Samba and the road to 100,000 users

Presented by Andrew Bartlett
Samba Team - Catalyst // 16 Jan 2017

catalyst
open source technologists

SAMBA TEAM

Samba is a member project of the
Software Freedom Conservancy

software freedom
conservancy

Andrew Bartlett

- Samba developer since 2001
- Working on the AD DC since soon after the start of the 4.0 branch, since 2004!
  - Driven to work on the AD DC after being a high school Systems Administrator
- Working for Catalyst in Wellington since 2013
  - Now leading a team of 5 Catalyst Samba Engineers
- These views are mine alone
- Please ask questions during the talk

In 2017, Samba is released Fast!

- New release cycle
  - **Strict** 6 months cycle
  - Supported for 18 months
- Support status
  - Samba 4.0 is 4 years ago old now
  - Samba 4.6 about to be released
  - Samba 3.6 **and** 4.0, 4.1, 4.2 already out of security support
    - Your distributor may be providing extended support for Samba

But also much faster as an AD DC

- In a two-hour benchmark adding users and adding to four groups:
  - Samba 4.4: 26,000 users
  - Samba 4.5: 48,000 users
  - Samba 4.6: 55,000 users
  - With pending patches: 85,000 users!
    - The first 55,000 added in just 50mins

Scale is important to us

- Every user account implies a computer account also
  - Computers are domain joined and get ‘user’ objects
- Samba3 deployed widely using OpenLDAP for the hard work
  - Ideally we need to match that scale!
- We really want to remove barriers, both real and perceived to Samba’s use
  - Not reasonable to ask that Samba be deployed on the very edge of its capability

Rebuilding Samba for performance

- Once we started looking at performance, we quickly found things to fix
- Performance issues, not bugs, are now the biggest area of work!
  - Customers deploying Samba at scale
  - Customers growing and very keen to keep Samba
- Very glad to be the backbone of some multi-national corporate networks!

Replication as a performance bottleneck

- So what if it takes time to add 10,000 users or so?
  - Companies can't hire that fast anyway
  - Even students don't enrol that fast!
- Biggest bottleneck is adding new DCs to Samba domains
  - e. g. opening a new office

Improvements in Samba 4.5

- Samba 4.5 addressed major issues with the client-side of replication
  - 3 of the 4 $O(n^2)$ loops removed
  - Critical as these were under the transaction lock
- Turned on graph (rather than all to all) replication by default
  - Previously every Samba DC would notify every other Samba DC about changes
  - This could trigger a short replication storm

