

Data from the LIGO I Science Run

Albert Lazzarini

*LIGO Laboratory
California Institute of Technology, Pasadena, CA 91125
LIGO-P010002-00-E*

Abstract. The LIGO¹ I Science Run is planned to begin in mid-2002. The characteristics of the data stream, data volumes, data products, and data availability are discussed. The data analysis activities will be undertaken by the LIGO Scientific Collaboration (LSC²). These activities include operating dedicated on-site pipelines at the LIGO observatories. In addition, a dedicated off-site facility for will be dedicated to melding data from different interferometer datastreams (both LIGO and eventually those of other international projects as part of a network-wide analysis effort). Exploratory university-based research on LIGO data will likely be supported in part by the nascent US computing grid. LIGO Laboratory and the LSC are working on grid computing efforts within the GriPhyN (Grid Physics Network) collaboration research activities.

LIGO DATA CHARACTERISTICS

LIGO interferometers produce audio-band time-series data digitized at a number of frequencies that are powers of two (refer to **TABLE 1**). The data are acquired as 2-byte integers from the acquisition system. They also include computed data, which are stored as 4-byte real numbers. In total, the three LIGO interferometers will generate between 7 – 9 MB/s of data as time series from 2000+ parallel channels. Because the data rate is quite high and because the data are all time series, LIGO, along with its international partner projects, has developed a format³ for acquiring and archiving the raw data. This so-called Frame Format captures all the data from a single interferometer that were collected during the same epoch. The frame duration is arbitrary and is typically between 1s and 32s in length. Figure 1 shows a schematic of the frame format and its organization. All data are stamped with the GPS (Global Positioning System) time, which is accurate to <100 ns, at the time of acquisition. Data from multiple detectors will be merged into composite frames spanning the same GPS epoch for coincidence analysis.

LIGO Data Volume

The datastream comprises less than 1% (3 channels at 32 kB/s each) of strain or so-called “science-channel” data that will contain the astrophysical signatures for discovery. The rest of the data stream constitutes ancillary or auxiliary channels. These channels are grouped in two broad classes. The first group (all the other interferometer channels and health/status) is used to monitor the behavior of the many electrical-mechanical-optical servos that are required to maintain the very complex

LIGO interferometers in a state of maximum sensitivity. This group corresponds to roughly 80% of the acquisition bandwidth. When the instrument is operating properly, these channels should show no anomalous behavior: a gravitational wave strain signal should not show up in these other channels. On the other hand, an anomalous instrumental artifact would be present in both the auxiliary channel data and the strain channel, allowing instrumental “glitches” to be detected and vetoed by suitably processing the auxiliary (“health & status”) interferometer channels.

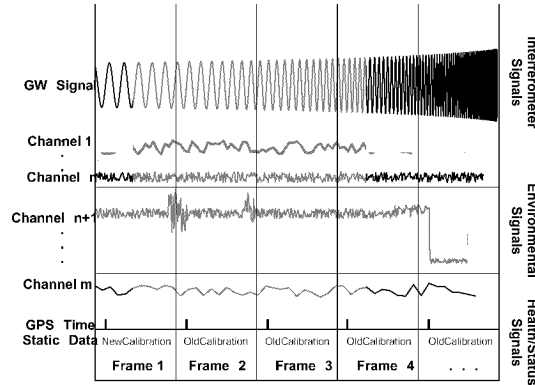


Figure 1. The LIGO datastream constitutes very many channels acquired in the time domain with sample rates up to 16384 samples/s. The data are organized in C-structures called frames which capture all the data collected by the LIGO instruments during a particular epoch. This encapsulation is intended to ensure a self-consistent set of raw data for archival.

The second class of channels is dedicated to monitoring the local terrestrial environments at the observatories. This group corresponds to roughly 20% of the acquisition bandwidth. These channels come from a variety of non-interferometric sensors that will be used to monitor at high sensitivity a large variety of geophysical and local phenomena of non-astrophysical. This is critical to the operation of LIGO because many geophysical and anthropogenic will likely be detectable by interferometers operating at sensitivities of $\sim 10^{-18}$ m RMS. The detection of the physical disturbances by other means will be used to establish vetoes for these types of real signals that are not of astrophysical significance.

