Present And Future File Serving With Samba

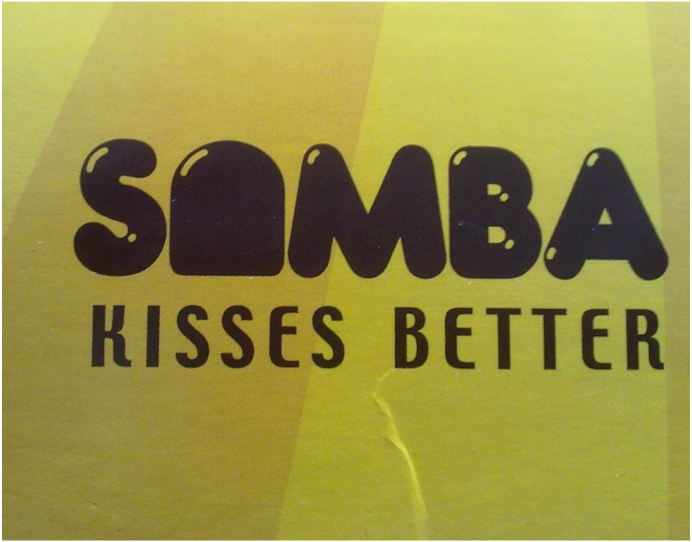
LinuxCon Europe 2014

Michael Adam

Samba Team / SerNet

October 14, 2014

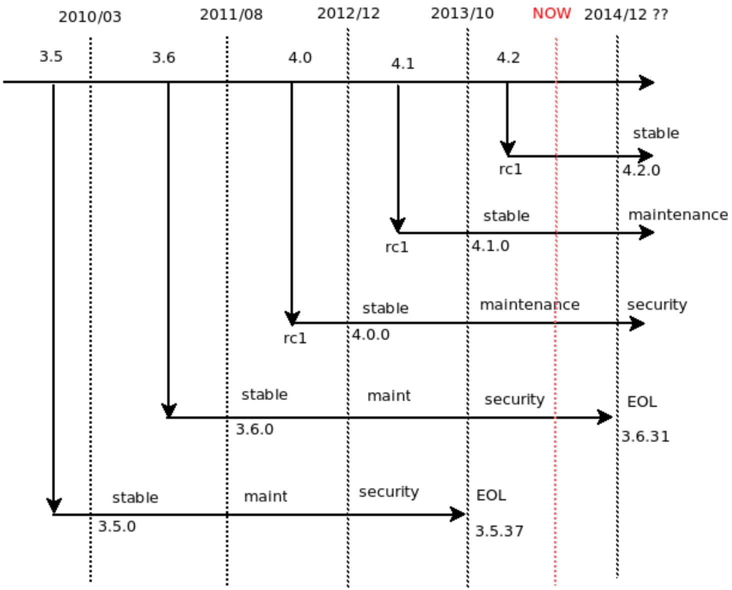
Samba...



Short History

- 1.9.17: 1996/08
- 2.0: 1999/01: domain-member, +SWAT
- 2.2: 2001/04: NT4-DC
- 3.0: 2003/09: AD-member, Samba4 project started
- 3.2: 2008/07: GPLv3, experimental clustering
- 3.3: 2009/01: clustering
- 3.4: 2009/07: merged S3+S4 code
- 3.5: 2010/03: experimental SMB 2.0
- 3.6: 2011/09: SMB 2.0
- 4.0: 2012/12: AD/DC, SMB 2.0 durable handles, 2.1, 3.0
- 4.1: 2013/10: stability
- 4.2: soon: AD trusts, performance, scalability, CTDB included

