Clustered Samba Scalability Improvements

SNIA SDC EMEA 2020
Tel Aviv

Volker Lendecke

Samba Team / SerNet

2020-02-04

Samba architecture

- For every client Samba forks a new process
- Distinct memory spaces in every process
- MS-SMB2 and MS-FSA suggest a lot of shared tables
  - Lists of clients, tree connects, open files
- Samba can't use any of those data structures directly
- Samba shares data structures via shared key/value stores
  - TDB is a memory-mapped hash table
  - Protection via fcntl locks or shared mutexes
- TDB provides a clean separation layer
  - This made clustering initially possible
  - Process separation extended to nodes

SMB history

- SMB semantics date back to DOS single-user OS
  - Every application by definition had exclusive file access
- SHARE.EXE maintained illusion by blocking concurrent access
- Network-aware applications could explicitly permit sharing
  - Different modes of access permitted on a per-open basis
- Posix opens only have to read metadata
  - Permissions, file location etc
- Inherent scalability problem through share modes
  - SMB opens need to examine all other opens

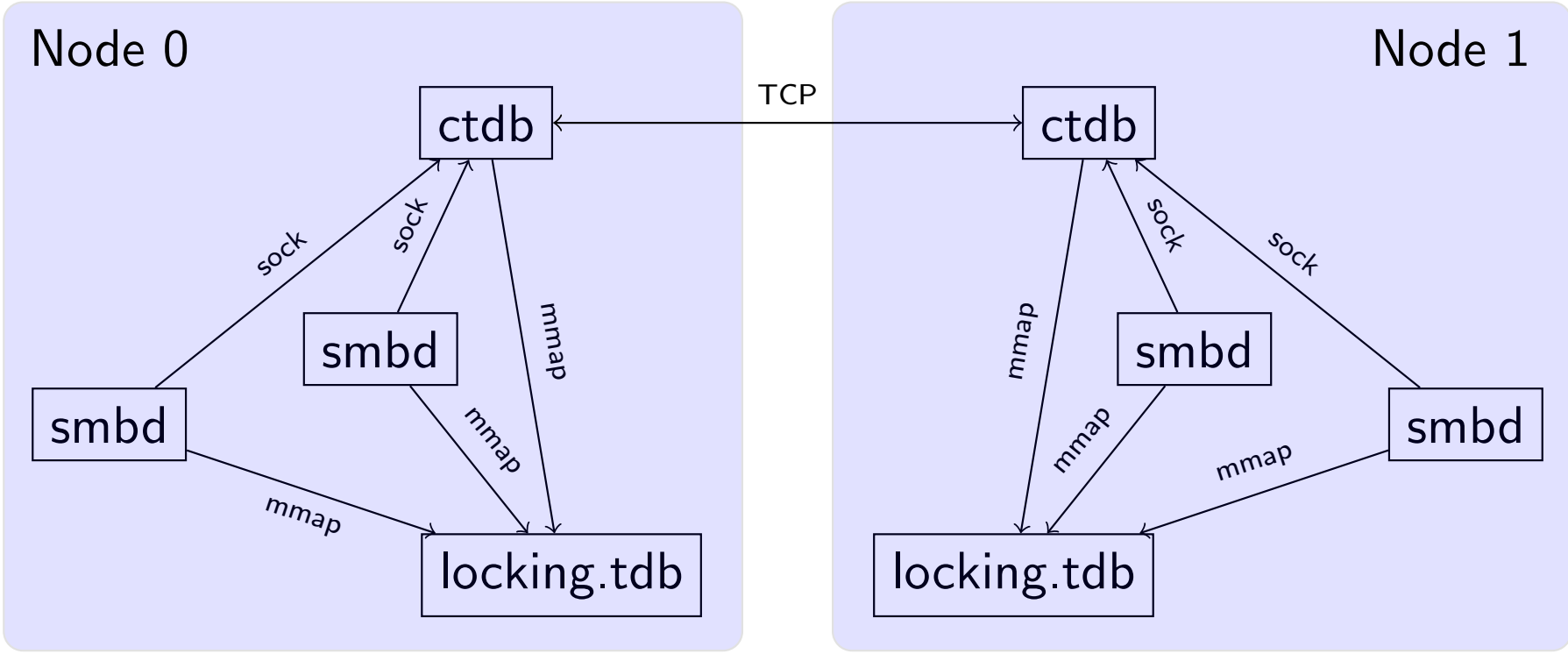
SMB share modes

- Every open call requests access permissions
  - READ, WRITE or DELETE (among others)
- Every open call allows other permissions
  - Concurrent READ, WRITE or DELETE permitted
- First come, first serve
- Samba stores an array of sharing information per inode in locking.tdb
  - Marshalling that array used to be very costly
- Restructure data structures to eliminate array NDR marshalling

