Technical Note | LIGO-T0500XX-0X-Z | 2009/11/05

**LSC Data Analysis Software Planning Document**

Data Analysis Software Working Group

April Fool’s Meeting participants: Stuart Anderson, Kent Blackburn, Patrick Brady, Duncan Brown, Kipp Cannon, Junwei Cao, Philip Charlton, Jolien Creighton (telephone Fri and Sat), Rejean Dupuis, Ben Johnson, Dan Kozak, Scott Koranda, Albert Lazzarini, Ed Maros, John McNabb (telephone Fri), Brian Moe, Craig Robinson (telephone Fri and Sat), Igor Yakushion (consulted by phone), John Zweizig

---

DRAFT

This is an internal working document of the LIGO Scientific Collaboration

---

California Institute of Technology  
LIGO Project, MS 18-34  
Pasadena, CA 91125  
Phone (626) 395-2129  
Fax (626) 304-9834  
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology  
LIGO Project, Room NW17-161  
Cambridge, MA 02139  
Phone (617) 253-4824  
Fax (617) 253-7014  
E-mail: info@ligo.mit.edu

WWW: [http://www.ligo.caltech.edu](http://www.ligo.caltech.edu)  
## Contents

1 Overview 2  
1.1 Broad Issues to Address before S5 2  

2 LSC Data Replicator 3  

3 Segment Database and Tools 4  

4 File and Segment Metadata 5  

5 Onasys 5  

6 Data Monitor Tool 7  

7 LDAS 7  

8 tclGlobus 9  

9 LSC Data Grid Client/Server 9
1 Overview

The Data Analysis Software Working group has been charged to coordinate the development, deployment, documentation and testing of software for data analysis within the LSC. This document summarizes development and deployment coordination discussions for the period April 2005 – April 2006. This document will be reviewed (at least) annually to identify places where effort is needed to meet the LSC’s analysis software needs. The main iterations and targets are outlined in Table ??.

1.1 Broad Issues to Address before S5

The following issues were raised as concerns and needing action at the DASWG April Fools 2005 meeting:

1. Calibration file publishing and versioning. The publication of calibration files to LSC-dataFind was initiated in S4; it works. The following items need to be addressed:
   - Calibration committee sign-off on procedures
   - Streamline frame generation at the observatories
   - Validation procedure prior to publication
   - Automated publication of validated frames
   - Versioning of software and models

2. Calibrated data
   - Code walk-through of Xavier’s time-domain code
   - Validation of the method
   - Validation of data from S4
   - Calibration committee sign-off

3. Matlab compiler problems depending on O/S, compute center, and code
   - Decide on limited compiler set
• Deploy default (and alternatives where needed) and test compile
• Freeze compiler at start of S5

4. Stability of the O/S and clusters
• Identify an O/S committee
• Examine FC4 for kernel bug fixes and stability
• Recommendation at end of June 2005.
• Dedicated maintenance times
• Dedicated hardware for testing

5. Automatic retrieval of hardware injections
• Identify Sim tables for this purpose
• Add features to appropriate piece of code
• Set up default location for download or publish?

6. Automatic data quality generation and insertion into database
• Segment table design is addressed below
• Tools putting info into database are addressed below
• Monitors and other tools need to generate info

7. Carrot?
• Where, when, how?

8. Expert panel to help identify code bottlenecks for users; bring to Analysis chairs for advertising

9. Require codes to be tagged

2 LSC Data Replicator

This is a grid tool layered over the globus toolkit to enable large scale data transfer around the LSC computing centers. This software is deployed at CIT, LHO, LLO, MIT, PSU, UWM, AEI and will come on-line at Cardiff and Birmingham. During S4, this software replicated all data to CIT and reduced data to other Tier-N centers with latencies of a few hours. Scaling of the data discovery component remains the major risk issue for S5+.

The main LDR developer is Scott Koranda (75%) with other participants being Kevin Flasch (25%) and Brian Moe (50%).