Some improvement in 4.6

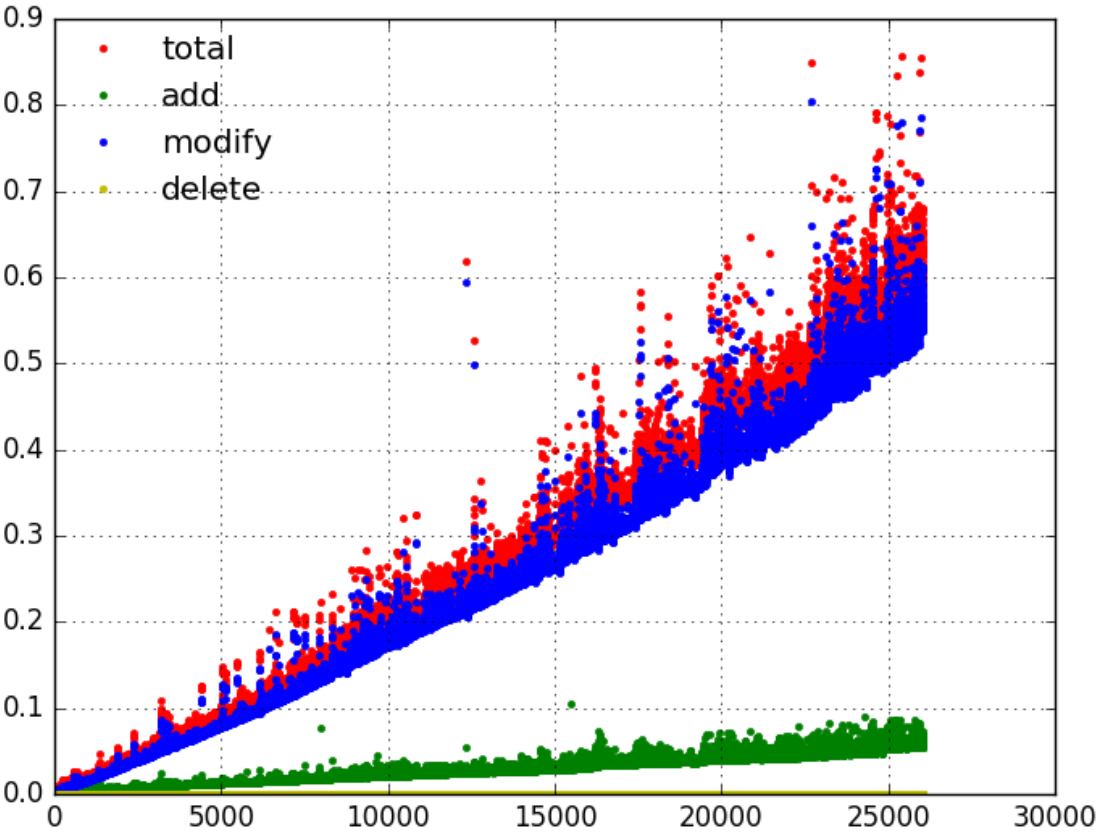
- Samba 4.6 will avoid over-replication of links
  - When replicating from server A, we also ask is what changes it got from B
  - That means we don't need to ask B for changes directly
  - We did this for attributes, but didn't do this for links previously
- Faster parsing of links also improved performance around 20% for some tasks
  - Avoid sscanf() and malloc()

Two steps forward, one back

- But there appears to be a regression with Samba 4.5 and 4.6
  - When joining new Samba 4.5 and 4.6 DCs we request ‘parents before children’
  - Sorting the DB (for initial replication) appears to take longer than a timeout period
  - Stefan Metzmacher from SerNet has a likely fix
  - My team is looking to isolate and confirm or improve that fix

