LIGO Data Grid

Patrick R Brady
University of Wisconsin-Milwaukee

What is the LIGO Data Grid?

- The combination of LSC computational and data storage resources with grid-computing middleware to create a distributed gravitational-wave data analysis facility.

- Compute centers at
  - » LIGO Hanford Observatory
  - » LIGO Livingston Observatory
  - » Tier-1: Caltech
  - » Tier-2: MIT, UWM & PSU
- Other clusters in Europe
  - » Birmingham, Cardiff and the Albert Einstein Institute (AEI)
- Grid Computing software
  - » e.g. Globus, GridFTP, and Condor and tools built from them

Why do we need the LIGO Data Grid

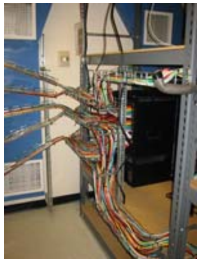
- Low latency analysis is needed if we want opportunity to provide alerts to astronomical community in the future
- Maximum scientific exploitation requires data analysis to proceed at same rate as data acquisition
- Computers required for flagship searches
  - » Stochastic = 1 unit (3 Ghz workstation day per day of data)
  - » Bursts = 50
  - » Compact binary inspiral = 600 (BNS), 300 (BBH), 6,000 (PBH) ......
  - » All sky pulsars = 1,000,000,000 (but can tolerate lower latency & .....)

LIGO's scientific pay-off is bounded by the ability to perform computations on the data.

→ LIGO Data Grid

History of the LIGO Data Grid

- Pre 2000:
  - » Commodity cluster computing shown to be ideally suited to LIGO data analysis needs in prototype analysis
  - » Trade study shows that clusters also provide best performance per dollar spent for LIGO data analysis
- 2000:
  - » Grid Physics Network (GriPhyN) funded via ITR program; LIGO is one of the founding experiments
  - » R&D program to prototype and develop grid-computing paradigm for data intensive experiments; LIGO portion funds development of LIGO Data Replicator
  - » UWM deploys Medusa cluster (funded by MRI) “a system for quick turnaround exploration, and development”