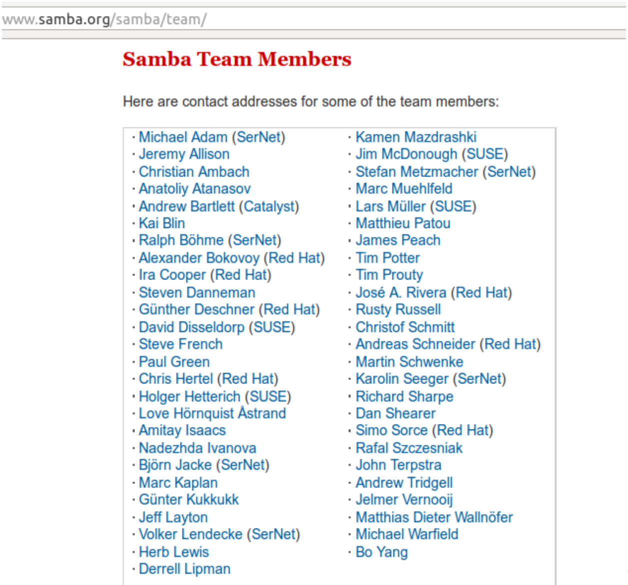
Release Stream



Release Planning

`https://wiki.samba.org/index.php/Samba_Release_Planning`

Samba Team



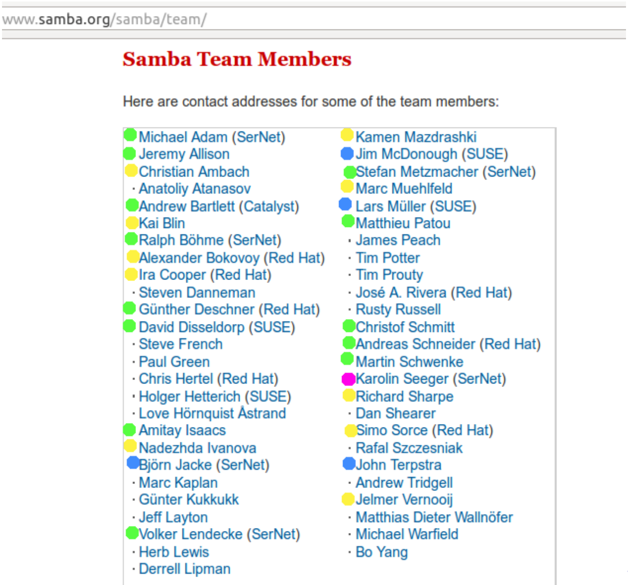
www.samba.org/samba/team/

Samba Team Members

Here are contact addresses for some of the team members:

- Michael Adam (SerNet)
- Jeremy Allison
- Christian Ambach
- Anatoliy Atanasov
- Andrew Bartlett (Catalyst)
- Kai Blin
- Ralph Böhme (SerNet)
- Alexander Bokovoy (Red Hat)
- Ira Cooper (Red Hat)
- Steven Danneman
- Günther Deschner (Red Hat)
- David Disseldorp (SUSE)
- Steve French
- Paul Green
- Chris Hertel (Red Hat)
- Holger Hetterich (SUSE)
- Love Hörnquist Åstrand
- Amitay Isaacs
- Nadezhda Ivanova
- Björn Jacke (SerNet)
- Marc Kaplan
- Günter Kukkukk
- Jeff Layton
- Volker Lendecke (SerNet)
- Herb Lewis
- Derrell Lipman
- Kamen Mazdrashki
- Jim McDonough (SUSE)
- Stefan Metzmacher (SerNet)
- Marc Muehlfeld
- Lars Müller (SUSE)
- Matthieu Patou
- James Peach
- Tim Potter
- Tim Prouty
- José A. Rivera (Red Hat)
- Rusty Russell
- Christof Schmitt
- Andreas Schneider (Red Hat)
- Martin Schwenke
- Karolin Seeger (SerNet)
- Richard Sharpe
- Dan Shearer
- Simo Sorce (Red Hat)
- Rafal Szczesniak
- John Terpstra
- Andrew Tridgell
- Jelmer Vernooij
- Matthias Dieter Wallnöfer
- Michael Warfield
- Bo Yang

Samba Team



www.samba.org/samba/team/

Samba Team Members

Here are contact addresses for some of the team members:

- 🟢 Michael Adam (SerNet)
- 🟢 Jeremy Allison
- 🟡 Christian Ambach
- · Anatoliy Atanasov
- 🟢 Andrew Bartlett (Catalyst)
- 🟡 Kai Blin
- 🟢 Ralph Böhme (SerNet)
- 🟡 Alexander Bokovoy (Red Hat)
- 🟡 Ira Cooper (Red Hat)
- · Steven Danneman
- 🟢 Günther Deschner (Red Hat)
- 🟢 David Disseldorp (SUSE)
- · Steve French
- · Paul Green
- · Chris Hertel (Red Hat)
- · Holger Hetterich (SUSE)
- · Love Hörnquist Åstrand
- 🟢 Amitay Isaacs
- 🟡 Nadezhda Ivanova
- 🔵 Björn Jacke (SerNet)
- · Marc Kaplan
- · Günter Kukkukk
- · Jeff Layton
- 🟢 Volker Lendecke (SerNet)
- · Herb Lewis
- · Derrell Lipman
- 🟡 Kamen Mazdrashki
- 🔵 Jim McDonough (SUSE)
- 🟢 Stefan Metzmacher (SerNet)
- 🟡 Marc Muehlfeld
- 🔵 Lars Müller (SUSE)
- 🟢 Matthieu Patou
- · James Peach
- · Tim Potter
- · Tim Prouty
- · José A. Rivera (Red Hat)
- · Rusty Russell
- 🟢 Christof Schmitt
- 🟢 Andreas Schneider (Red Hat)
- 🟢 Martin Schwenke
- 🟣 Karolin Seeger (SerNet)
- 🟡 Richard Sharpe
- · Dan Shearer
- 🟡 Simo Sorce (Red Hat)
- · Rafal Szczesniak
- 🔵 John Terpstra
- · Andrew Tridgell
- 🟡 Jelmer Vernooij
- · Matthias Dieter Wallnöfer
- · Michael Warfield
- · Bo Yang

CHILLI
SAMBA
FLAVOUR*
'14

Samba File Server Topics / Challenges

1. performance: scalable file server
   - scale-up: exhaust powerful boxes
   - scale-out: flexible all-active clusters
   - scale-down: perform well on low-end boxes
2. interop: multi-protocol access (nfs, afp, ...)
3. server workloads / SMB features
   - tune for: small # of connections, threaded applications
   - Hyper-V, ...
   - SMB3 (clustering, RDMA, ...)
4. special file systems support (gluster, ceph, gpfs, btrfs, ...)
5. cloud / openstack?...

Performance

Performance - low end systems

Reduction of CPU usage for low profile platforms like arm (SMB2)

- Samba 4.0:
  - didn't saturate 1G nic (arm), CPU 100%
- reduced memory allocations
- instrument SMB 2.1 multi-credit / large MTU
- Samba 4.2:
  - saturates 1G nic (arm), CPU $<$ 100%
- ⇒ continuing

Performance - DB performance

TDB

- trivial database
- used for IPC (smbd processes)
- cluster (CTDB): local copies

hot databases

- `locking.tdb` (open files)
- `brlock.tdb` (byte range locks)
- `notify_index.tdb` (for change notify)

Performance - DB performance

problem 1

- fcntl byte range locks for record locks
- contention via single kernel spinlock

solution

- alternative to fcntl: pthread robust mutexes
- ⇒ massive speedup
- ⇒ included in TDB 1.3.1, Samba 4.2

Performance - DB performance

problem 2

- freelist:
    - single chain, contended (`locking.tdb`)
    - gets fragmented (singly linked)
- especially a problem in ctdb-cluster: vacuuming

improvements

- make use of small per-record freelists (dead records)
- add automatic defragmentation upon traversal
- ⇒ included in TDB 1.3.1, Samba 4.2

Performance - DB performance

problem 3

- change notify not scalable

first improvement

- restructured `notify.tdb` to
  - global `notify_index.tdb` and
  - local `notify.tdb`
  - ⇒ better but still not good enough for some workloads

next steps

- replace DB-approach by new scalable, async notify daemon using messaging
- some false positives do not harm
- ⇒ TODO