TABLE 1: LIGO I Data Channel Count by Acquisition Rate for Hanford Observatory (for 2 interferometers)

| Acquisition Rate, samples/second (16 bit) | Number of Channels |
|--|--------------------|
| 16834 | 124 |
| 2048 | 532 |
| 256 | 109 |
| 64 | 205 |
| 16 | 208 |
| <hr/> | |
| Total No. of Channels: | 1178 |

Additional LIGO datasets will be generated from the raw data. In all, it is expected that three levels of frame data will be created (Levels 1, 2,3) and distributed for data analysis. These levels, their sizes, and their expected uses are shown in **TABLE 2**. Access to Level 1 data should be infrequent and will serve primarily to verify or veto putative detections by enabling scientists to look broadly across all channels during an epoch of detection. The higher levels of data constitute progressively more refined datasets that will likely be of interest to the science teams who are analyzing the data for astrophysical signatures.

TABLE 2: LIGO I Data Products and Volumes

| Mode | Raw and Derived Data for On-line Diagnostics | Level 1 Full (100%) frame data for archiving | Level 2 Strain and data summary, QA channels | Level 3 Strain best estimate |
|--|--|--|--|---------------------------------|
| Uncompressed Rate (MB/s) | LHO: 9.479 LLO: 4.676 Total: 14.155 | LHO: 4.698 LLO: 2.278 Total: 6.975 | Total: 0.300 | Total: 0.006 |
| w / 50% Hardware Compression, MB/s onto tape media | - | LHO: 2.349 LLO: 1.139 Total:3.488 | Total: 0.150 | - |
| Data growth rate, per year of integrated running, TB/yr. | - | LHO: 74 LLO: 36 Total:110 | Total:9.5 | Total: 0.200 |
| Total including redundant 100% backup, TB/yr. | - | LHO: 148 LLO: 72 Total:220 | Total:19 | - |
| Purpose | For on-line monitoring of interferometers | Deep permanent archive | Science analysis, data exchange | Science analysis, data exchange |
| On-site look-back time | Must use real-time control and monitoring system (CDS) disk caches | LHO Disk cache: 28 d LLO Disk cache: 28 d | - | - |
| Off-site look-back time | - | As long as required | In perpetuity | In perpetuity |

Level 2 data correspond to the strain channel and those channels which may have sufficiently strong cross-correlations with the strain channel to enable regression and removal of the cross-correlations to be performed. In addition, a data-quality (QA) channel will be useful to quickly exclude from further analysis those data segments that have exceedingly poor sensitivity. The Level 2 datasets will be of interest to both astrophysicists who are looking at the data in detail and to interferometer scientists who are trying to understand interferometer behavior and terrestrial correlations.

Level 3 data correspond to a smaller subset, consisting of the strain channel, data-quality and possibly a few other measures of overall interferometer (or environmental) characteristics. Level 3 data will be the sets that are exchanged and merged among different detector projects for network analysis. It is likely that idiosyncratic instrumental channels specific to a particular instrument will not be of much utility to a network analysis team composed of members from many projects.

LIGO Databases

In addition to the raw time-series frame data, LIGO has developed a capability to store important metadata into a relational database archive. The relational database is composed of a number of inter-related database tables⁴ that are used for a number of functions. Figure 2 present a schematic showing the database tables and their inter-relationships. The database provides a tabulation of those instrumental performance