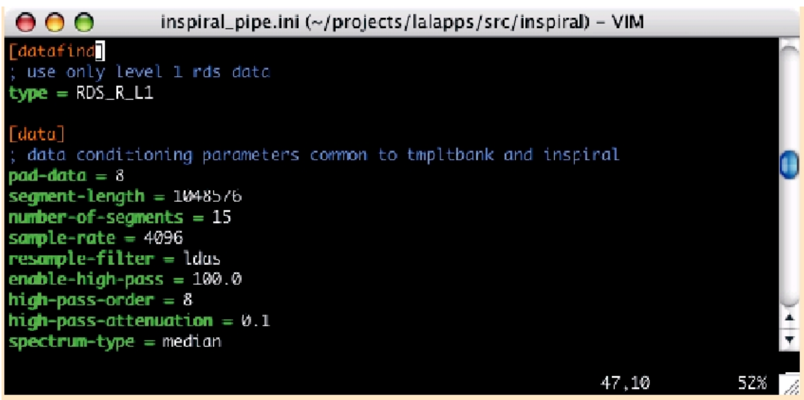


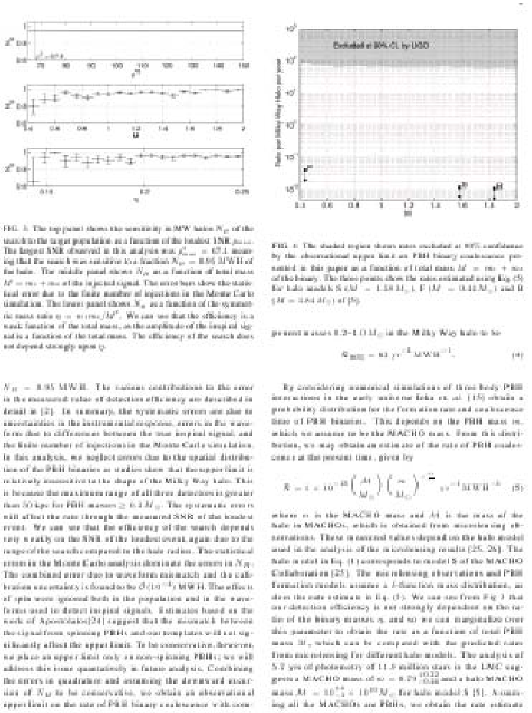
History of the LIGO Data Grid

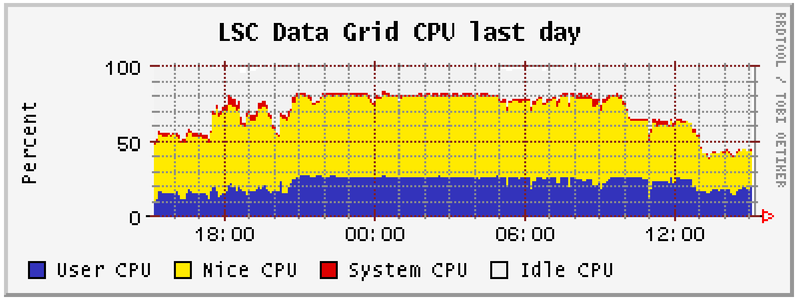
- 2001:
  - » International Virtual Data Grid (iVDGL) funded via ITR program
  - » Deployment of a Grid test bed for data intensive experiments
  - » LIGO portion funds deployment of Tier 2 center at PSU and enhancement of storage capabilities at UWM
- 2003:
  - » “Deploying the LIGO Data Grid; Grid-enabling the GW community” proposal by the LSC to transition from R&D to production deployment and use of the LIGO Data Grid.
- LIGO Data Grid now:
  - » Consists of 2000 (US) and 1000 (EU) CPUs with total peak performance ~5 TFLOPS, 2 TB RAM, and 500 TB of distributed mass storage, in addition to 1.2 PB of tape storage at Caltech.
  - » Provides dedicated computing support for data-intensive gravitational-wave research by 200 scientists of the LSC.

Users and Usage

- LIGO Data Grid (LDG) supports ~200 LSC scientists
  - » Demand for resources is growing rapidly as experience increases and more data become available



