

## 7. LSC Software Development

### Overview

The LSC science analysis pipeline will be implemented from modular components that are validated and controlled as part of the LIGO/LSC Analysis Library ([LLALLAL](#)). All delivered software will conform to a standard that has been defined jointly by the LIGO Lab and the LSC (ref. LIGO-T9900030).

The LSC Software Coordinator has the principal responsibility for defining and managing the LSC software development effort. Verification and validation of [LLALLAL](#) components will take place at three levels: (i) compliance to standards, (ii) piecewise component tests and (iii) integrated tests of the analysis pipeline through Mock-Data Challenges.

Software, database definitions, and other data representation standards, once adopted, shall be controlled by a Software Configuration Control Board (SCCB) that is chaired by the Software Coordinator. In the case of changes to a standard that has been adopted by the larger GWIC community (e.g., the frame data format), the proposed changes must be approved by the GWIC representatives from the LSC.

### The LIGO/LSC Analysis Library

The [LLALLAL](#) configuration is managed by the LSC Software Coordinator, who coordinates regular releases of the [LLALLAL](#) library with and between LDAS releases. Major releases will be scheduled to coincide with major LDAS releases ( 1 in 2Q1999, 2 in 4Q1999, in 4Q2000 and V1.0 in 4Q2001). These will test LDAS functionality and support the development and testing of analysis pipelines. Intermediate releases will take place quarterly to correct bugs and provide incremental increases in functionality and performance.

All LIGO data analyses involve filtering operations --- either linear or non-linear --- on time series consisting of weak signals in the presence of additive noise. These analyses can all be described as compositions of "atomic" operations on a small number of rigidly structured data types. Typical atomic operations include linear algebra and filtering, signal processing methods and descriptive statistics; typical data types are time series, frequency spectra and linear filter transfer functions. [LLALLAL](#) consists of these atomic operations acting on these structured data types.

All [LLALLAL](#) software development will conform to style specified in T990030, which describes coding rules, documentation standards, software diagnostic and test requirements.

We expect that ~~LLALLAL~~ will evolve and grow with accrued data analysis experience. Changes to ~~LLALLAL~~ will be authorized by ~~a~~ the Software Change Configuration Control Board (the SCCB, introduced above) whose members are appointed by the LSC and the LIGO Lab. Proposed changes will be weighed for relevance, impact on existing systems and resource, and benefits offered.

### Software Verification and Validation

Software verification tests the behavior of individual components. LSC component software verification involves documentation, component tests, and run-time diagnostics. Documentation describes in detail what the component is supposed to do, how it is supposed to do it, error conditions and how they are handled, and accuracy requirements or guarantees. Each ~~LLALLAL~~ software component will include documented test code which tests the component for fault tolerance, accuracy and correctness of implementation as described in the documentation. Finally, each component is required to return at run time a status structure, which reports on the component's current functioning and provides diagnostic information in the event of an error condition. All these components --- the documentation, the test suite, and software status reporting and error handling --- are the responsibility of the LSC member(s) who supply the software component. It is critical to the integrity of the integrated final products that all software modules be tested thoroughly at the module level by individuals who are not themselves the code developers. A thorough procedure of independent validation at the software component level is necessary to preclude the appearance for the first time of accumulated errors downstream at the integration level.

Software Validation tests s that the software components can be integrated into analysis pipelines that can perform ~~that the~~ analyses described in the science goals (Section 1 of this document) with the requisite speed on the target hardware platform (*i.e.*, the on-site and off-site LIGO Beowulfs).

Software system integration is tested at several levels. The ~~LLALLAL~~ has a hierarchical, modular design, with increasingly sophisticated analyses built upon a base of more primitive library calls: *e.g.*, power spectrum estimation by Welch's method involves sub-division of a time series into sequential overlapping components, the generation and application of a window function, discrete Fourier transform of the windowed sub-sequence, term-by-term modulus of the DFT results, and summing and normalizing the resulting frequency series. Each of these operations is a low-level library function that must properly integrate to compute successfully a power spectrum estimate.