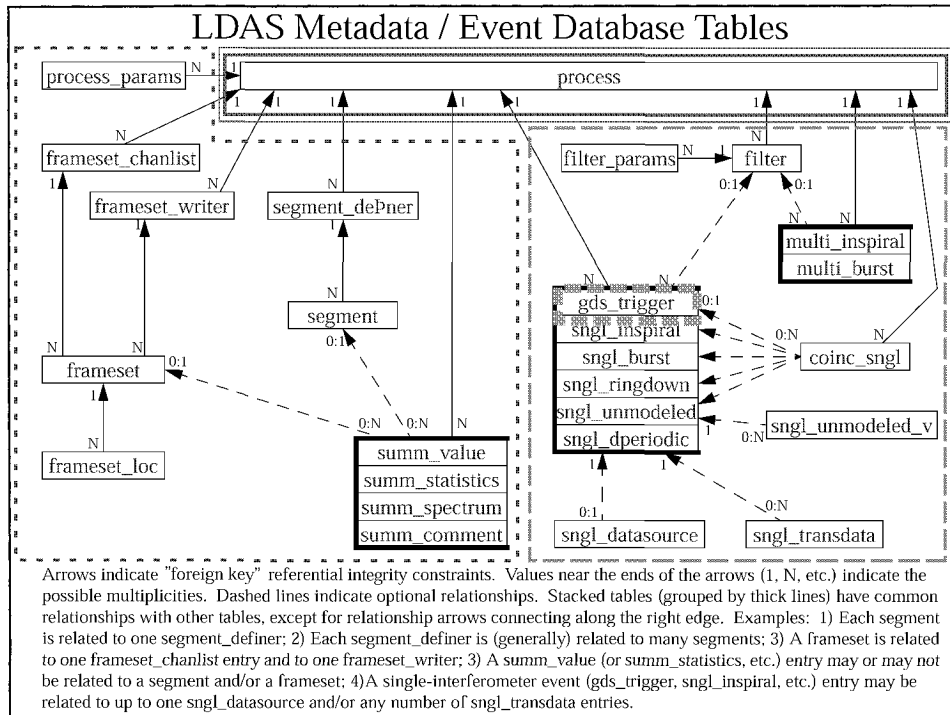


Figure 2. LIGO relational database design for storing metadata and events/triggers generated by pipeline analyses. Dotted box: tables used to summarize and index into time-series frame data. Dashed box: tables used to capture events or triggers generated by search analysis pipelines. Small dashed box (gds_trigger): table used to capture instrumental artifact triggers for later vetoing of data. Double thin-lined box (process): table used to capture information about software that was used to generate metadata.

metrics that are useful for a first-look survey of the available frame data. Examples include statistics (RMS, mean, peak-to-peak) for key channels. In addition,

representative spectra or other graphic elements may be stored as binary large objects (BLOBS). Another intent of the design of the database is to facilitate later use of LIGO data for data mining purposes. A remote client software package has been developed (GUILD⁵) to enable researchers to query and retrieve information stored in the database.

LIGO Data Flow Model

LIGO Laboratory constitutes four geographically isolated sites: two observatories in Washington⁶ and Louisiana⁷, and two university laboratories at Caltech and MIT⁸. Data acquisition in large volume takes place at the observatories and the data then migrate first to Caltech and MIT, and then a number of LIGO Scientific Collaboration institutions. Figure 3 presents a schematic of this data flow. The large on-line disk arrays systems are used to store the data as they are produced. These arrays are common to the acquisition system and the data analysis systems. Data will become available on-site as soon as they are produced. The on-line global diagnostics system is part of the real-time control and data systems (CDS⁹). Real-time software monitors are used to track instrumental performance and to generate trigger data that are sent

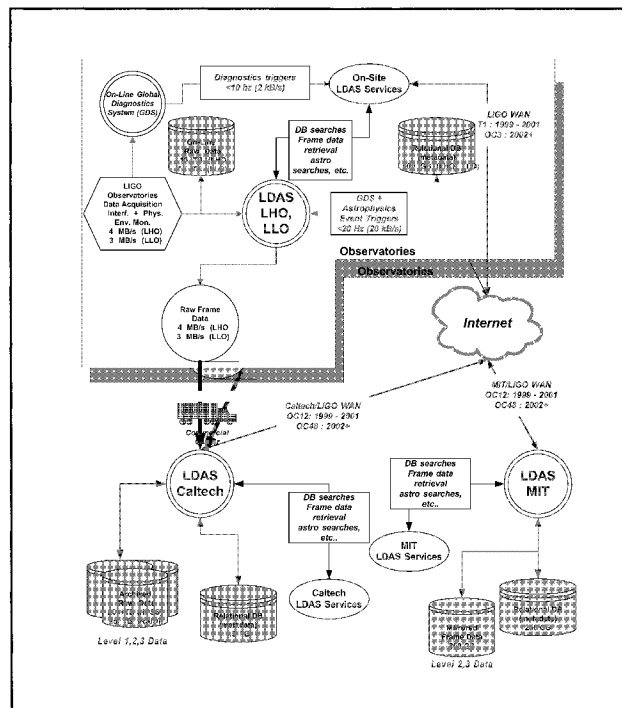


Figure 3. LIGO data flow from observatories to the repositories at Caltech and MIT.

for storage to the data analysis system (LDAS¹⁰) database server. In addition, pipeline search algorithms running in synchrony with the rate at which data are acquired sift through the data. These employ a number of digital filtering techniques to look for events of astrophysical interest. The complexity and computational efficiency of the search algorithms depends on the frequency-time volume that needs to be searched for different classes of events¹¹.

Raw frame data will be written to data tape and they will be sent to Caltech, where they will be ingested and made available through a deep archive (HPSS¹²). The rate of growth and volume of the metadata databases will be sufficiently small so that they can be migrated to the archive over the LIGO wide area network (WAN). The MIT databases are small mirrors of the deep archive at Caltech. All data products will be available across the LIGO Scientific Collaboration institutions that are performing data analysis studies.

Initially, the main emphasis of the collaboration-wide astrophysics searches will be development and maintenance of the pipeline searches. Many of the search algorithms require dedicated computational resources running nearly all the time. The Laboratory is providing these resources for coordinated searches that are formed within the LIGO Scientific Collaboration.

