End-to-end thumbnail of computing model and needs

Patrick R Brady

University of Wisconsin-Milwaukee

LIGO Scientific Collaboration

G050586

# End-to-end thumbnail of computing model

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

- 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

- Requirments for flagship searches (~2 x fudge-factor)
  - » 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
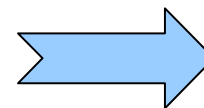
History of the LIGO Data Grid

- ## 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
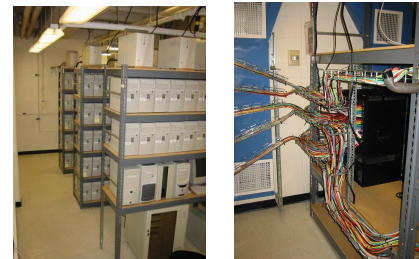
- **2001:**
  - » International Virtual Data Grid (iVDGL) funded via ITR program
  - » Deployment of a Grid test bed for data intensive experiments
  - » LIGO portion funds deployment of Tier 2 center at PSU and enhancement of storage capabilities at UWM

- **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.

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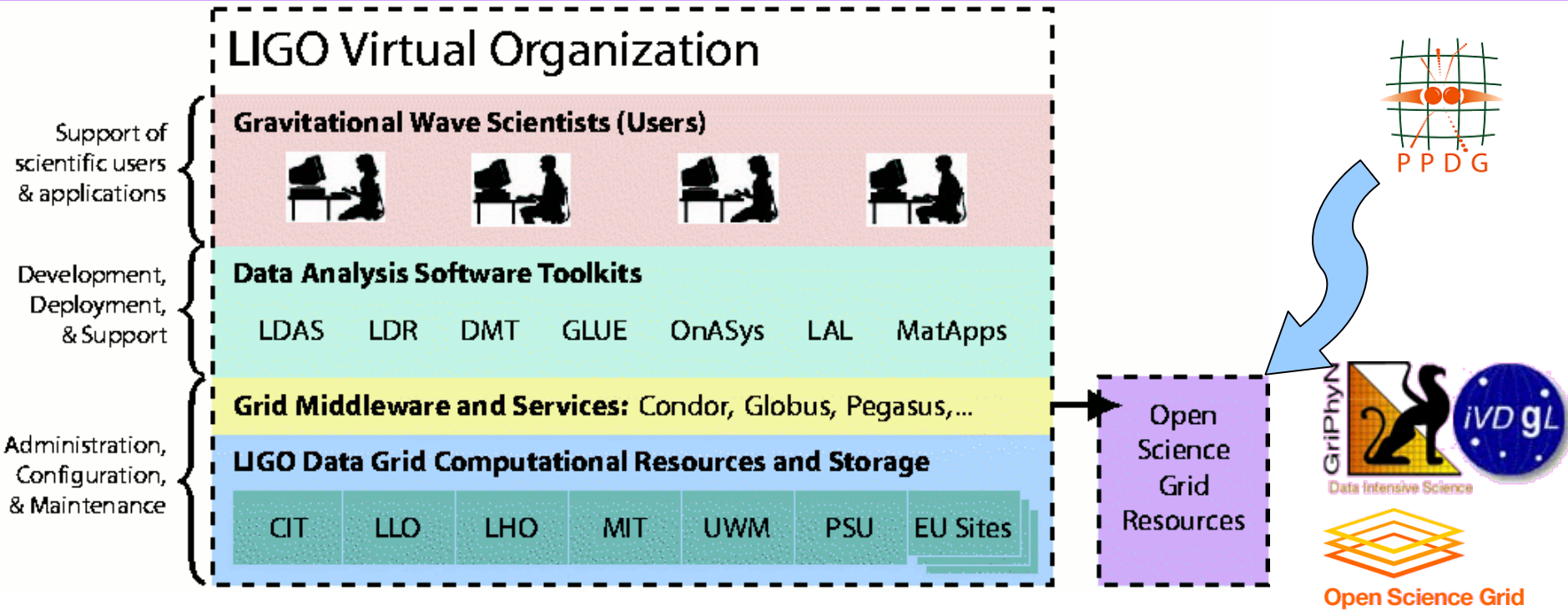
- **Cyberinfrastructure for the LIGO VO**
  - » Hardware - administration, configuration, maintenance
  - » Grid middleware & services - support, admin, configuration, maintenance
  - » Core LIGO analysis software toolkits – support, enhance, release
  - » Users - support