Clustered TDB ctdb

- ctdb extends tdb files beyond a single machine
- ctdbd is a daemon to move records around
  - smbd requesting a record gets a local copy
  - ctdb maintains the most recent record location
- locking.tdb can be lossy
  - Share mode state valid only for open file handles
  - A crashed node's file handles are closed by definition
- ctdb record access is like NUMA with extreme node distance

ctdb Architecture



Cleanup in share_mode_data

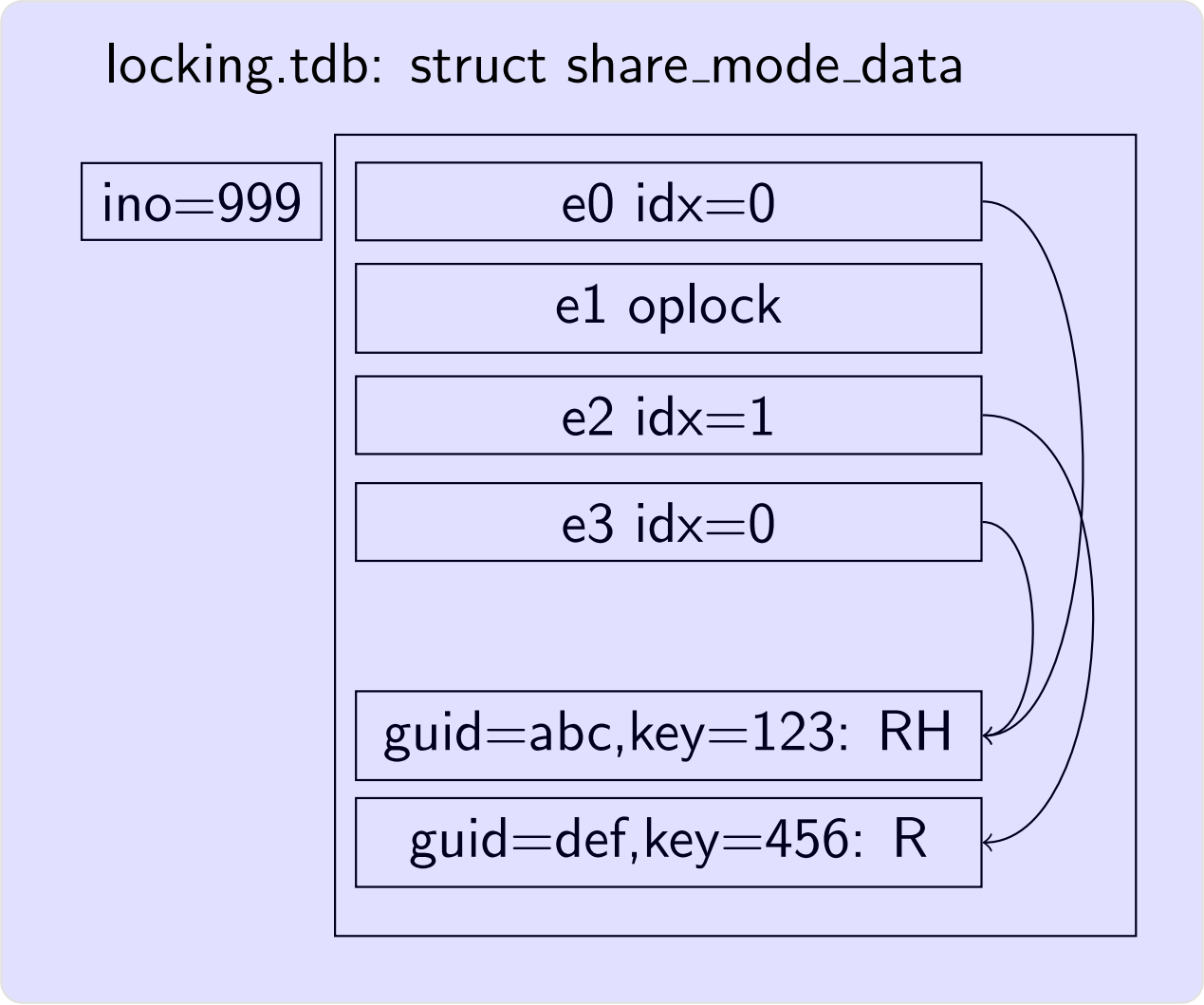
- One share_mode_data record in locking.tdb per inode

```
share_mode_entry share_modes[];
share_mode_lease leases[];
```