At higher levels, system integration, performance and analysis goals are tested through "Mock-Data Challenges" (MDCs). In a MDC data of known character (*e.g.*, noise of known statistical properties possibly superposed with a signal of known character) is

passed through the system, whose response is observed and compared to the expected response. MDCs of increasing sophistication are carried out first on sub-systems and finally on the full system in different configurations.

System integration and performance testing will involve a single LSC/LDAS team that both generates test data and characterizes the system's performance. End-to-end tests of an analysis pipeline will be carried-out single-blind by two teams: one team generates data, which may include signals, and a second team analyses the data and reports back the conclusions. The two teams operate independently, with only the data (but no details of its character) passing between them. The system's ability to handle the analysis goals will be verified statistically by comparing the conclusions reached by the second team with the known character of the input data, generated independently by the first team.

These final MDCs require the ability to generate data streams with the statistical character of LIGO data. This characterization comes from the LSC detector characterization effort, described above, and involves the LIGO End-to-End modeling effort.

MDCs will be performed on an incremental basis. MDCs will be coordinated with each of ~~LAL~~LAL and LDAS major release; additionally, there will be MDCs in between major releases, continually testing the software in different configurations. MDCs are organized by the Software Coordinator in collaboration with the LIGO Laboratory LDAS team.

## 8. Computational Infrastructure and Usage Model

The computational infrastructure required for data analysis is determined by the emerging LIGO/LSC user/usage model. The model is based on a hierarchically arranged infrastructure of computational and storage resources. Three tiers of infrastructure are envisioned and each has a specific role. The tiers include:

~~This model defines several different physical locations where data analysis must be supported and the types of usage supported at each location.~~

~~Data analysis computations will take place at three distinct types of sites:~~

- ~~• IFO Lab Sites (LIGO/WA and LIGO/LA);~~ Tier 1: LIGO Laboratory. The four Laboratory sites (CACR at Caltech, Hanford, Livingston, and MIT) constitute a distributed Tier 1 Center.
- ~~• Non-IFO Lab Sites (CIT and MIT);~~ and Tier 2: Dedicated LSC Sites. It is expected that there will be between 3 and 5 centers established at LSC institutions, other than Caltech and MIT that will be operated and maintained for access by all LSC institutions. In addition, Caltech and MIT will have to comparable computational

resources to support their scientists and scientists at universities in their immediate region.

- ~~Non-Lab LSC Sites~~ **Tier 3: University research group resources.** Individual research groups will have stand-alone computational hardware available to them for local autonomous use. The distinction between Tier 2 and Tier 3 is basically one of scale and the fact that Tier 3 resources are dedicated to the local group, while all or most of resources at a Tier 2 center are subject to LSC scheduling.

~~Non-lab LSC sites may eventually number in the tens.~~

Three broad categories of usage are also defined:

- **Local Processing/Local Data/Low-bandwidth WAN.** This type of usage involves workstation-based analysis and analysis development activities using local data files. Typical activities will involve requesting small (1-10~MB) data files from the archive over the net (e.g., T1, T3, or DS3), or larger ones (1-100~GB) via tape, and analysis using programs running on local workstations. The analysis environment may or may not involve the LDAS software environment. It is expected that a large fraction of the LSC software development and instrument characterization will fall under this model. This mode of operation will typically involve local Tier 2 or 3 resources.