Exploratory prototyping by individuals will be mostly performed at the home institutions of members of the LIGO Scientific Collaboration. Data distribution will be provided to enable individuals to have access to datasets for their research. The LIGO data use model is still evolving as this new collaboration learns to look at the data from the instruments that are just now coming on-line.

Access to LIGO science data is available through membership in the LIGO Scientific Collaboration. The collaboration believes that instrumental idiosyncrasies that will always be present in data from the interferometers will require intimate working knowledge of the instruments before the data can be analyzed or interpreted for astrophysical content. This resembles more the model of data access in the high-energy physics community than the astronomical community. The LIGO Scientific Collaboration is open to interested scientific researchers who establish a memorandum of understanding (MOU¹³) with LIGO Laboratory for access to the data archive.

GRID PHYSICS NETWORK (GRIPHYN) AND LIGO

The Grid Physics Network (GriPhyN¹⁴) is a collaboration among high-energy physics experiments, astronomy, gravitational physics, and computer science whose mission is to mobilize large-scale information technology resources for scientific research. The high-energy physics experiments that are involved include the U. S. members of the CMS¹⁵ and ATLAS¹⁶ collaborations who are developing experiments at CERN for the Large Hadron Collider (LHC¹⁷). The astronomy members are from the SLOAN Digital Sky Survey (SDSS¹⁸) and are developing technologies for the National Virtual Observatory (NVO¹⁹). LIGO Laboratory and its Scientific Collaboration represent gravitational wave astrophysics. The four physics programs each face challenging computational and database needs that place emphasis on massive data movement over high speed networks. They also need large-scale

distributed computing resources in order to analyze the data products, which are generated by the experiments. The collaboration includes a number of key computer science institutions that have been pioneering the concepts of a computational grid (in a very rough analogy to the electrical power grid).

Within the GriPhyN data grid concept, there exists a hierarchy of levels, or tiers, of data repositories and computational resources. Referring to Figure 4, at the highest level there is a Tier 0 center (for high-energy physics only). Within the U. S., there will be one Tier 1 center per physics project. This center represents the archive and dedicated computing resources that are required to perform essential data analysis functions common to large groups of researchers. At the next lower level are a number of the Tier 2 centers – regional centers that serve as a data mirrors for the Tier 1 center and that have computational resources that will be available to the respective collaborations. The next level, Tier 3, corresponds to the resources available to individual university research groups. Finally, at the lowest level, Tier 4, there are the individual workstations available to individuals to perform small-scale analysis and prototyping of algorithms.

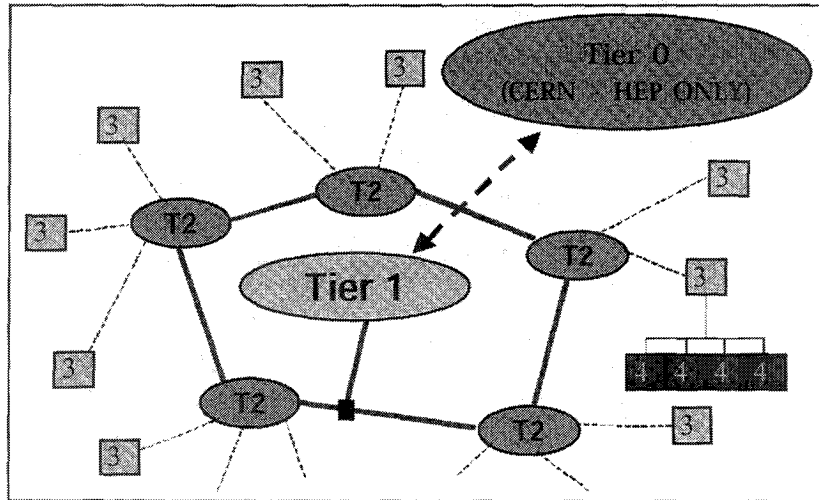


Figure 4. Schematic of the GriPhyN hierarchy of computer and database resources within the U. S.

Within the LIGO Scientific Collaboration, there will be between 3 and 5 Tier 2 centers. These will provide a number of services that will help to off-load the Tier 1 center within the Laboratory. Examples of services that will be provided include the following:

Access to Distributed Computing Power. This provides access to the collaboration of additional computational resources that are not dedicated to specific tasks.