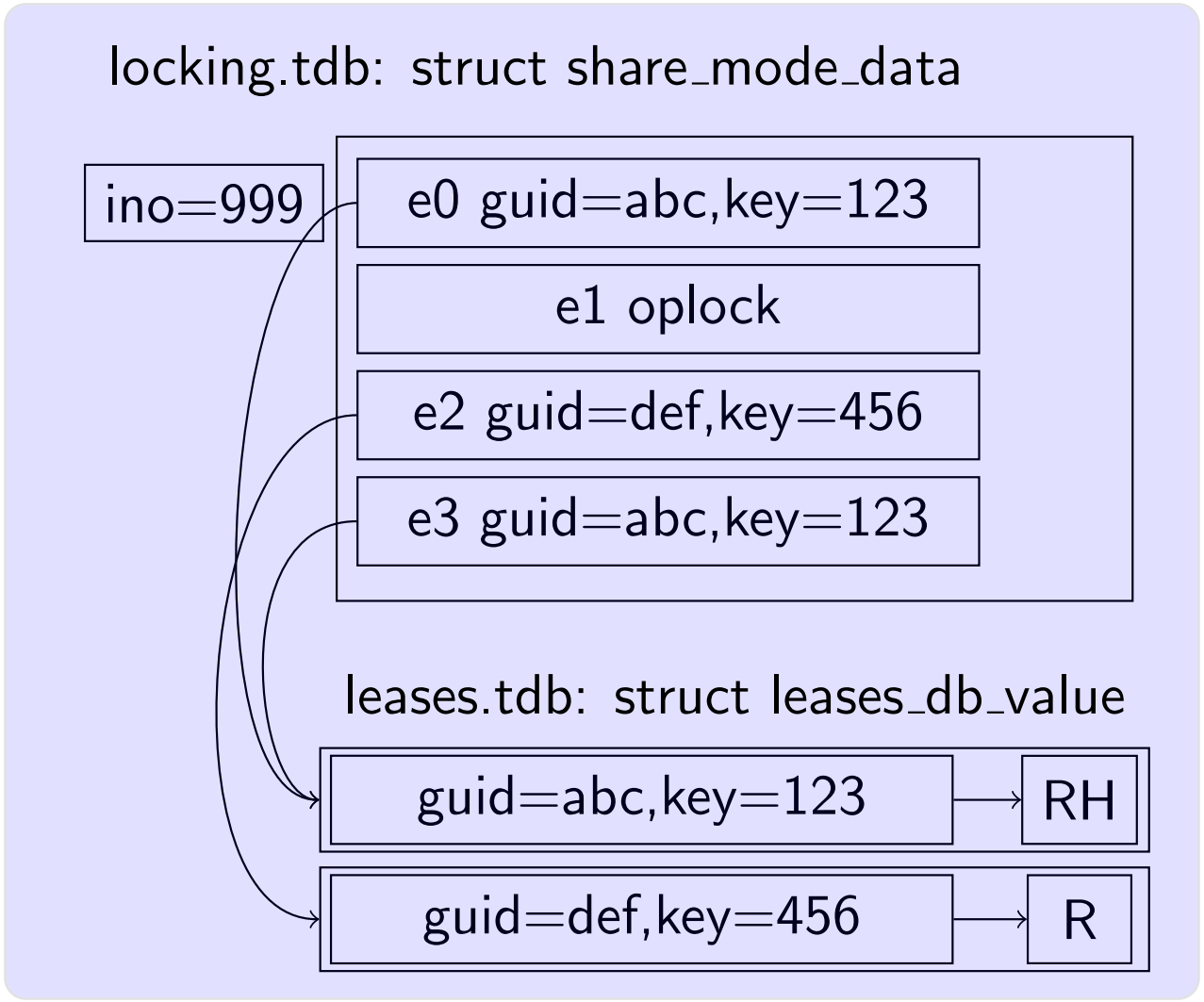
- Every share_mode_entry represents an open handle on a file
- share_mode_entry→lease_idx references the lease array
- struct share_mode_lease:

```
GUID client_guid;
smb2_lease_key lease_key;
smb2_lease_state current_state;
```