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

G050586

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

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

Task	Fiscal Year (Starts 1 Oct. of previous year)									
	2007		2008		2009		2010		2011	
	Req.	Curr.	Req.	Curr.	Req.	Curr.	Req.	Curr.	Req.	Curr.
System Administration	3	9	5	7	5	7	5	7	5	7
LIGO Data Grid Services	4	4	7	1	7	1	7	1	7	1
Core Analysis Software	5	9	9	5	9	5	9	5	9	5
Analysis Profiling/Help	1	0	1	0	1	0	1	0	1	0
Totals	13	22	22	13	22	13	22	13	22	13
Total Need	35		35		35		35		35	

PIF Proposal to Fund Positions   LIGO Ops Funded Positions

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
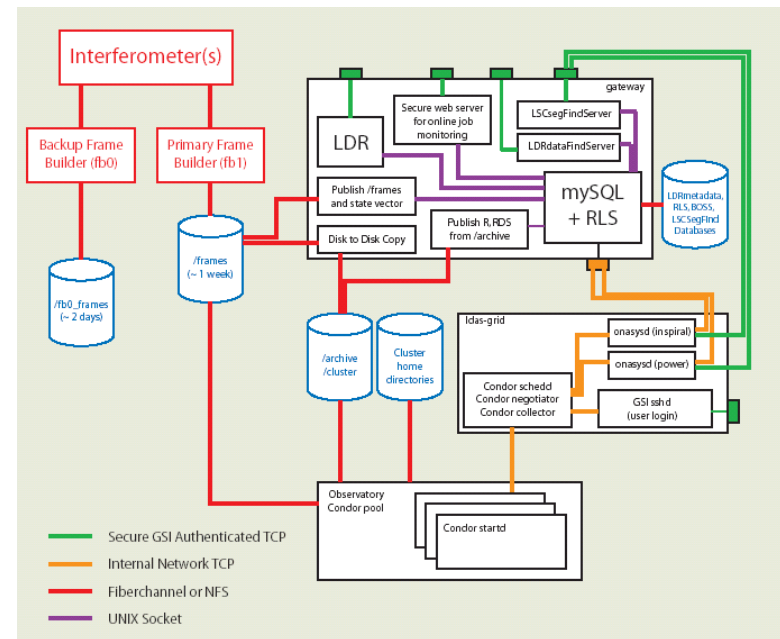
- ## Complicated workflows

  - » to perform all steps to search data from four LSC detectors

  - » workflow generation built on top of GLUE & LALApps analysis codes
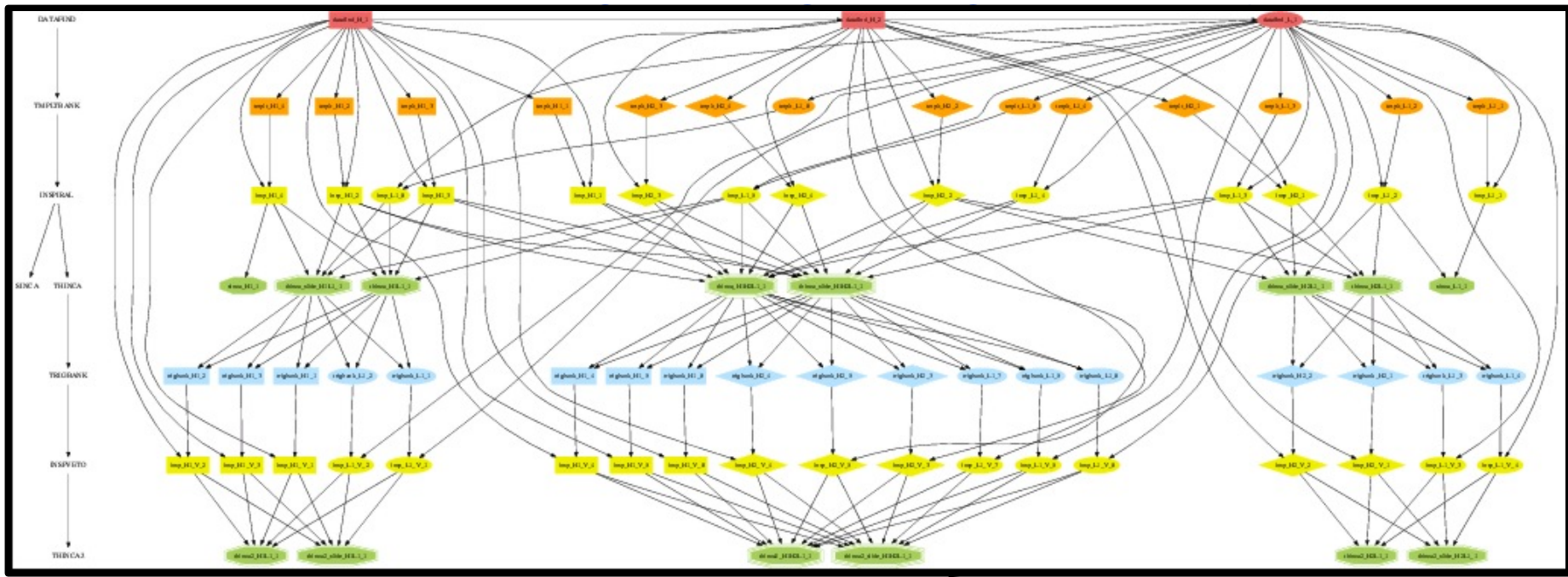
Part of binary inspiral workflow
full analysis workflows have over 10,000 nodes

