Experience running a clustered Samba gateway for CERNBox

Giuseppe Lo Presti, Aritz Brosa Iartza
CERN, IT Dep.
The Large Hadron Collider and its friends
Data Distribution in the Grid

- Global transfer rates regularly exceeding **60 GB/s**

- **830 PB** and **1.1B files** transferred until end of LHC Run 2 (2010-2018)

- Main **challenge** is to have the **useful data close** to available computing resources
  => match storage/compute/network

Running jobs: 365644
Active CPU cores: 807139
Transfer rate: 21.54 GiB/sec
Storage solutions for the HEP Community

- 7 years of dev & ops
  - 5K+ monthly active users, 37K users in total
  - 6PB+ data, 1.7B+ files, 110K+ shares
- Sync&share + online access
- Consolidating “home dirs” into CERNBox
  - Samba gateways instrumental to support Windows users
- Central Hub for CERN Data and Apps

https://cernbox.web.cern.ch

Powered by EOS
- Open Source in-house storage solution
- 10+ years of dev & ops
- Serving the LHC storage and throughput requirements
  - 100s of PBs
  - Disk & Tape

https://eos.web.cern.ch
Samba for CERNBox

- Cluster of 4 nodes, ctdb-driven setup
  - 192 GB RAM, 25 Gbps NIC
  - **Samba 4.11.16** on CentOS 8.3
  - A small /cephfs mount is used to share the state

- Distributed Storage (EOS) is FUSE-mounted
  - Multiple separated *instances*, all exposed via `\cernbox-smb\eos\...`

- Windows Domain (AD) joined in dedicated keytab mode
  - Authc performed by winbind, Authz performed by EOS

- File locking supported across all gateways
  - A must to support Office concurrent usage notifications
  - *Credits to the Samba community for the suggested solution*
Timeline

- **Production service as of September 2019**
  - Samba 4.8 on CentOS 7, 4 nodes with 64 GB RAM
  - Windows Terminal Servers configured to use it for *roaming profiles*

- **Usage growth in Q2 2020**
  - Upgraded to 4.10, then reverted because of too much pressure on our underlying FUSE-mounted storage

- **New cluster commissioned in October 2020**
  - Samba 4.11, improved EOS FUSE implementation, *very stable service*
  - Compiled in-house with a custom Gitlab CI (*latest 4.11 releases not available upstream*)

- **“Coming Soon”**
  - Upgrade to Samba 4.13.*latest* + deployment of a VFS module to support *RichACL*-based permissions
  - Started looking at Samba 4.14, *but*… compilation breaks on CentOS Stream because of dependencies
Monitoring and Alerting

• Based on a custom probe pushing data to InfluxDB/Grafana
  • Continuous parsing Samba syslog-formatted logs + actively testing backend

Long run trends in terms of FUSE latency can be analyzed.

Faulty behavior detected in FUSE logs can be alerted to administrators.
Usage Evolution

- Close to peaks of 3K connections, average has doubled compared to ~ a year ago

- Significant usage also in terms of I/O ops sustained by FUSE
  - Rates of ~10 kHz seen on a regular basis
  - Windows clients often insist on some specific files!
Steering the Development

- A Samba over FUSE stack is extremely latency-sensitive
  - Substantial efforts invested in our storage to address latency
  - Lots of tracing (strace, wireshark) analysis to identify bottlenecks

Time to stat each SMB mount from a Windows Server, daily
Coming next

- **Further usage growth** expected ahead:
  about to migrate more Windows-based use cases, in particular
concerning *shared project areas used by engineering applications*

- Possibly need to isolate the most demanding use cases
  in a separate storage backend, optimized for low latency operations
Conclusions: the bigger picture

- Multiple access paths . . .
  - Windows Desktop/Documents/… system folders
    - Either synchronized, or mapped to Samba
  - Samba became a first-class citizen among the available access methods to CERNBox
    - Significant usage, critical service for many workflows in the user community
  - . . . aiming at a coherent cross-platform UX
Thanks for your attention! Questions?