Data preservation & SAM

Robert Illingworth

8 August 2012

SAM & data preservation

- Short term plan: keep existing systems running (until ~2015?)
- Long term needs more thought
  - What SAM data do we need to preserve?
  - What functionality do we need to preserve?

Metadata

- The metadata catalogue holds information about the data
  - Properties of files; lineage; processing history; etc
- Important for understanding what is in the data
- Currently 140 M entries for D0, 36 M for CDF

Location catalogue

- Stores location of all physical files in the system
  - Important for finding files (pnfs location)
- In principle, can be reconstructed from enstore, but a slow task
- Seems like something we want to retain, and keep up to date
  - Enstore migration does not affect access paths; only a concern if data moves to some other system, but preferable to update tape labels

SAM at D0

- Heavy user of metadata & SAM dataset definitions
- Use full cache management features of SAM station for file deliveries

SAM at CDF

- Less use of metadata and SAM dataset definitions (tagged CDF datasets instead)
- Uses dCache for file caching on the CAF rather than SAM cache – more limited use of SAM station features

Option 1

- Keep full system running
  - All existing clients still work
  - Depends on long term viability of software products like omniORB (CORBA middleware)
  - Significant support load (need to keep old services functioning)

Option 2

- Keep existing metadata and file location access, but turn off SAM stations
  - If someone needs to access the data, can create file list from DB, but requires alternative access method (direct dCache or Encp?)
    - More impact on D0 than CDF
  - Still dependent on maintaining significant parts of the existing codebase
  - Slightly less support load than option 1

Option 3

- Turn all SAM services off, but keep databases
  - Zero effort required
  - Unless somebody wants to use it
    - Would need to write DB queries to retrieve metadata and handle all file access themselves

Option 4

- Convert to using IF SAM
  - New HTTP based interface in development
    - But not compatible with existing client code
    - Existing scripts, sam_manager/diskcache_i, dataset definitions, are not compatible
  - Low support level required, as using same software as IF experiments
    - Provided IF implementation does not diverge too much
    - And that we don't decide to migrate to some other DH system

Option 5

- Turn system off, but be prepared to reactivate it if needed
  - Zero effort required if no usage
  - Might turn out to be a lot of work to reactivate
    - Even if documented, knowledge gets lost quickly

Final thoughts

- We want to keep as much in common of the SAM infrastructure between CDF and D0 as we can
- Many of the existing servers can be moved from current hardware to VMs
  - Won't need dedicated hardware
- SAM Database requires Oracle
  - Unlikely to be an issue as used by lab business services