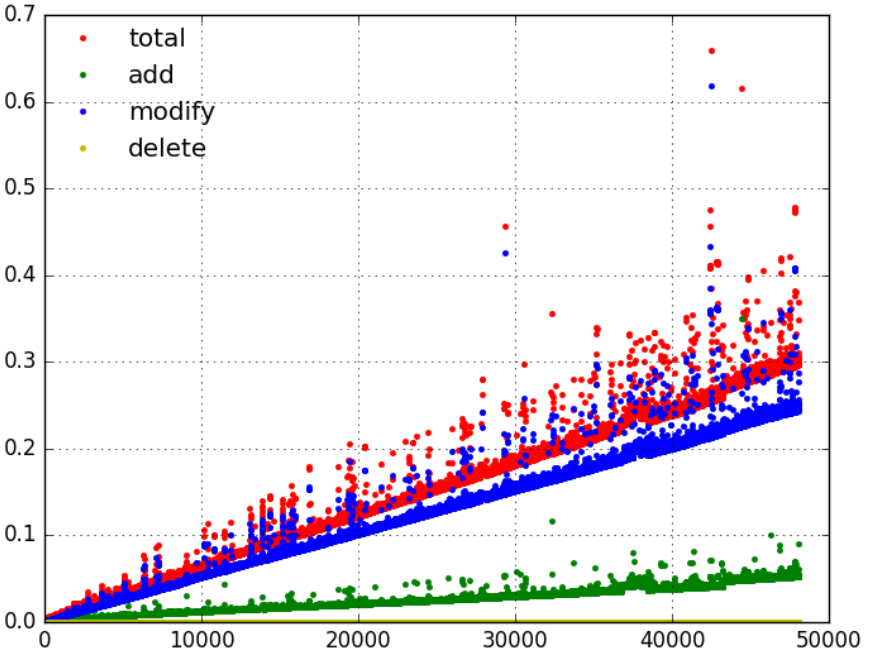
Before optimisation: Samba 4.4

- Adding a user and adding that user to four groups in a two-hour limit

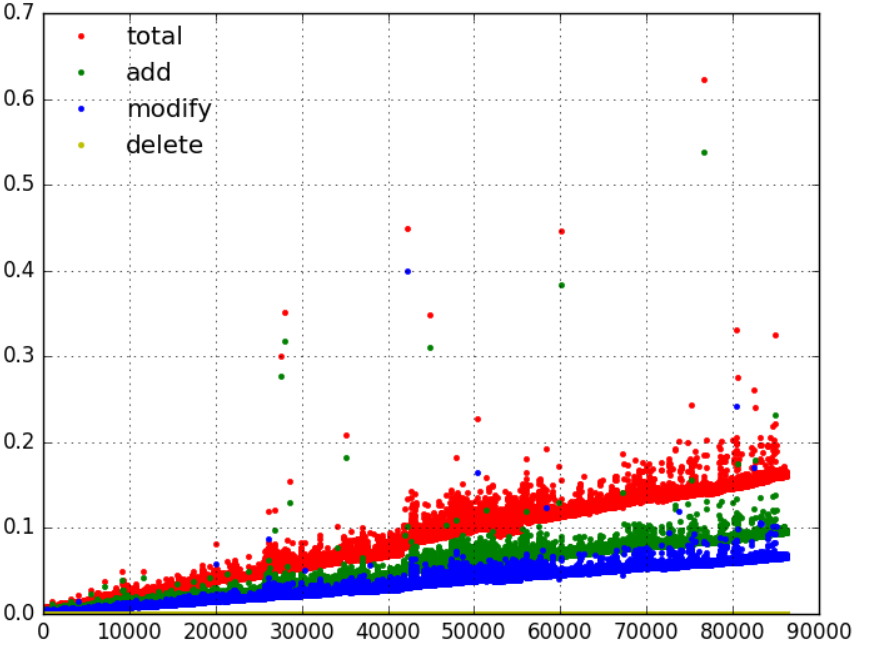


Much improved scale factors: two-hour limit

Samba 4.5



Pending patches for 4.7



Supporting more users on each DC

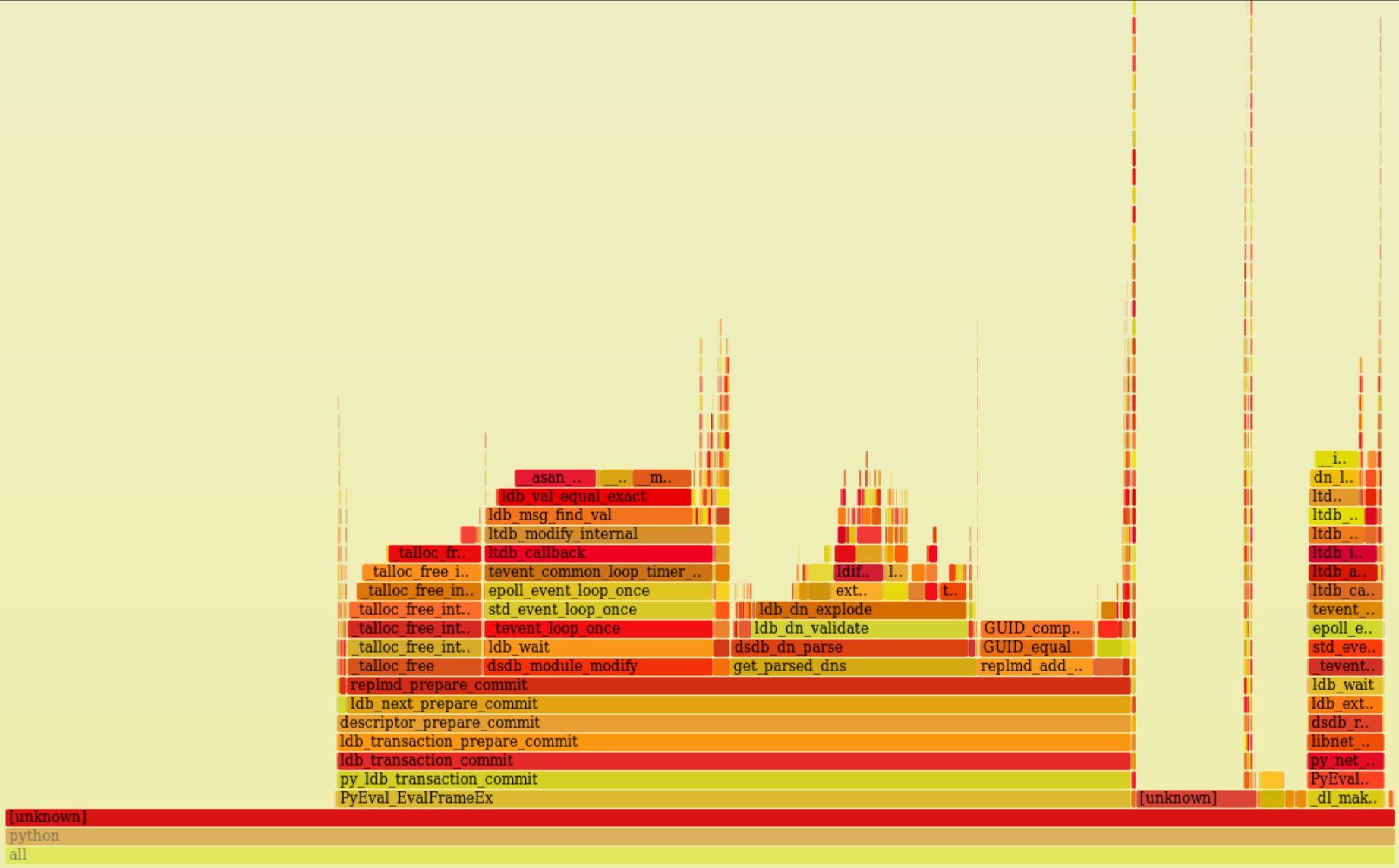
- Hoping to avoid needing to run extra DCs to spread the load
- Samba 4.6 removes single-process restrictions on NETLOGON
  - Really important for 802.1x backed wireless authentication
  - Unbreak the WiFi and watch the DC melt instead :-(
- Samba 4.7 will support a multi-process LDAP server
  - Easy to turn on in the code
  - Currently fork() and cleanup for exit() costs are too high

Finding a Bottleneck: Number of group members

- The slowest part of the code was not user or group objects
- Main cost was each group member (a link and backlink)
- As discussed, quite painful during replication
  - Client-side link processing is during the transaction lock
  - 4 different $O(n^2)$ loops found!
  - talloc() and talloc_free() quite expensive