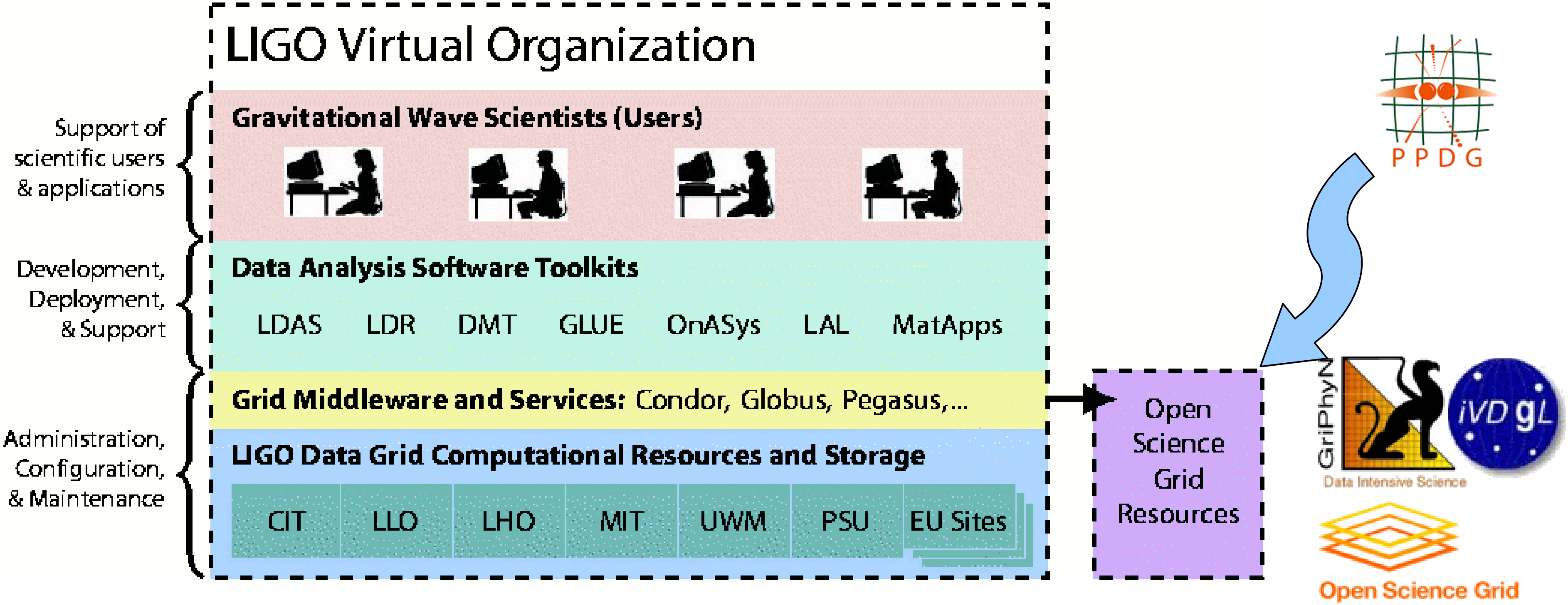




LIGO Data Grid Research & Development

- LIGO production computing tests grid concepts at both the high level of ideas & at the detailed level of tools.
- Example collaborative research with comp. sci.:
  - » LIGO runs a very larg dedicated Condor pool with a complex usage model that stress tests Condor & identifies areas that need to be more robust – excellent collaboration with Condor Team.
  - » LIGO's environment has identified the need for Pegasus to better support rapidly changing executables & to better integrate internal & external resources to appear seamless to the user – ISI Group
  - » LIGO's needs led to improved scaling of the Globus Replica Location Service (RLS) to handle many millions of logical filenames, improved reliability of the server, and an improved API for faster publication of available data and queries – Globus Team

LIGO Data Grid Overview



- Cyberinfrastructure for the LIGO VO
  - » Hardware - administration, configuration, maintenance
  - » Grid middleware & services - support, admin, configuration, maintenance
  - » Core LIGO analysis software toolkits – support, enhance, release
  - » Users - support

LDG System Administration

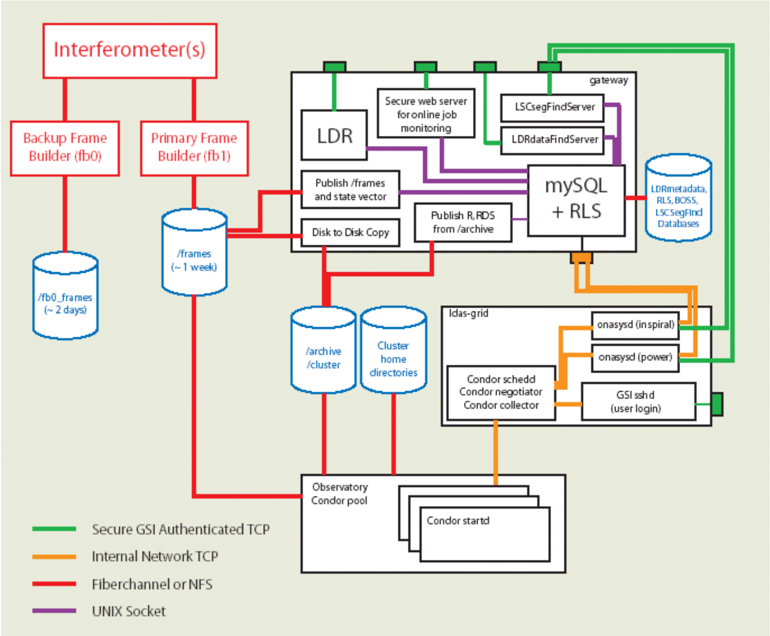
- Hardware and Operating System Maintenance
  - » Commodity hardware running Linux; track changes & enhancements
- Grid Middleware Administration
  - » Deploy LIGO Data Grid Server, configure Condor, LDR & other services.
- Data Distribution and Storage
  - » SAM-QFS, commodity storage on nodes
  - » LIGO Data Replicator to transfer data onto clusters before jobs are scheduled.
- User Support
  - » This is a big job because we have many inexperienced users who are prototyping analyses for the first time ever

Grid services administration and deployment

- LIGO Certificate Authority
  - » Under development, needs long term personnel commitment
- Problem tracking and security
  - » Crude problem tracking in place, needs effort to make useful
- Virtual organization management service
  - » With 200 users, this is an essential service
- Metadata services
  - » Data catalogs, instrument quality information, resource information ..
  - » Need better resource monitoring
- Data Grid Server/Client bundles built on VDT
  - » Bundling of tools for users and admins on LIGO Data Grid