Researchers needing a computational environment suitable for algorithm development or code validation will have a facility at their disposal. In addition, within GriPhyN, application interfaces will be developed to enable individuals to perform less general but commonly needed menu and parameter-driven processing requests that are useful to condition data. These functions are being built into the LIGO Laboratory's data analysis environment and they can be ported to GriPhyN. Last, massively parallel, compute-intensive background jobs can be run at Tier 2 centers if they are not time-critical. One example of this is an all-sky pulsar search ("*Pulsar@GriPhyN*" project) using compute cycles in an opportunistic fashion to process the LIGO data in order to perform a blind search (location unknown, spin frequency, and other dynamical parameters unknown). One challenge of this class of activities is the research required for rendering large libraries of analysis software portable within GriPhyN.

Providing and Tracking LIGO Virtual Data. The richness of the LIGO raw data implies there are many transformations that are possible to generate derived data. Examples include Fourier transformations, regressions, bandpass limited filtering and decimation. In general, the full spectrum of signal processing tools may be applied to LIGO datasets. Typically, these types of transformations are computationally costly while they are also generally useful. Every endeavor should be made to store and re-use "popular" transformed datasets. The ability to do this requires a grid infrastructure built upon such elements as data, catalog, reduced data archives, and data mirrors. While subsets of this functionality are called out in the design of the LIGO Laboratory's data analysis system, a general extension of these concepts to a national computing grid is a challenging computer science research topic. This research activity is being pursued by a group of LIGO Scientific Collaboration researchers who have teamed with USC's Information Sciences Institute to identify the needed enhancements to the existing functionality in order to support LIGO applications within GriPhyN.

In summary, the challenges facing LIGO researchers who want to access and use the data archive are in some aspects unique. The detector produces a datacube of raw data that need to be filtered many times in order to produce events (refer to Figure 5). The expected true signals are rare and false alarm rates likely to remain high during the early searches. The same data are processed repeatedly in different ways according to the source type of interest.

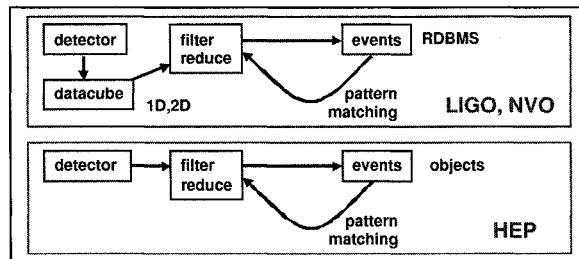


Figure 5. LIGO data analysis challenges involve a multidimensional *datacube* on which a large number of filtering operations are needed in order to generate potential events of interest.

ACKNOWLEDGMENTS

The work reported herein represents the collective efforts of the LIGO team who have been working hard in order to make LIGO become a reality. Without their tremendous spirit and dedication, I would have had little to report. I am indebted to all of them for the materials provided for this contribution. I wish to acknowledge my GriPhyN collaborators and especially Roy Williams of Caltech's Center for Advanced Computing Research (CACR) and Ewa Deelman from USC's Information Sciences Institute (ISI) for the conceptual matter relating to LIGO applications on GriPhyN resources.

The LIGO Laboratory is supported by the National Science Foundation under cooperative agreement PHY-9210038.

REFERENCES

- ¹ <http://www.ligo.caltech.edu>
- ² http://www.ligo.caltech.edu/LIGO_web/lsc/lsc.html
- ³ <http://www.ligo.caltech.edu/docs/T/T970130-D.pdf>
- ⁴ <http://www.ligo.caltech.edu/docs/T/T990101-02.pdf>
- ⁵ http://www.ldas-sw.ligo.caltech.edu/doc_index/user_apis.html
- ⁶ <http://www.ligo-wa.caltech.edu/>
- ⁷ <http://www.ligo-la.caltech.edu/>
- ⁸ <http://space.mit.edu/LIGO/Welcome.html>
- ⁹ <http://blue.ligo-wa.caltech.edu/>
- ¹⁰ <http://www.ldas-sw.ligo.caltech.edu/>
- ¹¹ See P. Brady's contribution to these proceedings for information on the scientific searches that will be conducted with LIGO data by the LIGO Scientific Collaboration. See also Brady's talk: http://www.physics.drexel.edu/events/astro_conference/
- ¹² <http://www.cacr.caltech.edu/resources/HPSS/>
- ¹³ http://www.ligo.caltech.edu/LIGO_web/mou/mou.html
- ¹⁴ <http://www.griphyn.org/>
- ¹⁵ <http://uscms.fnal.gov/>
- ¹⁶ <http://www.usatlas.bnl.gov/>
- ¹⁷ <http://lhc.web.cern.ch/lhc/>
- ¹⁸ <http://www.sdss.org/>
- ¹⁹ http://www.astro.caltech.edu/nvoconf/white_paper.pdf