Required upgrades to be done by May 1
1. MySQL to recent 4.1.x version using precompiled binaries
2. GTK 4.0 components RLS pyGlobus possibly GridFTP server and client
3. Python (possibly)

May 1

1. Completed Administration Guide
2. RLS scaling test to $10^8$ LFNs and $4 \times 10^8$ PFNs examine bloom filter size examine false positive rate

June 1

1. Begin testing of overhauled LDRSchedule and LDRTransfer
2. Increased control of remote site use
3. Ability to quickly adapt to configuration changes
4. Implementation of high- and low-watermarks for transfer agents
5. Fix bug in having multiple transfer agents per site (unless obsoleted)
6. Ensure design scales appropriately on collections

July 15

1. Finalize LDRSchedule and LDRTransfer
2. Publishing service (possibly)
3. LDRdataFindServer (possibly) better behavior on long queries more useful error messages
4. Logging with detail about site queueing

August 1

1. Monitoring facility - scope TBD

No Dates:

1. LDRVerify (possibly) (scope TBD?)

3 Segment Database and Tools

May 1 2005 Define DB tables for segments.

1. Segments defined as:
   (a) Immediate segments from frame channels (statevec, activity) - from publishing script
(b) DQ Segments from DMT monitors - from DMT monitors (via GC connection?, cluster DMT jobs?)
(c) ? Segments from Condor searches?

2. User access to segments: - i.e., tables designed/populated to allow
   (a) Segwizard-type interface
   (b) LSCsegFind -t Science,....... etc.
   (c) Something more DMT-native?
   (d) DTT access?

==== June 1 2005 ==== Complete client and server parts of the segment database.

1. Client
   (a) DMT ”class” and/or segment auto-insertion service of some sort.
   (b) Tool for inserting from cluster
   (c) Back-end for Segwizard?

2. Server
   (a) Main purpose is to serve as authenticator based on X509 certs.
   (b) Will abstract database from client tools, as a side effect.

4 File and Segment Metadata

Metadata Plan for S5

May 1 Decide on a workable Metadata facility
   1. Schema
   2. Services
   3. Replication protocol

June 1 Begin testing Metadata for ability to handle appropriately large abount of data

July 15 Finalize Metadata system

August 1 Monitoring facility - scope TBD

Aug 15 - Sep 15 Continue verification of metadata integrity and incremental validation

5 Onasys

2005-05-01

1. - new data segmentation algorithm
2. - track trigger intervals
3. - job recovery
4. - have Quill in our hands

2005-06-01
1. - replace BOSS with something of our own design
2. - rework segment database querying algorithm
3. - have tried Quill

2005-06-15
1. - final decision on whether or not to go with Quill
2. - if Quill is not to be used, add solving the connection problem in
dagdbUpdater to the TODO list

2005-07-15
1. - real log file, and log file rotation

2005-09-01
1. - package components separately
2. - scimon web page (incl. lag plots)

2005-12-01
1. - automated house-cleaning
2. - move as much of the daemon as is possible into the library

Other tasks without specific deadlines
1. - non-integer segment boundaries
2. - dag node log retrieval via web interface
3. - create a slave database for the web server
4. - eliminate the daemon’s need for internal state (lowest priority)
5. - allow DAG generator to be a directly-linked Python module (lowest priority)
6 Data Monitor Tool

There are two dmtoffline targets before S5 starts:

1 May 2005 There will be a test version of dmtoffline available before May 1 2005. Users should be able to check it out from GDS cvs and build it successfully on FC, RHEL and Solaris. Several tasks include:

1. Check in all the auto make files into the GDS cvs
2. Adapt auto make files to latest updates of DMT codes (e.g. using new version of framecpp, root, fftw3, ..)
3. Initial test it on FC, RHEL and Solaris
4. Hope some users would like to download it and participate in the test so that we can have more feedback to get it evolved.

