NFS Client Failover

Rob Thurlow

Sun

Blame The Automounters!

- Multiple choice feature in maps makes people expect real failover
- Accesses while mounted hang
- Heavily used data sets are paradoxically less available
- Need an alternate persistence choice - soft, hard, failover?

Solution Assumptions

- Don’t want to touch protocol!
- Synergy with future protocol changes (e.g. consistency)
- Prefer client-only changes
- New server daemon may be OK
- Need interoperable solution!

AFS/DFS Behaviour

- AFS volumes / DFS filesets
  - Designed for replication
  - Need Episode functionality or "captive" FS (yuck)
- FID same across volumes
  - Extra level of indirection helps greatly
- Replicas really look the same
- Replicas easy to schlep
  - Backup, dump, restore, move, duplicate, update
- Admin effort scales well

NFS - What’s Possible?

- There seem to be three possibilities:
- 1) Client uses automounter data
- 2) Client uses replication database
- 3) Servers present same filehandle for all replicas
- Also some variation on these themes

Sun

Client uses automounter data

- Pass replica list into kernel via autofs mount info
- Store pathname relative to VFS root
- Repeat partial pathname lookup with new server for failover
- Stored pathname introduces greater vulnerability to replica differences

Client uses replication database

- Central filehandle database maps endpoint/fh tuple to other tuples
- Client consults service to find other sites for file, polls servers to choose
- Client more tolerant of renames
- Is new database just another point of failure?

Servers present coherent filehandle

- Server agree to cooperate when replica is first created
- Servers all return same abstract filehandle to clients
  - Mapping from filehandle to dev/inode has extra step
- Clients still need to know or find list of replica sites
- Abstract filehandle has something that looks a lot like a volume ID!

Next steps

- Replication tools
  - Would like tools to create, update & verify replicas
  - Also need any database maintenance functionality
  - Would like something (anything!) better than "rdist"
- Consistency protocol
  - Will we ever have one for NFS?
  - If so, replication should fit right in

What do you think?

e-mail thurlow@eng.sun.com

Sun