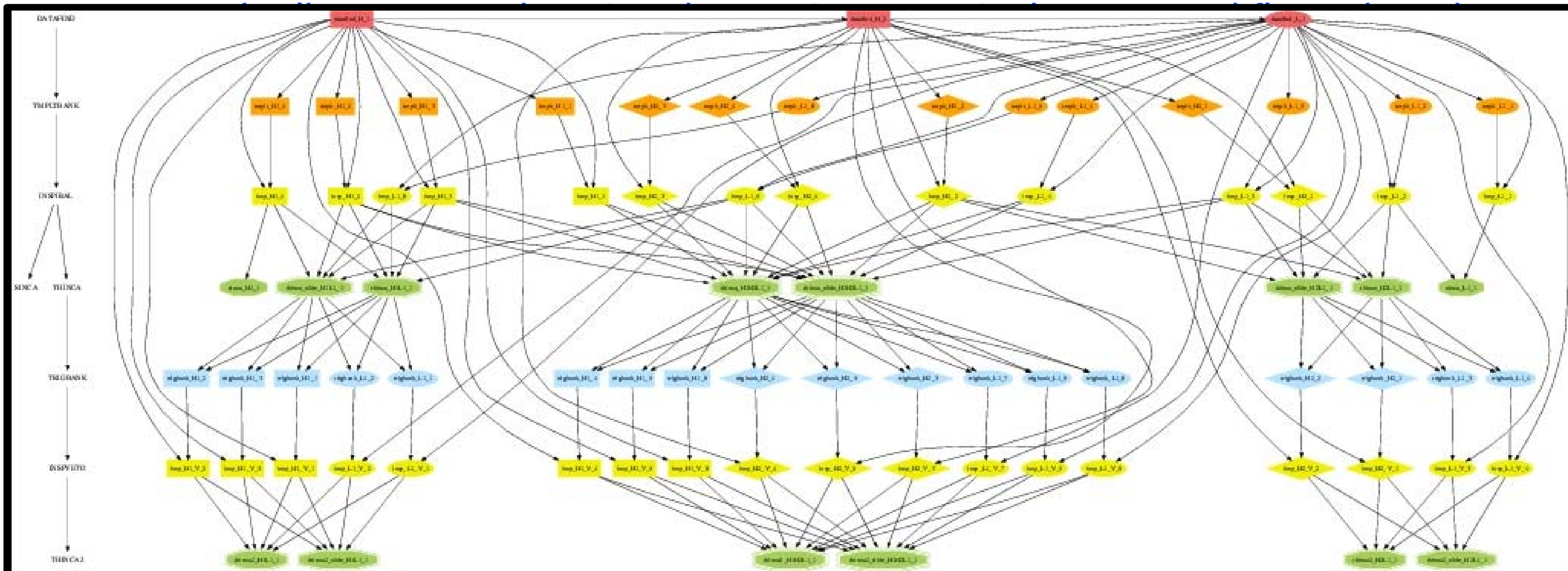
Grid LSC User Environment (GLUE)

- Provides high-level infrastructure for running applications on the LIGO Data Grid
  - » provides an infrastructure to simplify the construction of workflows by treating data analysis applications as modules to be chained together.
  - » use of metadata (e.g. data quality information) allows complicated workflows to be easily constructed.
  - » contains certain LSC specific metadata clients and servers, such as data discovery tools.