Performance - scaling

parellelism

- samba is multi-process:
  - smbd child process ↔ TCP connection
  - event-loop in one process
- within a smbd process:
  - pthread-pool jobs for potentially blocking syscalls
  - ⇒ parallelism for reads/writes
  - default for async I/O since Samba 4.0

Performance - scaling

messaging

- classical messaging:
  - messages.tdb and signals between processes
  - does not scale well
- new massaging in Samba 4.2:
  - fast and scalable messaging based on unix datagram messages
  - ⇒ WIP: integrate with AD/DC messaging
  - ⇒ features fd-passing for sockets (SMB3 multi-channel)
  - ⇒ TODO: integrate into CTDB inter-node-messaging

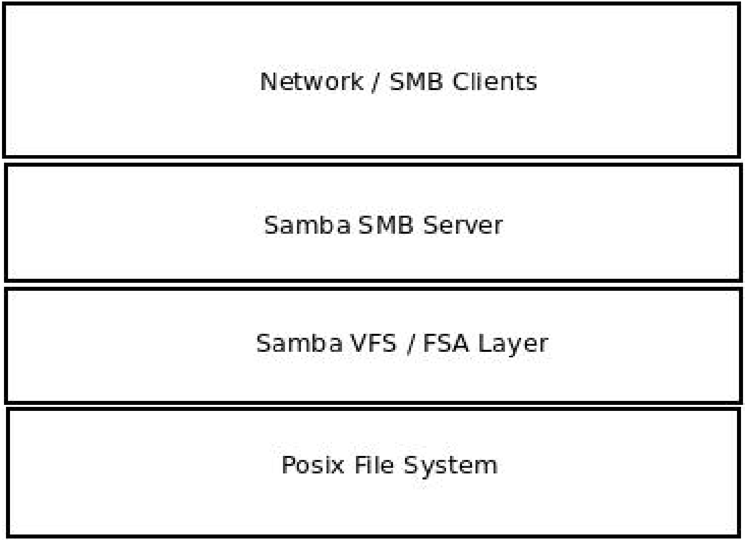
Interop

Interop-Central

multi-protocol access

- nfs (kernel, ganesha, ...)
- afp: netatalk
- local access
- SMB2+ unix-extensions

File Server Layout/Scope



Network / SMB Clients

Samba SMB Server

Samba VFS / FSA Layer

Posix File System

Interop - Fruit

- MacOS 10.9: SMB 2.1 preferred file protocol
- `vfs_fruit` - new module in Samba 4.2
- spotlight
  - indexed search
  - dcerpc service
  - ⇒ under review
- AAPL
  - SMB2 create context
  - speed up directory listings
  - ⇒ under review

Fruit Demo

SMB features

SMB features in Samba - SMB2

- SMB 2.0 (Vista / 2008):
  - durable file handles [4.0]
- SMB 2.1 (Win7 / 2008R2):
  - multi-credit / large mtu [4.0]
  - dynamic reauthentication [4.0]
  - leasing [WIP++]
  - resilient file handles [WIP-tracer]

SMB features in Samba - SMB3

- SMB 3.0 (Win8 / 2012):
  - new crypto (sign/encrypt) [4.0]
  - secure negotiation [4.0]
  - durable handles v2 [4.0]
  - persistent file handles [WIP.tracer]
  - multi-channel [WIP+]
  - SMB direct [designed/starting]
  - cluster features [designing]
    - witness [WIP]
  - storage features [WIP]
- SMB 3.02 (Win8.1 / 2012R2): [WIP]
- SMB 3.1 (Win10 / 2014): [ess.DONE]

