1.0.4 - LIGO Applications

Kent Blackburn
(Robert Engel, Britta Daudert)

Einstein@Home on the OSG

- Continuing to see an average throughput of 50 to 100K CPU-Hrs per day. Discernable trend towards lower numbers over the past few weeks.
- Running on roughly 30 sites around the OSG
- This is a sweet-spot in terms of hands-on oversight and productivity.
  - Robert Engel reports that he has less time to devote to this with the increased demands from the Documentation project.
- Other than University of Wisconsin Milwaukee, these are strictly opportunistic jobs.
- Unclear how increased competition and LHC ramping up will impact these numbers, but for the time being they've been holding up well.
- Detailed graphs at http://t2.unl.edu/gratia/vo?vo=ligo
- Einstein@OSG has been showcased numerous times in various OSG articles over the past few months

E@OSG Future Thoughts

- Some interest in developing support for SRM SE storage as opposed to the local storage currently used
- Probably greater interest in migrating to using a standard "pilot" method to replace LIGO/GEO home grown job submission/management solution
- Errors encountered have been documented in a report by Robert and given to OSG Production for improved usability for all communities.
- Unclear how much effort is available in the near term to make modifications/enhancements to E@OSG, but it is not a lot.

Binary Inspiral Workflows

- Focusing on several areas in an effort to achieve scientific contribution from the OSG
  - Use of storage elements to pre-stage (TBs) of data used by workflows designed by LIGO scientist
    - This is legacy based on the LIGO Data Grid paradigm of all the meaningful data at all LDG sites all the time.
    - Development of efficient data staging scripts that also register data locations in SE in RLS catalog and check integrity of transferred data sets.
  - Changes to both LIGO data analysis codes and Pegasus to make the two grids appear transparent to the workflow submitters
    - This requires scientific collaboration backing to accept methodologies and code changes/reviews associated with modifying publication class production codes.
      - Process for this is slow and of lower priority to collaborators in LIGO, often resulting in LIGO Data Grid code changes occurring before sufficient time and review has transpired to vet OSG specific changes … leading team back to the "drawing board" to try again with new code bases
    - Development and tracking time needs to be greatly reduced and overall differences in the workflow's "flow" resulting from being on the LDG and OSG removed to keep pace
- Almost all effort has taken place at two sites: Caltech/LIGO ITB cluster and Firefly.
  - Looking at using other sites where much less storage is available than would be scientifically interesting to the binary inspiral analysis team so that greater understanding of the storage element model and challenges can be explored

Binary Inspiral Future Thoughts

- Performance of workflows on Firefly are seen to be an order of magnitude slower than seen on the LIGO Data Grid (and on the Caltech/LIGO ITB cluster). Unclear why this is the case. Britta has been investigating, hopefully with inputs from Firefly admins.

- Seems very clear to me and a few (many?) others that usability and transparency of storage solutions on the OSG are resulting in loss of traction for analysis codes that demand lots of data out of the gate.

- Pegasus has historically performed early binding of environmental factors to the workflows. Some discussion and effort underway to explore and develop later bindings, hopefully avoiding some of the issues with information services used in early binding falling out of data or upkeep.

Conclusions

- LIGO's E@OSG application is in science production contributing to the E@Home analysis. This is a huge analysis requiring enormous amounts of computation globally before conclusion at the level of a scientific publication are mature … its will happen!
- The need for large data sets available to the jobs in the binary inspiral workflows has been challenging based on the current usability of storage on the overall OSG "Grid".
  - Performance has yet to be demonstrated that matches well with the LDG for jobs ⇔ data throughput on the OSG