Online Analysis System (Onasys)
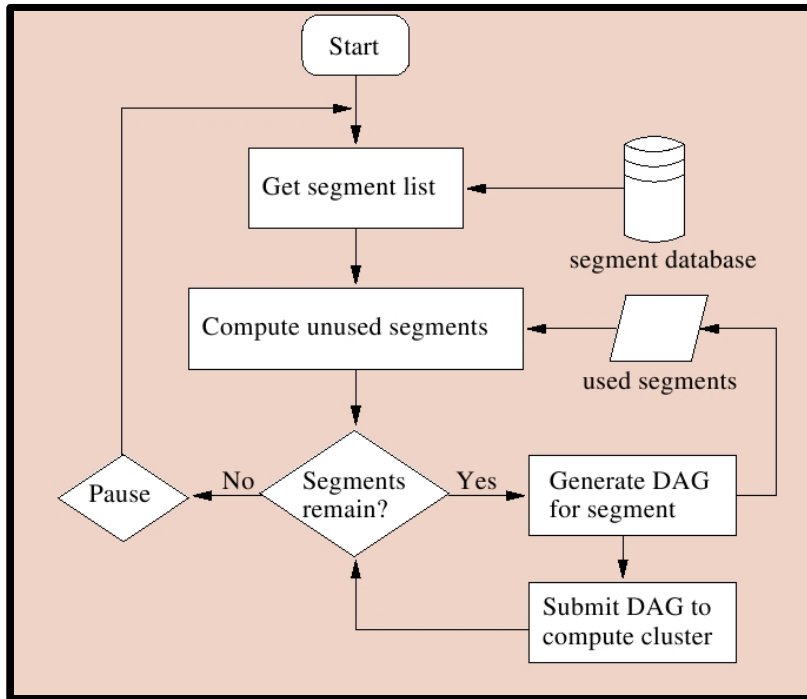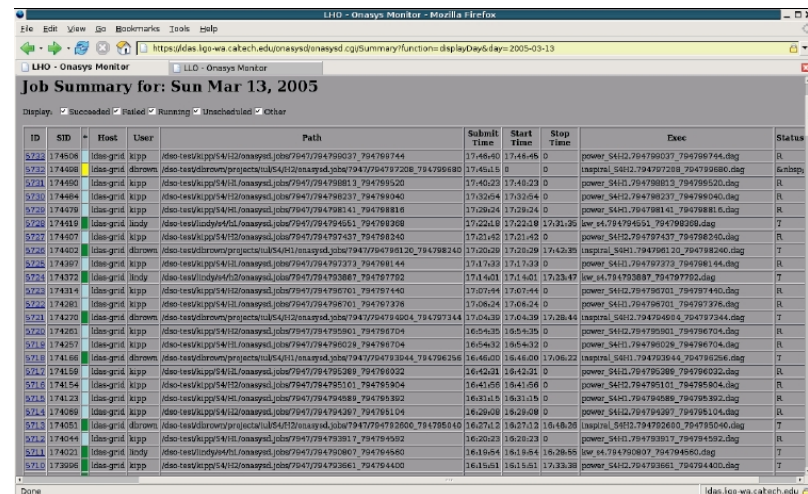
- **How it works**
  - » Identify data to analyze by query to GSI-authenticating instrument status & data quality service
  - » Find data with LSCdataFind
  - » Configure analysis pipeline using user defined pipeline construction tool
  - » Exescute on the grid

- **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

G050586

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 .....