Multi-Channel - Windows/Protocol

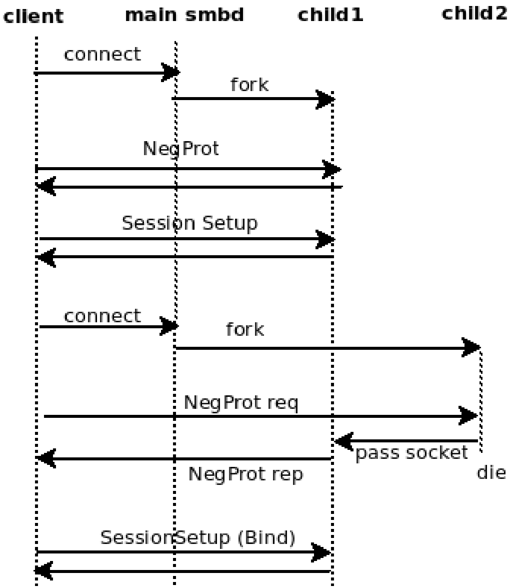
- find interfaces with interface discovery: `FSCTL_QUERY_NETWORK_INTERFACE_INFO`
- bind additional TCP (or RDMA) connection (channel) to established SMB3 session (session bind)
- windows: uses connections of same (and best quality)
- windows: binds only to a single node
- replay / retry mechanisms, epoch numbers

Multi-Channel - Samba

- samba/smbd: multi-process
  - process ⇔ tcp connection
  - ⇒ transfer new connection to existing smbd
  - use fd-passing (sendmsg/recvmsg)

- preparation: messaging rewrite using unix dgm sockets with sendmsg [DONE,4.2]
- add fd-passing [DONE,4.2]
- transfer connection already in negprot (ClientGUID) [ess.DONE]
- implement channel epoch numbers [WIP]
- implement interface discovery [WIP]

Multi-Channel - Samba

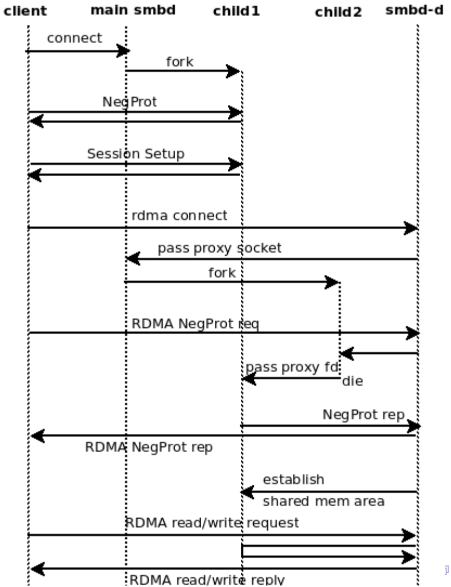


Multi-Channel Demo

SMB Direct (RDMA)

- windows:
  - requires multi-channel
  - start with TCP, bind an RDMA channel
  - reads and writes use RDMB write/read
  - protocol/metadata via send/receive
- wireshark dissector: [DONE]
- samba (TODO):
  - prereq: multi-channel / fd-passing
  - buffer / transport abstractions [TODO]
  - problem: libraries: not fork safe and no fd-passing
    ⇒ central daemon (or kernel module) to serve as RDMA "proxy"

SMB Direct (RDMA) - Plan



SMB features in Samba

https://wiki.samba.org/index.php/Samba3/SMB3

Misc

Misc

File Systems

- gpfs, gluster, ceph, btrfs...
- support through vfs modules
- fuse-based: avoid context switches
- instrument SMB3 storage features (fsctls)

Misc

Testing

- unprivileged selftest, autobuild
- selfcontained testing: wrapper
  - socket wrapper
  - nss wrapper
  - uid wrapper
  - resolv wrapper [new]
- externalized as separate projects:
  - ⇒ `http://cwrap.org/`
  - git on samba.org
  - ⇒ Andreas Schneider's talk

Forecast: Cloudy

Possible involvement with OpenStack

- SMB storage service for Windows (and other) VMs
- SMB3 storage backend for Hyper-V images
- also: chances for AD-integration into auth

Credits

especially but not exclusively

- Volker Lendecke
- Stefan Metzmacher
- Ralph Böhme
- Jeremy Allison
- David Disseldorp
- Andreas Schneider

Conclusion

Conclusion

Remember

- Samba 4.X is quite different from 3.Y

What's coming?

- Performance: the story continues
- Interop: strengthen strenths
- SMB(3) features: a lot to come (⇒ cluster, hyper-v, ...)
- Some clouds in the sky...

Thanks for your attention!

Questions?

```
obnox@samba.org
ma@sernet.de
```