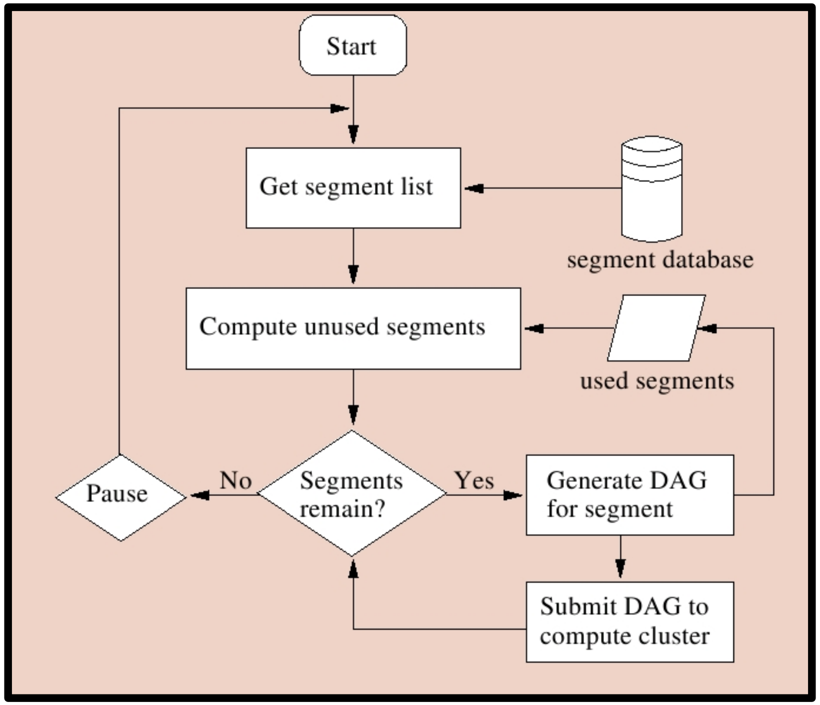
Pipeline Generation

- Complicated workflows
  - » to perform all steps to search data from four LSC detectors
  - » workflow generation built on top of GLUE & LALApps analysis codes



Part of binary inspiral workflow
full analysis workflows have over 10,000 nodes

Online Analysis System (Onasys)

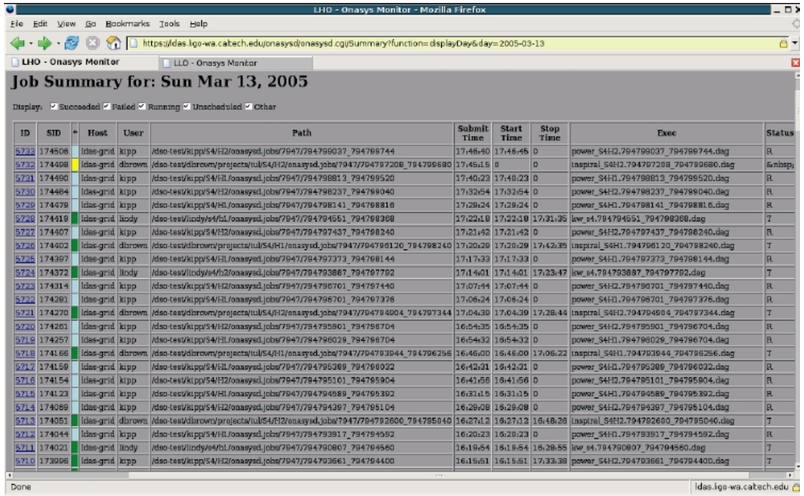


- How it works
  - » Identify data to analyze by query to GSI-authenticating instrument status & data quality service
  - » Find data with LSCdataFind
  - » Configure analysis pipeline using user defined pipeline construction tool
  - » Exsecute on the grid
- ONline Analysis SYStem
  - » Tools to automate real time analysis of GW data
  - » Built on top of GLUE
  - » Uses scientific data analysis pipelines from LSC users

Onasys

- Built on top of Condor, GLUE, Globus
- Database of job information maintained to track progress through workflow
- Online monitoring via a web interface which queries job information metadata database.

- LDG is a lean effort; LIGO specific software is built on Condor & VDT ...
- ... but has relied on much volunteer effort which cannot be sustained .....



