Data Management TEG Status

Dirk Duellmann & Brian Bockelman

WLCG GDB, 9. Nov 2011

DM TEG status

- The working group has the stakeholder representation we need. The only apparent omission from the first meeting is a representative from dCache and StoRM, but we will contact both projects explicitly. We are also already in contact with the Storage TEG to achieve an additional balance between experiments and infrastructure providers.
- The mandates are clear and agreed. Some members have expressed concern about the efficacy of such working groups; does the MB have any guidance how the resulting strategy documents will be used after the TEG's have disbanded?
- The working group has had its initial kickoff phone meeting. We will be holding biweekly meetings and a face-to-face meeting (collocated with the storage TEG) in late January (possibly also one in early December). We plan to focus initially on the needs and status of each experiment, then move into the examining the roadmaps of the infrastructure providers.
- Other issues:
  - We would like guidance on how to best coordinate between different TEGs. There's significant overlap with storage management, and some amount of overlap with security and operations.
  - We believe it could be difficult to have a fully-formed long-term strategy document in January. Our current goal is to put together a strong statement of the status quo by the end of the face-to-face meeting.

TEG Logistics

- Twiki page
  - https://twiki.cern.ch/twiki/bin/view/LCG/WLCGTEGDataManagement
- Email list
  - mailto:wlcg-teg-data-management@cern.ch
- Bi-weekly phone meetings
  - https://indico.cern.ch/categoryDisplay.py?categId=3772
- In person meetings being planned close to Storage TEG

The Story so far..

- We do have a working system
  - We do not design a new s/w system
  - It’s cost and complexity is exceeding what is strictly needed to do the job
- We should rather critically look at which elements of the current system are used and how? Which have been added and why?
  - Which (larger) functionality did not make it and therefore could be declared obsolete (time scale 2-5 years)?
- What is the impact on the experiment / site / dev project side to progressively drop complexity?

- This review needs a simple layered model of data management as we do it today and a list of boxes labeled with “crucial”, “not required”
- And one/few future evolutions of this model which we see as desirable target after the 2-5 years

The E stands for Evolution

- We should not start by collecting bug reports or feature requests for existing systems
- We should rather step back and look at main concepts behind recent strategy changes/additions on the experiment side?
  - Eg Data placement and Federation
- Which concepts allow to implement Storage Federation?
  - What are the implicit assumptions: eg namespace congruence, read-only files
  - Upcoming Federation workshop in Lyon will help TEG to collect input
- How and where can new technology come in and which of our “special needs” is preventing this today?
  - Clustered file systems at smaller sites
    - but we need SRM for disk?
  - Cloud storage
    - but we need a global namespace?

Requirements vs Strategy

- Just collecting a list of additional requirements will not be sufficient
- Need to document a few strategic target scenarios
  - main advantages in sustainability
  - allow storage providers to align / point out risks
  - define crucial validation tests
- Possible scenarios
  - POSIX: storage pools look like local file storage with publishing / deletion
  - Cloud storage: storage has constrained semantic a la S3

DM Model and Evolution

- A few key questions to guide input collection
  - Is the archive<->disk split an agreed strategy?
    - In other words: can we drop HSM?
    - If yes, several simplifications can be done and benefits can be evaluated. Eg Why SRM for disk-only pools?
  - Do we see “naked” clustered file systems or cloud storage coming in during the reviewed time scale?
    - Can we take our abstract DM model and fill its conceptual boxes with technology options for a give class of sites?
    - Are the experiment data services prepared for this?
  - What role do we see for Federation vs Data Push?
    - Eg Federation is complementing Push at small sites (T>1) but Push is still required to maintain control
    - Eg Federation/Push is far better and will be the strategic target

Summary

- DM TEG has (just) started
  - First meeting was mainly on logistics and scope
  - Mandate and environment for it are OK
- We should start by abstracting from what we learned and come up with DM model(s) and their expected evolution
  - .. rather than repeating sometimes painful experience of low level functionality definition groups
- Process will start driven by the experiment view and progressive include storage provider and site feedback
  - In close contact with other TEGs
  - Regular MB and GDB reports