15 June 2005 Binary distributions should be make available to users by June 15 2005. As required by the LSCsoft release mechanism, RPM should be made available. Several tasks include:

1. Include distribution mechanism into the auto make files
2. Binary release as tar balls
3. Binary release as rpms for Linux

7 LDAS

LDAS will prepare two release prior to the start of the "Astrowatch" end-to-end testing planned to begin July 15th. This first release will tentatively be the first of June. This release will primarily be driven by the Laboratories plans to upgrade to Solaris 10, requiring a new release of LDAS to assure proper software system behavior. In addition, this release will target:

1 June 2005

1. The ability to generate checksum information and store it to disk during the creation of RDS frames by LDAS.
2. Add logic to the createRDS driver script to handle the occasional "short" frame issue associated with needing to run multiple createRDS jobs in the Science Runs to keep up with the increasingly larger frames.
3. Improvements to error messages within the frameCPP I/O library as requested by the DMT developer.
4. Removal od dependencies on the wrapperAPI on LDAS development. This is to include replacing the current dataPipeline testsuite with dataStandalone and putStandalone testsuites removing the need to have a cluster of nodes for LDAS.
5. Enhance FrVerify utility to verify metadata (in addition to the checksums).

6. Add support for GEO frame types in the createRDS command, allowing for LIGO and GEO data to be merged into a single output RDS frame.

7. Add support for compression of all frames output by LDAS.

1 July 2005

The second release of LDAS will tentatively scheduled for the first of July. It will address other open problem reports of importance to the start of the S5 Science Run. These will include as a minimum the following:

1. Fix to the updating of history records by the frameAPI.

2. Decrease the size of the memory leak in the frameAPI to allow for a comfortable margin in a long duration science run (one year or more).

3. Improve performance of the createRDS command through changes in compiler technologies, software profiling or faster hardware in order to improve the "short" frame issue associated with having to run multiple createRDS jobs to keep ups with the raw frame generation.

4. Improve performance and scalability to number of subdirectories of the diskCacheAPI.

5. Port to 64 bit compiles on Solaris hardware to improve performance and memory address space available to the frameAPI process.

6. Remove dependencies on STLPort standard template library if the GCC standard template library can be made thread safe and stable.

7. Port to Linux Fedora Core 4 if the collaboration makes a decision in time for this second release (Note: current schedule carries making a decision on July 1 so this is unlikely to be in this release).

No deadline set

Other changes to LDAS that are targeted for post ”Astrowatch” testing that may make it into an LDAS release prior to the start of the S5 Science Run include:

1. Port to Fedora Core 3 in the event that it is not part of the second (July 1st) release of LDAS.

2. Upgrade LDAS to a newer version of Tcl Interpreter (8.4.x series).

3. Upgrade LDAS to a newer version of GCC compiler (4.1.x series).

4. Add certificate based authentication to user requests.

The hardware used by LDAS and other LSC software will be upgraded prior to the start of the S5 Science Run. The Solaris operating system will be the first change schedule for spring, prior to the first LDAS release. Other hardware changes may include additional CPUs in gateway servers and the possibility of increasing the size of the clusters.
8  TCLGlobus

The TclGlobus project will prepare two alpha releases of the software in the form of tarballs prior to the start of the "Astrowatch" end-to-end testing planned to begin July 15th.

1 May 2005 The first tarball release of TclGlobus will take place on May 1st and will support the following Globus 3.2.1 packages:

1. The Globus_IO package
2. The Globus_XIO package
3. The Globus_FTP_Client package

The second alpha release of TclGlobus will occur after application developers have had the opportunity to provide feedback and is tentatively scheduled to take place two months later. This release will also be a tarball and will include the same packages as the first release with any fixes and enhancements requested during the development of applications based on these three packages.

9  LSC Data Grid Client/Server

While it is difficult to estimate how much efforts have to be put into the LDG packaging, there are still two targets defined temporarily for S5:

LDG 4.0 - June 1 2005 I can figure out the exact estimation before June 1 2005. Some tasks include:

1. VDT update
2. using new pacman
3. new LSCdataFind
4. new RLS-Client
5. new LSC-VDS
6. fix bugs for LSCcertUtil

LDG 5.0 - September 1 2005

Something in future (not related to S5)

1. new LSC certificate
2. Mac support
3. improve error messages