LIGO Extensions

OSG Staff Retreat 2010

Wednesday July 21st (3pm Central)

Kent Blackburn (Call-In)

The Team

- Kent Blackburn
  - LIGO Senior Scientist, OSG Council Co-chair
  - ~30% OSG
- Britta Daudert
  - Primary liaison between LSC Binary Inspiral Group|Pegasus|OSG Grid Technologies
  - Some involvement with LSC Periodic Group
  - Some support of OSG EOT
  - ~100% OSG LIGO Extensions
- Robert Engel
  - Primarily involved with Documentation
  - Oversees administration of Caltech ITB cluster
  - Provide operational oversight for Einstein@OSG production running
  - ~25% OSG LIGO Extensions, remaining devoted to sysadmin/documentation
- Misc. low level support from LIGO Lab and LIGO Scientific Collaboration members
  - System administration recommendations
  - Binary inspiral code evolution in support of OSG as a target grid
  - Typically less than a few hours a week
- USC-ISI Pegasus Developers
  - Evolution of Pegasus workflow planning and diagnostics to merge LIGO Data Grid / Open Science Grid into a common user viewed platform for running binary inspiral workflow
  - USC-ISI has an MOU with LIGO establishing sharing of resources in support of Pegasus development
  - ~ 25% FTE of support

The Resources

- Caltech is home to an OSG Integration Testbed Cluster
  - OSG Software Stack (CE & SE)
  - 104 cores
  - 25TB utilizing Hadoop & BeStMan SE
  - Primarily used for Binary Inspiral Workflow testing on OSG software stack
  - Four Virtual Machines provide services for job submission, CE and SE gateways (recently added OSG-MM)
- LIGO-MIT is establishing (renewed) OSG Production Site at new Bates Center
- LIGO Data Grid is unique from the OSG but leverages the VDT and has many similarities with the OSG

Focus Areas of Work 2010

- Sustained production running of Einstein@OSG
- Continued development of LIGO Binary Inspiral Workflows for the OSG
- Interface with Pegasus workflow developers to develop grid specific transparency between LDG and OSG for LSC users
- Understand capabilities and limitations of OSG storage elements in the context of the binary inspiral workflows
- Develop long-term sustainable OSG-MM plan for LIGO
- Deploy native packaging technologies

Einstein@OSG

- Migration to Condor-G resulted in significant up-scaling
  - Running in production mode on 31 sites in a typical week
- Significant effort has gone into tracing down site-by-site specific error conditions and customizing error handlers
  - This seems to have transitioned to very stable, robust production running with current OSG release software stack
    - Typically only requiring about 30 minutes per week to supervise
- A few sites still have unresolved challenges (ask Robert for details)
- Some competition for opportunistic cycles seen, not only across VO's but also internal to LIGO VO
  - Firefly has been a significant contributor to all LIGO apps, including the Einstein@OSG

OSG Contribution to E@OSG

- LIGOis the largest non-HEP VO using opportunistic resources on the OSG
- 5,000 jobs running simultaneously on 31 grid resources in average, peak at 9,000
- 100,000 cpu hours per day in average, peak usage at 200,000 cpu hours per day
- 1 to 2 million science credits per day in E@H database
- Currently listed as number 1 contributor in terms of recent average credits contributed to E@H in world
- This area is in very good shape and contributing to the overall science mission of LIGO

Binary Inspiral (CBC) Workflows

- Computationally challenging analysis for LIGO, “BUT” in scope
  - LIGO Data Grid resources sufficient to address primary milestones for the project
  - Some higher order statistics and corner cases could benefit from not taxing existing LIGO Data Grid schedules
- Workflows consist of tens of thousands of jobs with anticipation to reach hundreds of thousand of jobs in the near future
- Each job take order 1-10 minutes, but can be longer in corners of parameter space
- Each workflow requires large local storage of LIGO data, on the order of few - ~50 terabytes
- This is considered LIGO’s most likely detection analysis strategy – important!!
  - Modifications to the code are under strict control and highly regulated for verification on the LIGO Data Grid.
  - Support from the LIGO Scientific Collaboration to migrate off LDG is anticipated as requiring an almost doubling of effort due to the highly regulated nature and significance of this code base
- Best path for migration is to utilize workflow planning on both the LDG and the OSG for the same common analysis code base
  - Years of effort have gone into this using Pegasus Workflow Planner from USC-ISI
  - Some progress (within the last year) at standardizing code to use Pegasus for both LDG and OSG, but a few minor changes needed in LIGO’s code, vetting these changes takes months

Challenges to CBC in 2010

- In a nutshell, complexity of using opportunistic storage has been the greatest challenge
  - Not seen as being globally supported
- To date only two sites have been successfully utilized by the binary inspiral workflow
  - Caltech integration testbed site
  - Firefly production site
  - Both are Hadoop / BeStMan (?)
  - Pegasus development may improve this in the future
- The CBC code base is not static – there is a synchronization challenge with the branches of code used between the LDG and the OSG

Expectations for 2011

- LIGO is ending its current science run this fall (2010)
- The instrument will see multiple years of upgrading to provide a ten times improvement in sensitivity, resulting in a thousand time greater event rate.
- No new data will be available during this period, analysis activities will likely see significant tailing off until data collection begins a new
  - 2011 will likely see data analysis efforts continue at pre-decommissioning levels as paper are being pushed forward for publication on the current science run data sets.
- There is also likely to be an upgrading of the LIGO Data Grid close to the start of the new data taking

Recommended Focus for 2011

- Einstein@OSG:
  - Continue the current minimal effort required for large scale production runs
    - Periodic data analysis could use nearly infinite computing and will likely continue through decommissioning
  - Contribute to the effort to provide a standard test suite for the OSG
  - Migrate to pilot based job submissions that will eventually solve some of the current problems by utilizing a standard test suite (see above point)

Recommended Focus for 2011

- Binary Inspiral Workflow:
  - OSG-MM was very recently set up as a service at Caltech. More work to do to become familiar and understand the long term role of this technology as a placeholder in the OSG
  - Effort began just last week to explore glide-in technology for the binary inspiral workflows. This may solve some of the performance issues that have been faced when running these large workflows through job managers
  - A LIGO Storage Task Force met for the first time last week
    - Looking at how to make data transfers more effective
    - Exploring the challenges of opportunistic storage for large datasets
    - Should continue and expand beyond LIGO applications
    - Global solutions needed!
  - There is an interest in utilizing the OSG for "quick and dirty" looks at the data hot off the instruments. This will be of greater importance in the new (Advanced LIGO era) when event rates are 1000 times greater

Closing Remarks

- Some slightly more coherent than random efforts have gone into other types of LIGO data analysis.
  - Most recent of these have involved a different type of periodic search (sister to E@OSG) that requires large amounts of pre-staged data similar to the binary inspiral analysis
  - Other types of analyses have the potential to require significantly larger computing resources as the sophistication evolves.
- Grids (both LDG and OSG) have not been "plug and play" the way many have wanted
  - New technologies (Condor-C, Clouds, glide-ins) may help
- This is the business of data analysis … storage is essential!!!
- Likely to see open data policies for LIGO Data in the Advanced LIGO era! (public data needing access to "public" computing)