CohortFS

Restart/Continue pNFS Metastripe '89

CohortFS

Metastripe: pNFS for Metadata

- Adds cooperating MDS servers, scale-out metadata, and parallel metadata operations to NFSv4, using new pNFS metadata layout type(s)
- Adapts and extends the proposal by Eisler, in 2010
- First presented at NFSv4 WG IETF 86, Orlando

Recap: changes from prior drafts

- Reduced layout state overhead
  - Avoid requirement to take metastripe layouts on regular files
    - Propose *stripe hints* (attribute)
      - Coarse-grained layout (file-system)
      - Deviceid hint (attribute)

CohortFS

Filehandle Striping (new name!)

- Filesystem “singleton” layout
  - Holds device list and stripe pattern
  - Permits clients to request stripe hint
    - Recommended attribute
  - Used for:
    - Open, GETATTR, LOCK

CohortFS

Directory Striping (new name!)

- All other metastripe (layouts on directories)
    - Essentially just like original metastripe
    - Used for:
        - Name-based operations (CREATE)
        - Directory enumeration (in parallel)

Simplify metadata layout slightly

- Remove layout filehandles (try to)
- Simplified device model
  - Define one device structure and layout device presentation
    - always use it
  - Keep “offsets” opaque

CohortFS

Changes since 01

- Layout sub-type names changed
  - Structure and semantics unchanged
- Improved language in several areas
- New directory striping algorithm (CEPHFRAG) added (to appear in next draft)

New Items for discussion

- Layout subtyping (update)
- Opaque data in LAYOUTGET (still needed)
- New directory striping algorithm (CEPHFRAG)

Layout sub-types

- Currently, there is one new layout type, LAYOUT4_METADATA, with filehandle striping and directory striping sub-types
- LAYOUTGET iomode argument overload to specify a desired layout subtype
- At IETF 88, we discussed splitting out filehandle striping and directory striping layouts, to avoid overloading
  - (Because LAYOUTGET lacks layout-specific data)
  - No one liked this

MDN_ALG_CEPHFRAG

- The CEPHFRAG algorithm describes the Ceph algorithm for placing new directory entries on “fragments”
  - A striping algorithm based on recursive hashing and splitting
  - Shows generality of the mechanism, frag trees are typed seed data already provided for
- The next metastripe draft introduces the new code points and description. We plan to push these changes when draft submission re-opens.

F CohortFS tion

- Ganesha
  - Provisionally complete, WIP source available
    - https://github.com/linuxbox2/nfs-ganesha (metastripe)
- PyNFS
  - Nearly complete set of initial tests, WIP source available
    - https://github.com/linuxbox2/pynfs (master)
      - Soon!

Current draft

http://tools.ietf.org/html/draft-mbenjamin-nfsv4-pnfs-metastripe-02

Next draft

https://github.com/linuxbox2/metastripe

CohortFS

Q/A