Flame graphs

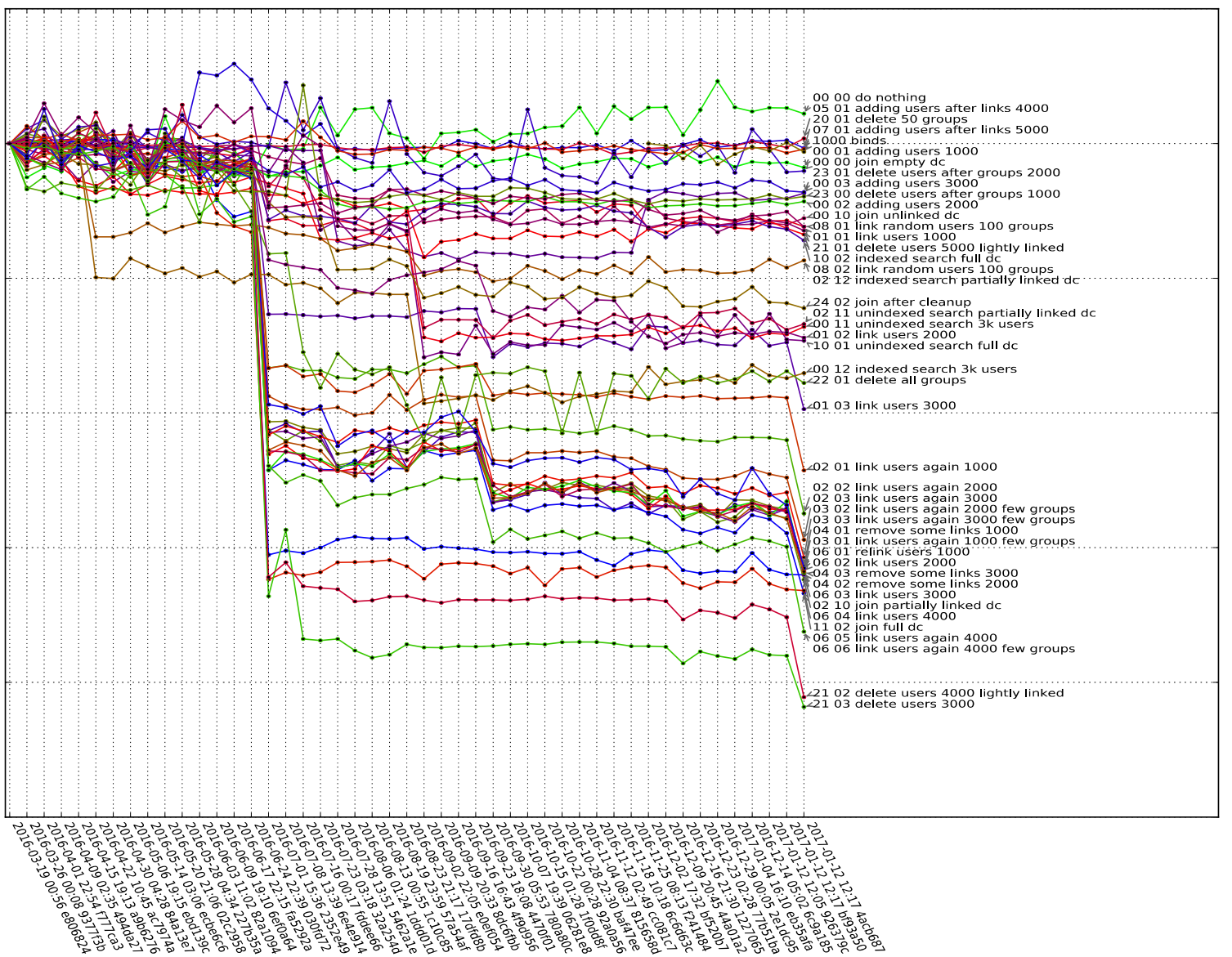
- We used linux perf and Brendan D. Gregg's Flame Graphs
- If you have performance issues with Samba:
  - Install the linux-perf tools
  - Clone https://github.com/brendangregg/FlameGraph
  - Follow http://www.brendangregg.com/FlameGraphs/cpuflamegraphs.html
  - Send us the Interactive SVG
    - Sensitive user data not included, just function names!


