Data and Storage GridPP5 Ideas

Wahid Bhimji

General statements

- David said to present suggestions and also FTE estimates. Some discussions in storage group (and Fri technical meeting) - otherwise my opinion, FTE not current but future minimums.
- We manage (well) but experiment data activities are still very inefficient in many areas and challenges will grow.
- Measuring effort across GridPP not just dedicated posts:
  - Could be done in spare time by site admins but if it takes up the bulk of their time then essentially the same
- A list of current gridPP storage activities:
  - https://www.gridpp.ac.uk/wiki/StorageGroupActivities

More general statements

- 5 years: technology exists now: but specific predictions always go wrong (see some from Gridpp4 in backup)
- “Big-data” is much bigger than b4: we can big-up our activity
- Maybe “common” technologies or “industry” solutions (clouds ; hadoop appliances) will rise. Whatever way, case for moving our activity “up” to making use of these technologies
  - Operations
  - Thin-layer, future-looking, middleware
  - Future-looking data access interfaces
  - Interoperation with LHC and wider “big-data” communities

“Operations”

- Not always glamorous but responsive operations activity is mainly what we do
  - Installing, configuring storage (hardware and middleware): 0.25
  - Experiment demands (dark data, space manage etc.) : 0.75 FTE
- We have an excellent reputation for responsiveness (I think)
- This activity could increase (but balance is probably right)
- LHC to non-LHC VO balance is never quite clear
- Spread over many people - (and probably thus grossly underestimated) but anyway it needs to be done

Middleware

- A HEP-specific storage layer of some sort will continue to be needed (in next 5 years) for HEP use of distributed storage resources
  - Some requirements may be relaxed: e.g. SRM
  - Some not: accounting; authentication (and data access peculiarities: WAN access, random access etc.)
- Interoperation with “common” (modern) technologies is something we should support
  - So long as we can deliver what HEP experiments need

Middleware: practicalities

- DPM/DMLite is evolving in "future-looking" direction.
  - Through a growing and successful "collaboration" we are
    - integrated into "dev" team - influencing and contributing.
  - Small fractions of FTE could support some roles - e.g. tools within DMLite; docs; support rota; testbeds
  - However someone with some dedicated time needed to keep our influential role (and CERN has put strong and ongoing support on the back of community support)

0.75 FTE overall

- Storm and dCache we play unofficial role in testing / pushing dev.

0.25 FTE

  - Storm future uncertain - but maintaining our options in all directions important.

Future-looking interfaces

- Xrootd ; WebDav; Cloud; Storage Federation etc.
- Some of this is “now” (and growing) - some will come
- One could lump in to this section our work on metadata management (gFal2 etc.)
- And our work on “data processing” - ie. making good use of interfaces
- Dedicated people on these technologies not bad idea
  - but lets not pick a winner. ..... 0.5 FTE

Engagement with LHC and wider (e.g. 'Big data')

- I think we should have dedicated effort for a technical-level bridge with industry and other academic players.
  - Spread our expertise in networking, data transfer, storage, data processing etc etc.... generate "impact"
  - Learn useful things for us from these huge communities
  - 0.5 FTE again doesn't have to be one person -
  - Could be considered to subsume current activities of "technology tracking"; hardware industry interaction ;
- LHC "Engagement" (VOs, WLCG etc.) should be lead roles
- Publicity and publications - should be high-quality (e.g journals, mainstream computing conferences)

Networking

- Important: was going to get its whole own talk.
- Much of the gridpp4 tuning done by non-storage folk (most reported into the storage group)
  - Again the capacity to do that work has to exist somewhere
- Operations - FTS etc. needs some support
  - In addition to T1 support
- 0.5 FTE

Summary

- Data operations: exp support (0.75) and infrastructure support (0.25)
- Data middleware: DPM support and dev (0.75), Other middleware support (0.25)
- Data technology-transfer and engagement (0.75)
- Data interfaces and processing (0.5)
- Data transfer and networking (0.5)
- 3.75 FTE (operations or closely related activities are > 2.5)