Linux and Samba

Andrew Tridgell
Samba Team

Semantic Mapping

- Providing CIFS file services on Linux is an exercise in “semantic mapping”. The detail of mapping that is needed depends on the role the server needs to play
    - most detailed as a NAS box
    - dual-mapping for multi-protocol server
- A good example of the semantic mapping problem is the CIFS equivalent of open(), called NTCreateX().
    - takes 11 parameters and returns 14

CIFS meta-data

- File meta-data in CIFS is more complex than in POSIX
    - 4 settable times (POSIX has “2 and a half” time fields)
    - DOS attributes, ACLs and SIDs
    - separate allocation size
    - 8.3 names
    - file IDs
    - alternate data streams
- Unfortunately applications do end up relying on all these bits of meta-data
    - the perils of a software monoculture

Some bits already done

- Some bits of CIFS semantics have already been added to Linux for the 2.4.0 kernel and above
  - oplocks
  - simple share modes
  - directory notify
- These have helped a lot for Samba, but some have caused maintainence headaches for the kernel
  - How to integrate future CIFS features with less headaches?

Case-Insensitivity

- CIFS needs to be able to export a case insensitive view of a filesystem. The problem is doing this efficiently.
    - very contentious issue
    - problems with charsets
    - NT is not quite UTF-16
    - kernel maintainers have proposed a possible solution
        - smbd to kernel dcache coherence mechanism
    - log(N) lookup important?

Locking

- File byte range locking is rarely used in POSIX
  - works badly, so programmers avoid it
  - few users of it, so not priority to fix it
- CIFS needs more sophisticated byte range locking
  - true 64 bit (not 63 bit or 31 bit)
  - no brain-dead “close loses locks on other fds” features
  - mandatory locking (needs hook in read/write path)
  - lock stacking
- Just solve in user space?
  - works, but not good for multi-protocol file servers

File access control

- CIFS users expect full NT ACLs
    - impossible to correctly map to POSIX ACLs
    - needs SIDs for task security context
- Solve via LSM module?
    - Samba LSM module
    - NT ACLs and other attributes in an EA
    - has sufficient hooks for share modes and locking as well?

Sendfile

- Sendfile seems like an obvious fit for Samba, but there are potential problems
  - header sent first, what to do when sendfile returns short?
    - maybe doesn't matter as NT gets it wrong too
  - what happens with WinXP SP2 and mandatory packet signing?

Async IO

- Samba4 is designed around asynchronous operation, whereas Samba3 is very synchronous in nature
    - How do we do async filesystem requests, like open(), rename() etc?
    - do we have to use pthreads? What about pthread performance overheads
        - see thread_perf.c benchmark - doesn't look good
    - can we use direct clone() wrappers, bypassing glibc?

EAs and ACLs

- Samba4 will make extensive use of EAs and ACLs, for a closer semantic match to CIFS
    - use EAs for alternate data streams?
    - EAs limited to 64k. What to do about large streams?
    - do we really have to read all or nothing? nasty.
    - what about performance?
        - nobody benchmarks filesystems with ACLs and EAs!
        - some filesystems don't journal inode operations involving EAs
- Can we hook all this into LSM sanely?
    - Looks like we can

Alternate Data Streams

- NTFS and CIFS have “file streams”
    - arbitrarily named additional streams of data in files
    - mostly used for meta-data now, like who wrote it
    - WinXP SP2 uses them for “security zone” information
        - this makes streams much more urgent
- How should we store them?
    - In EAs?
    - in dot-files or a dot-directory?
    - what about large streams?

Content Indexing

- A core part of longhorn and WinFS
    - Maybe can be summarised as “open by content”
    - real-time indexing essential
    - can be very quickly deployed by Microsoft
        - Win2k implementation uses periodic indexing
- We need this in Linux!
    - Users could quickly become addicted to it
    - must be supported on network drives
    - uses a strange pipe format