__i..
__asan_..
__..
__m..
dn_l..
ldb_val_equal_exact
ltd..
ldb_msg_find_val
ltdb_..
ltdb_modify_internal
ltdb_..
_talloc_fr..
ltdb_callback
ltdb_i..
_talloc_free_i..
tevent_common_loop_timer_..
ldif..
l..
ltdb_a..
_talloc_free_in..
epoll_event_loop_once
ext..
t..
ltdb_ca..
talloc_free_int..
std_event_loop_once
ldb_dn_explode
tevent_..
_talloc_free_int..
_tevent_loop_once
ldb_dn_validate
GUID_comp..
epoll_e..
_talloc_free_int..
ldb_wait
dsdb_dn_parse
GUID_equal
std_eve..
_talloc_free
dsdb_module_modify
get_parsed_dns
replmd_add_..
_tevent..
replmd_prepare_commit
ldb_wait
ldb_next_prepare_commit
ldb_ext..
descriptor_prepare_commit
dsdb_r..
ldb_transaction_prepare_commit
libnet_..
ldb_transaction_commit
py_net_..
py_ldb_transaction_commit
PyEval..
PyEval_EvalFrameEx
[unknown]
_dl_mak..
[unknown]
python
all

Flame graphs are interactive

- When used in a web browser
- I blogged about this for catalyst:
  - Burning Samba with perf and FlameGraph
  - https://catalyst.net.nz/blog/burning-samba-perf-and-flamegraph