Job Summary for: Sun Mar 13, 2005

Display: Succeeded Failed Running Unscheduled Other

ID	SID	*	Host	User	Path	Submit Time	Start Time	Stop Time	Exec	Status
5733	174506		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794799037_794799744	17:46:40	17:46:45	0	power_S4H2.794799037_794799744.dag	R
5732	174498		ldas-grid	dbrown	/dso-test/dbrown/projects/iul/S4/H2/onasysd.jobs/7947/794797208_794799680	17:45:15	0	0	inspiral_S4H2.794797208_794799680.dag	
5731	174490		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794798813_794799520	17:40:23	17:40:23	0	power_S4H1.794798813_794799520.dag	R
5730	174484		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794798237_794799040	17:32:54	17:32:54	0	power_S4H2.794798237_794799040.dag	R
5729	174479		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794798141_794798816	17:29:24	17:29:24	0	power_S4H1.794798141_794798816.dag	R
5728	174419		ldas-grid	lindy	/dso-test/lindy/s4/h1/onasysd.jobs/7947/794794551_794798368	17:22:18	17:22:18	17:31:35	kw_s4.794794551_794798368.dag	T
5727	174407		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794797437_794798240	17:21:42	17:21:42	0	power_S4H2.794797437_794798240.dag	R
5726	174402		ldas-grid	dbrown	/dso-test/dbrown/projects/iul/S4/H1/onasysd.jobs/7947/794796120_794798240	17:20:29	17:20:29	17:42:35	inspiral_S4H1.794796120_794798240.dag	T
5725	174397		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794797373_794798144	17:17:33	17:17:33	0	power_S4H1.794797373_794798144.dag	R
5724	174372		ldas-grid	lindy	/dso-test/lindy/s4/h2/onasysd.jobs/7947/794793887_794797792	17:14:01	17:14:01	17:23:47	kw_s4.794793887_794797792.dag	T
5723	174314		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794796701_794797440	17:07:44	17:07:44	0	power_S4H2.794796701_794797440.dag	R
5722	174281		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794796701_794797376	17:06:24	17:06:24	0	power_S4H1.794796701_794797376.dag	R
5721	174270		ldas-grid	dbrown	/dso-test/dbrown/projects/iul/S4/H2/onasysd.jobs/7947/794794904_794797344	17:04:39	17:04:39	17:28:44	inspiral_S4H2.794794904_794797344.dag	T
5720	174261		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794795901_794796704	16:54:35	16:54:35	0	power_S4H2.794795901_794796704.dag	R
5719	174257		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794796029_794796704	16:54:32	16:54:32	0	power_S4H1.794796029_794796704.dag	R
5718	174166		ldas-grid	dbrown	/dso-test/dbrown/projects/iul/S4/H1/onasysd.jobs/7947/794793944_794796256	16:46:00	16:46:00	17:06:22	inspiral_S4H1.794793944_794796256.dag	T
5717	174159		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794795389_794796032	16:42:31	16:42:31	0	power_S4H1.794795389_794796032.dag	R
5716	174154		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794795101_794795904	16:41:56	16:41:56	0	power_S4H2.794795101_794795904.dag	R
5715	174123		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794794589_794795392	16:31:15	16:31:15	0	power_S4H1.794794589_794795392.dag	R
5714	174069		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794794397_794795104	16:29:08	16:29:08	0	power_S4H2.794794397_794795104.dag	R
5713	174051		ldas-grid	dbrown	/dso-test/dbrown/projects/iul/S4/H2/onasysd.jobs/7947/794792600_794795040	16:27:12	16:27:12	16:48:26	inspiral_S4H2.794792600_794795040.dag	T
5712	174044		ldas-grid	kipp	/dso-test/kipp/S4/H1/onasysd.jobs/7947/794793917_794794592	16:20:23	16:20:23	0	power_S4H1.794793917_794794592.dag	R
5711	174021		ldas-grid	lindy	/dso-test/lindy/s4/h1/onasysd.jobs/7947/794790807_794794560	16:19:54	16:19:54	16:28:55	kw_s4.794790807_794794560.dag	T
5710	173996		ldas-grid	kipp	/dso-test/kipp/S4/H2/onasysd.jobs/7947/794793661_794794400	16:15:51	16:15:51	17:33:38	power_S4H2.794793661_794794400.dag	T

Results of LIGO Data Grid Research

- The LDG has enabled 13 results papers on searches for gravitational waves from
  - » binary neutron star and black hole systems,
  - » isolated spinning neutron stars
  - » a gravitational stochastic background from the early universe
  - » gravitational-wave bursts from supernovae or other energetic events
- The LDG has also enabled more than 17 technical data analysis papers
- Six (6) students have received PhDs based on results obtained using these resources; Fifteen (15) more are in pipeline

LIGO Data Grid needs

- Need personnel committed multi-years to support, enhance and deploy this dedicated cyber-infrastructure
- System administration: 12 FTEs
- LIGO Data Grid Services: 8 FTEs
- Core analysis software: 14 FTEs
- Analysis profiling/help: 1 FTE