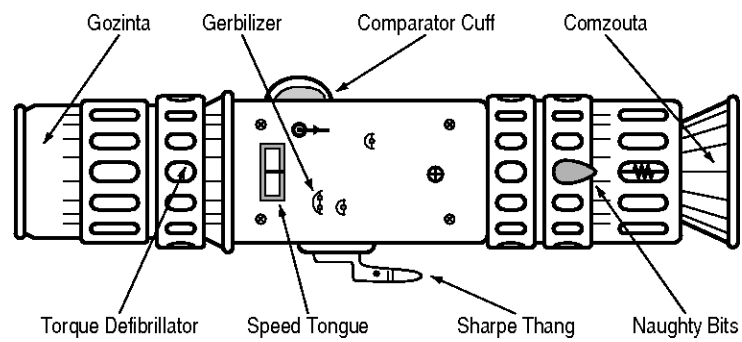
Samba and Gluster and FUSE

(oh my)





Samba/Gluster/FUSE



**Now we put the pieces together
and see how well they fit.**



Samba/Gluster/FUSE

“Samba is a semantic translation machine”

—We said that already

- 🚲 Samba translates from Windows to POSIX and back again
- 🚲 Samba expects POSIX behaviors:
 - ⦿ Read/Write Coherency
 - ⦿ POSIX byte-range locking support
- 🚲 Samba also wants extra FS features:
 - ⦿ Extended Attributes
 - ⦿ “POSIX” ACLs
 - ⦿ RichACLs (would be nice)





Samba/Gluster/FUSE

Gluster

...is adaptable, and cool.

- ▶ Can add support for SMB-specific features:
 - Windows ACLs?
 - OpLocks and Leases?
 - Windows timestamps?

- ▶ “Translators” stack like Samba VFS modules



There are possibilities to explore here.



Samba/Gluster/FUSE

FUSE

...presents a problem.

- 🔑 A generic mount point
- 🔑 No way to enable/disable per-access-method features

FUSE provides a single standard interface, but effectively locks away Gluster internals.

What's a coder to do?



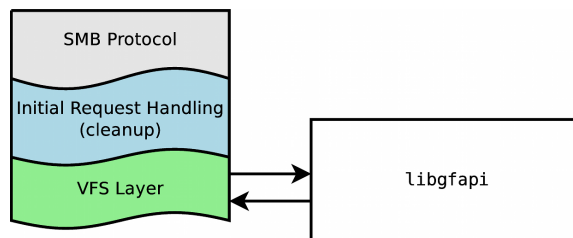


Samba/Gluster/FUSE

vfs_glusterfs

- New code, recently submitted
- Similar to the Ceph VFS

Writing the first draft VFS module took less than a week. (Credit not mine.)





Enter CTDB





CTDB Basics

CTDB offers three basic services:

- 👁 Distributed Metadata Database
- 👁 Node Failure Detection/Recovery
- 👁 IP Address (Service) Failover

CTDB allows multiple Samba servers to operate synchronously, over the same file system.

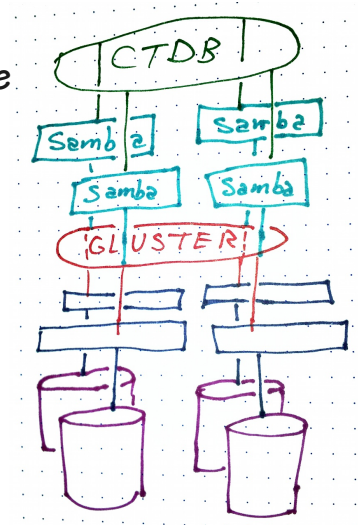




CTDB Basics

CTDB Forms a Samba Cluster

- ☘ Separate from the underlying cluster
- ☘ May duplicate some activities
 - Messaging
 - Heartbeat
- ☘ Flexible configuration





What We Found





What we Found

POSIX Byte Range Locking



"To be blunt, unless your cluster filesystem is called GPFS, locking is probably completely broken and should be avoided."

— From Samba-Technical, 29-Mar-2013

FIXED: F_GETLK return value bug.






What we Found

Cache Coherency

 Stock config fails ping_pong.

 Caching occurs in multiple locations in Gluster.

 Turning off caching seems to solve the problem, but performance suffers.




SMB has strict locking and consistency requirements.




What we Found

CTDB Node Banning

Under Heavy Load, CTDB permanently bans a running node.

 Recently discovered.

 May be related to a known (and fixed) bug in the version of CTDB we are testing.

Will re-test with a newer CTDB version.





What we Found

Slow Directory Lookups

Samba must do extra work to detect and avoid name collisions.

 WiNdoWs IS cAse-INsensiTive

 POSIX is case-sensitive

FUSE calls to `getdents(2)` were taking a very long time, and there were a lot of calls to `getdents(2)`.

FIXED: ...by using `vfs_gluster`.







The End

