MogileFS Summit

2006-09-19

Format

- Who's Who?
- Overview of MogileFS
- What's new in 2.0
- Wishlist / Future
- Mounting Mogile
- Plugins
- …..
- (whatever)

Who’s Who?

- ?

Overview: History

- History
  - Storage company sales people intolerable
    - Milking little guys for all we're worth
    - (asking us our yearly revenues before quote)
  - How hard can it be?
    - First version working after a weekend
    - Not hard.
  - Needed it for photo storage site

Overview: Parts

- Clients
- Trackers
  - mogilefsd
  - All queries / maintenance
- Storage Nodes
  - mogstored (and/or lighttpd/apache)
  - Just HTTP/DAV server w/ disks.
- Database
  - hopefully HA (more than 1, acts as 1)

New in 2.0

- all backwards compatible
- code cleanup
  - OO, file per class, ...
  - well-defined protocol between all processes
  - Start of test suite
- faster. no arbitrary sleeps.
  - Slowly adds sleep, otherwise stays busy

New in 2.0

- much more robust to error conditions during copying
  - detects difference between replicating failing reading from src vs dst.
- watch dog for processes
- configurable replication policies. removed from the mechanics.
  - Geographic,
  - Even/odd,
  - Whatever you want.
- can MKCOL now, so works w/ any DAV server (apache and lighttpd tested)

New in 2.0

- Fsck job (partially done)
  - Double-check all replication decisions / locations
  - 3-4 levels of checking
    - Db location only
    - Stat (HEAD request)
    - Get contents, compare
- Replication now uses new data structure (fids_to_replicate)
  - No longer queries files's devcount
  - Means replication plugins can say:
    - "Put it here for now, but I'm not happy about it (not ideal)… revisit layout later."

Future

- Coming
- Wishlist
- ….?

Future

- Mogstored --http=lighttpd, --http=apache, etc
  - Just do iostat and df interface (usage.txt files)
  - Smartd interface?
- Mogstored improvements (aio isolation)
- UUIDs, automounting in core (LVM integration?)
- Move LJ's out-of-band device reweighting using iostat into the core, so everybody gets it
  - Use it in more places
  - Add rebalance job, to move files around?

DB improvements

- Partitioned DB
  - Necessary? So small.
- Docs on MySQL HA
- Or, mogilesd supporting master-master
  - Add DLM (ddlockd?) to mogilesd
    - Use for lock on namespace changess

Parallel meta files

- 0/00/000/00000001.{fid,meta}
- for fsck if lose db
- Add checksumming info in there
- Large file support (auto chunking?)

HTTP in mogilefsd

- Add DAV server to mogilefsd
  - PUT / DELETE / GET (w/ redir or not)
  - Use any HTTP client / command line
  - Makes for easier mounting
    - Google “wdfs fuse”