Performance graphs from March 2016

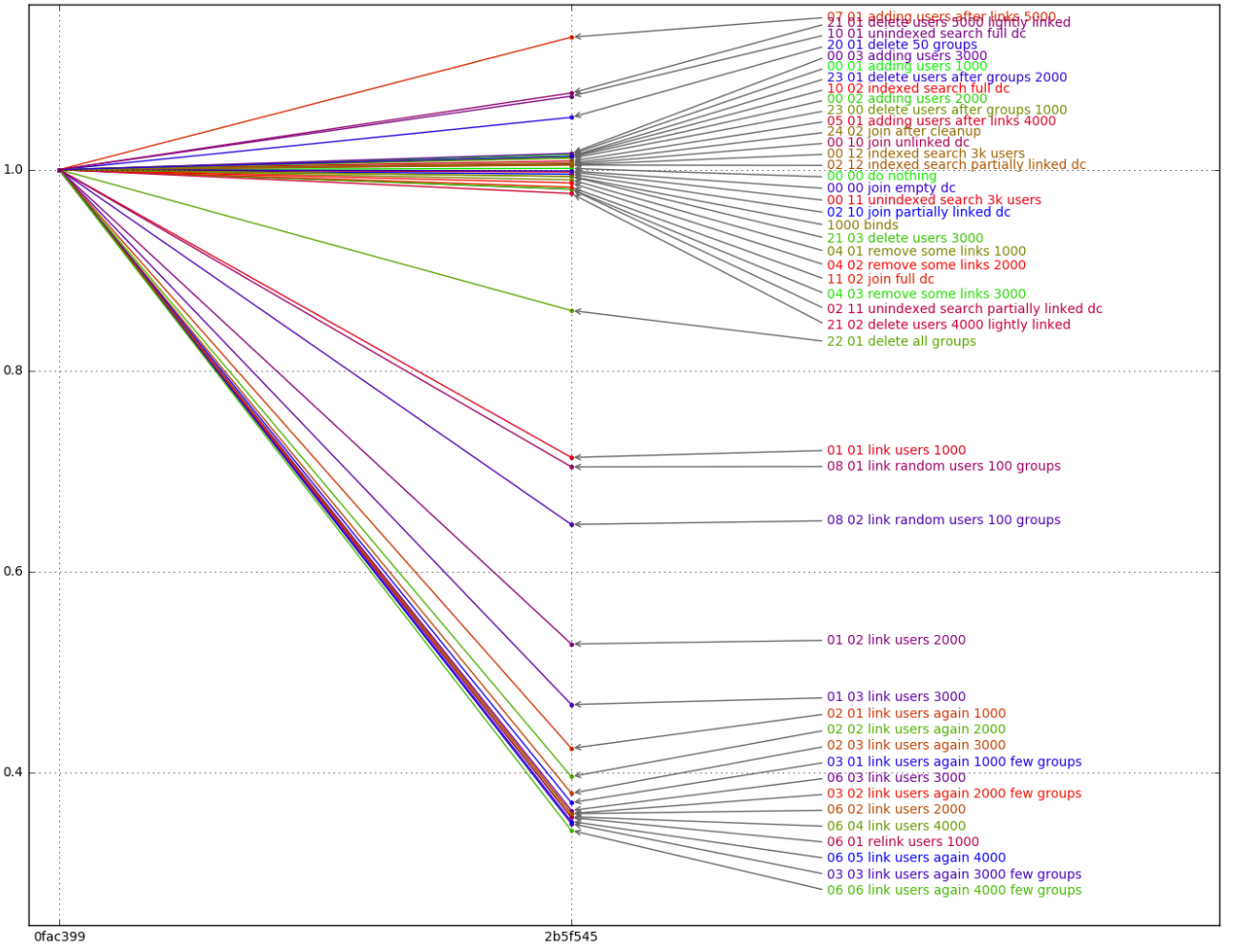


The difference a sorted list makes!

- Our code needs to find group members to support add/delete/modfy
- Previously, we had to parse every link
  - member: <GUID=a57fda98-631c-4897-8b2d-e3d8517d44f7>;
    <RMD_ADDTIME=1312841678300 00000>;
    <RMD_CHANGETIME=131284167830000000>;<RMD_FLAGS=0>;
    <RMD_INVOCID=a0a5a67 8-5114-4e30-bede-691df820b485>;
    <RMD_LOCAL_USN=3723>;<RMD_ORIGINATING_USN=3723 >;<RMD_VERSION=0>;
    <SID=S-1-5-21-734207269-1740946421-976543298-1103>;
    CN=testallowed,CN=Users,DC=samba,DC=example,DC=com
- Now we sort by GUID, and so can do a binary search

Pending changes for sorting links

- Over a 60% drop in time for some tests



The future for performance

- Remove other O(n) and O(n^2) operations
  - Multi-valued attribute handling
- Better index handling
  - Our current index code is still very much a first pass
  - Proposal to move to a GUID based index
- Reaching the limits for the current DB:
  - memcpy() and memmove() from ldb_tdb transactions are 20% of the time

Lightening Memory-mapped Database from Symas

- The company behind OpenLDAP
- Built by Howard Chu to make OpenLDAP fly
- LMDB backend prototyped by Red Hat for sssd
  - Appears to be 3 times faster for some operations
- Garming Sam has been working on reimplementation
  - Preparing it in a way that could be submitted
  - Based more tightly on the TDB LDB backend
- As of Friday it successfully ran provision!

Maintaining Performance and scale

- Large scale operation needs to be part of Samba's autobuild
- Project to develop a new performance metric for Samba domains
  - Currently awaiting client approval
- Ongoing graphing of performance measurements
  - Try to spot regressions before they get too old

Help wanted!

- For the performance metric tool I need to calibrate it
- I need volunteers running AD willing to run a tshark script
  - Windows or Samba AD welcome
  - What does your busy hour look like?
  - What is the pattern of requests?
- E-mail abartlet@samba.org if you can help

Beyond performance

- Inter-forest trusts
  - Because sometimes sharding the data is really the right approach
  - Initial support in Samba 4.3 but more work needed
- Inter-domain trusts
  - To allow migration from per-department Samba domains
  - Still pending further development
  - Most companies move to one domain, one forest

Beyond just pure AD

- What would make Samba compelling for your networks?
- Can be integrate better with POSIX systems?
  - Become the natural directory for Linux networks too?
  - Can Mac OS X be better supported?
- Samba 4.5 includes a Samba-specific password sync extension

