Flex Files:
A New Layout Type

Tom Haynes
loghyr@primarydata.com

What is a layout?

- What data protocol is being used?
  - NFSv4.1
  - Blocks
  - Objects
- How does the client reach the storage devices?
  - LUN
  - Netaddr

What is Flex Files?

- A new layout type for pNFS
- Control protocol could be NFS
- Data protocol is NFS
  - v3
  - v4.x
- Client side mirroring

What is the metadata protocol?

- NFSv4.1
  - New layout types are allowed
- NFSv4.2
  - Provides ability to return stats and errors before LAYOUTRETURN

An old elevator pitch

- MDS is from Vendor A
- DS is from Vendor B
    - It only speaks NFSv3
    - Reuse storage investment

But what is it really?

- Data mobility
  - Move the data without touching the namespace
- Provide multiple copies of the file
  - Pick the local one for reading
  - Client controls updates
    - Every mirror has to be updated for a write to be valid

Coupling

- Tightly coupled - explicit protocol between MDS and DSes
  - Fencing
  - stateid
- Loosely coupled - shoehorn semantics into an existing protocol

Fencing

- MDS recalls the Layout
- Client does not respond
- MDS tells the DS to stop servicing the client via the control protocol
    - Flex Files might not have an explicit control protocol
        - MDS is Primary Data
        - DS is a stock RHEL 6.5 server

Synthetic uid/gid

- MDS provides client with synthetic ids
- uid is presented for writes
- gid is presented for reads
- Client is trusted to cache pages correctly
    - It does the access checking locally for different users

Example fencing

- MDS file

```
-rw-r--r--    1 loghyr  staff    1697 Dec  4 11:31 ompha.c
```

- DS file

```
-rw-r-----    1 19452   28418    1697 Dec  4 11:31 data_ompha.c
```

- Fenced off

```
-rw-r-----    1 1066    1067     1697 Dec  4 11:31 data_ompha.c
```

Cons

- Fencing occurs for all clients, not just the problematic one

Client-side Mirroring

- READs
  - Client picks the best mirror to get a copy
    - Server may hint
    - Client can override
- WRITES
  - Each WRITE has to succeed over all mirrors in the layout or the client reports an error
    - Returns the layout
    - Asks MDS for a new one

What does MDS do?

- Determine which DS(es) are out of sync
- Issue new LAYOUTs with only the good copies
- Resilver the bad copies
  - Must also get any modifications the clients are making
    - I.e., the clients have no clue about DSes not in the layout
- Add the copies back to the layout when resilvered

Where can you learn more?

- https://datatracker.ietf.org/doc/draft-ietf-nfsv4-flex-files/