4.10 locking.tdb



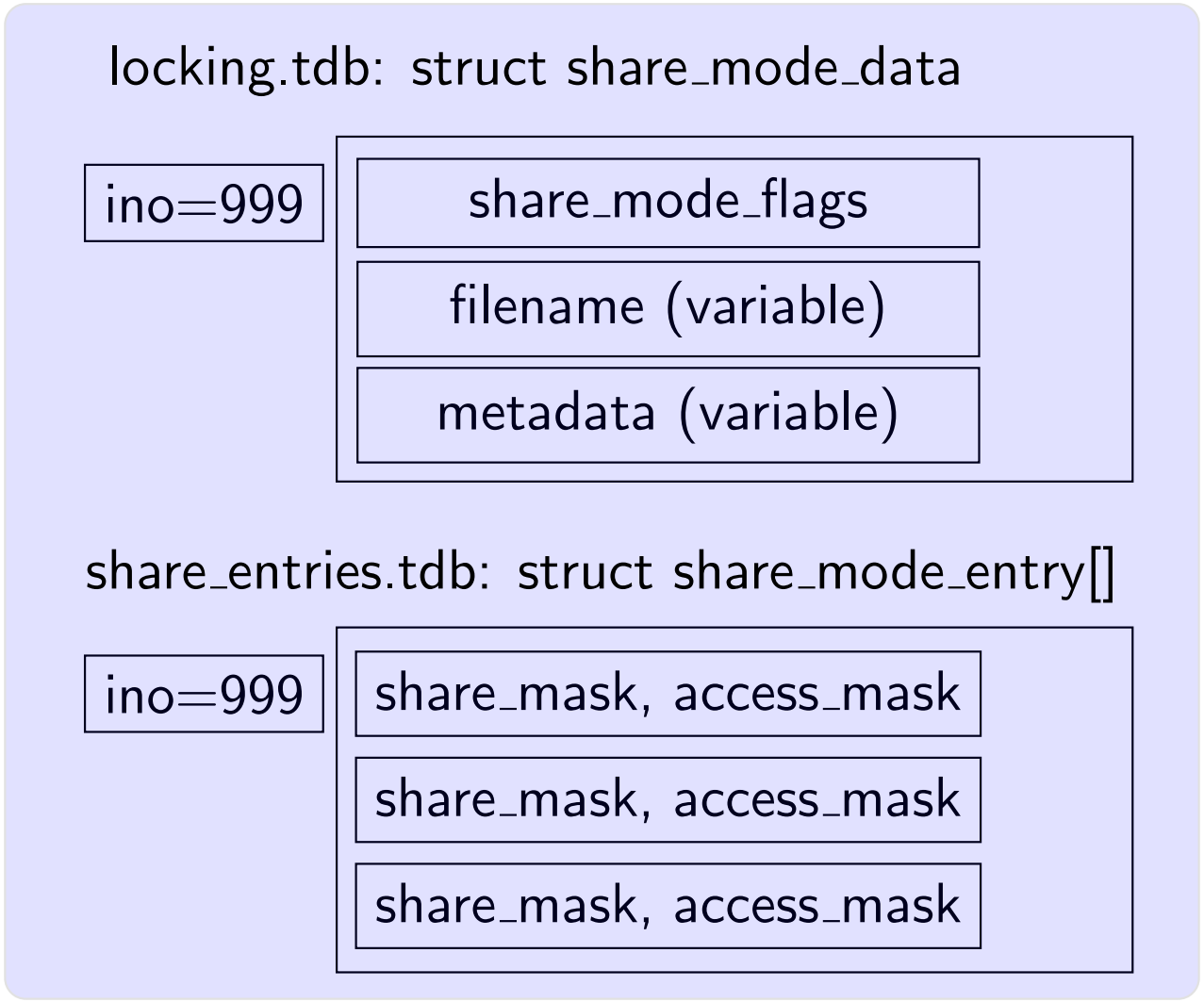
4.11 locking.tdb



Separating the share_modes array

- Roughly 30 places in Samba 4.11 reference share_mode_data→share_modes[]
- Most of the references walk the array
  - Lease breaks, durable file handling, file rename notification, etc
- Introduce share_mode_forall_entries() with a callback
- Introduce share_entries.tdb with sorted share_mode_entry arrays
  - share_mode_entry is fixed size, no variable components
  - Finding a record with binary search
  - Closing a file down from $O(N)$ to $O(log(N))$
  - Opening still $O(N)$

4.12 locking.tdb



Avoid walking the share mode array

- Share mode conflict:
  - I want to write, but someone else did not grant FILE_SHARE_WRITE
  - I don't grant FILE_SHARE_WRITE, but someone already writes
  - Same for READ and DELETE
  - First come, first serve
- Byte range locking cleanup introduced a 1-bit flags field
  - SHARE_MODE_HAS_READ_LEASE
- Extend that field to hold most restrictive share mode
  - Intersection of all share modes granted
  - Union of all granted access
- Opening a file just checks the per-file summary
- If there's a conflict, recalculate the truth

Demo Time

DEMO

Next steps

- Move share_entries.tdb back into locking.tdb
  - Non-contended file access got slower (3 instead of 2 records)
  - Now that the logic works, we can optimize data structures
- Base locking.tdb on g_lock.tdb technology
  - Avoid tdb locks while doing open/close/unlink/rename etc
  - Improve parallelism, reduce contention
  - Enable ctdb recovery while cluster file system is stuck
- Spread locking.tdb across per-node per-inode records
  - Parallel case (no share mode conflicts) only looks at one record
  - Conflicting case must take all records into account

Questions?

vl@samba.org / vl@sernet.de
http://www.sambaxp.org/