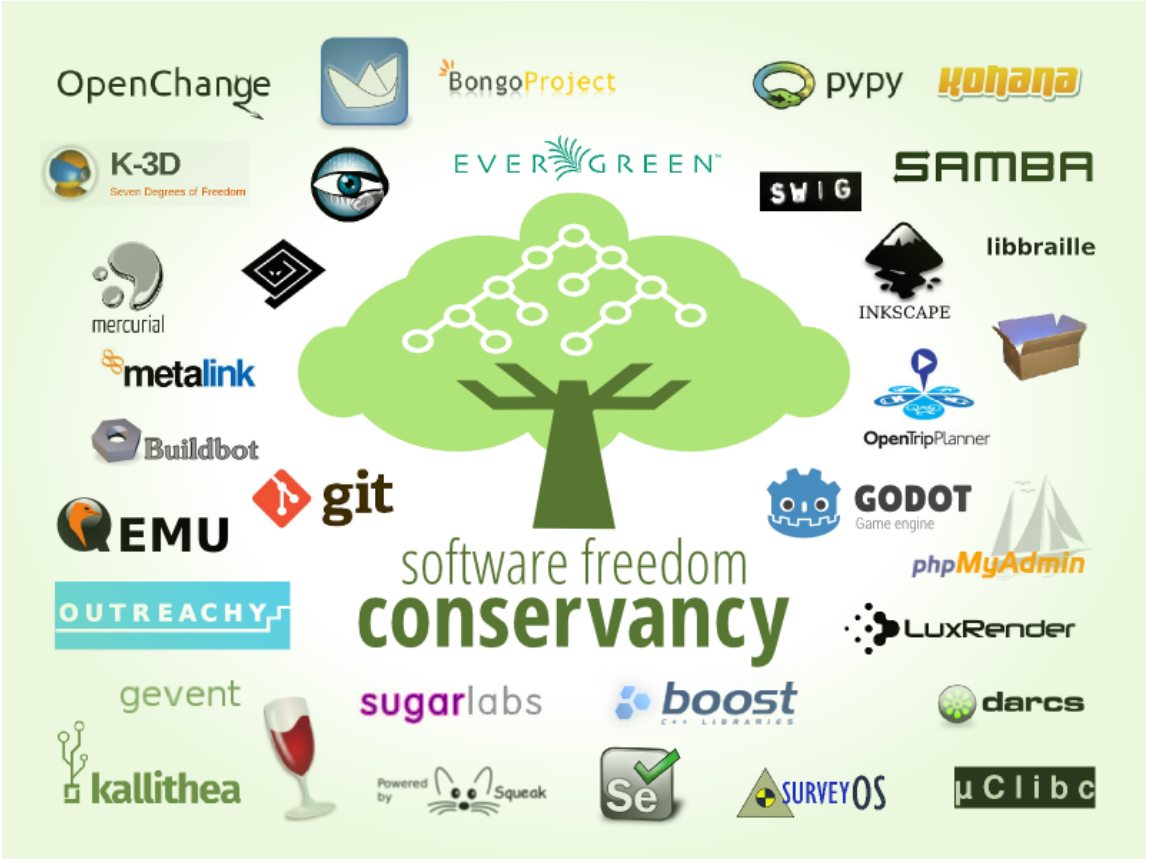
MIT Kerberos

- Blocking Samba being a part of SLES and RHEL
- Still in progress
- Very important as Heimdal Upstream only just restarted releases
- I'm hoping to update the Heimdal copy as well
  - 5 year old security code is not a great thing

OpenLDAP backend

- The original ‘make samba faster’ proposal
- Sadly little progress other than a presentations in 2015
- No public code
- I’m hesitant about another lift-and-shift like MIT Kerberos
- Prefer to fix one identified, isolated issue at a time
- Incremental progress can pay off now

Become an OFFICIAL CONSERVANCY SUPPORTER!



Catalyst's Open Source Technologies – Questions?

Want to join Catalyst? We love Linux-passionate Sysadmins and our Samba dev workload is growing: talk to me in the hallway track!