- 
- **Remote Processing/Remote Data/Low-bandwidth WAN.** This model describes development and analysis using significant LSC resources accessed via the net through a browser or X-window interface. A typical example would be a LSC scientist connecting from their home institution to the LDAS system at any of the Tier 1 or Tier 2 Centers. ~~the CIT archive and performing an analysis on a multi-gigabyte data set using the LIGO/CIT Beowulf cluster.~~ Analysis will take place principally within the LDAS software environment. Code validation, Monte Carlo analyses, as well as a large fraction of the computational intensive science analysis are expected to fall under this model. This mode of operation will be particularly intensive during the commissioning phase for the LSC components of the LDAS software, particularly during the first 1-2 years of engineering and science data runs. During this period, the distributed LSC scientific community will be continuously improving their algorithms. Much (but not all) of this activity will require access to increasingly extensive data sets and issues and problems (like false triggers) become more subtle and rare. During this period, it is essential to provide access to data sets at the Tier 1 and 2 centers through X windows, with sufficient bandwidth and low enough latency that the user is not continuously aware that (s)he is a continent away. The computing capacity at the Tier 1 and 2 centers must be adequate to support the anticipated usage

during this phase. The number of simultaneous X window sessions that must be supported by a Tier 2 Center in order to accommodate the LSC computing needs is at present TBD; experience during the engineering runs and early during the Science Run will be used to determine and to tune this parameter.

- **Local Processing/Remote Data/High-bandwidth WAN.** This usage model encompasses analysis on a local workstation or supercomputer using remote data files provided via high-bandwidth (OC-~~4~~3 or greater) from the LIGO archive. Usage under this model is not expected initially; however, it is expected to play an increasingly large role in the future as high-bandwidth network connections and increasingly powerful local computing resource become more common. This mode of use can be supported by Tier 1, Tier 2 (or even Tier 3) resources, depending on the local resources.

### *Usage at sites* Role of the Tier 1 Center

#### *Observatories* IFO Sites

Operation of the interferometers and ~~reduction-storage~~ of ~~data from Level-0-1 to Level-1~~ is the highest priority activity at the IFO sites. In addition, local pipeline analyses will be operated to continuously monitor the strain channel datastreams for a class of astrophysical waveforms having relevance for real-time detection. ~~Correspondingly,~~ The on-site computing infrastructure is oriented toward local-access, and local processing with access from off-site strictly-controlled and given a lower priority. Three LANs will be supported: CDS/GDS, LDAS and general computing.

#### *Non-IFO Lab Sites* Universities The LIGO Data Center at Caltech

~~CIT.~~ The CIT-CACR at Caltech site is home to the LIGO data archive center. Its principal roles ~~is-are~~ to provide access to archival data and support detailed science analysis on the combined ~~IFO-multiple-interferometer~~ data set. The reduction of Level 1 data to Levels 2 and 3 will be performed as data are ingested into the archive Remote user support will include searching the archive and selecting archival data for analysis. Analyses may be carried-out on the LIGO/~~CIT~~-Caltech workstations or Beowulf clusters, or transferred to a remote site via network or tape. The LIGO/~~CIT~~-Caltech LDAS is ~~scoped~~-designed to provide support for five simultaneous high-bandwidth users, assuming a mix of tape and disk data transfers.

MIT will be equipped with a Beowulf cluster for software development and local data analysis. MIT will act as a mirror for the Level-2 data product, in which case it will support use in the Remote Processing/Remote Data/Low-bandwidth WAN mode using

the LDAS software environment. Functionally, MIT will be sized as a Laboratory-supported Tier 2 Center.

The Tier 1 Center will provide those computational resources for Laboratory scientific staff to carry out their research that does not require the analysis pipelines. This includes exploratory R&D on algorithms, analysis techniques, and detector characterization.

### ~~Non-Lab LSC Tier 2 LSC Sites~~Centers

~~Non-Lab LSC sites will operate in either the Local Processing/Local Data/Low-bandwidth WAN or the Remote Processing/Remote Data/Low bandwidth WAN mode. High-bandwidth connection to Lab sites is not currently a requirement; however, efficient remote access to data and LSC computational resources for code validation and data analysis is.~~

~~Some non-Lab LSC sites may obtain or have access to significant computing resources for LIGO analyses. These resources should be available for scheduled use by remote LSC users, operating in the Remote Processing/Remote Data/Low bandwidth WAN mode.~~

For the foreseeable future, LIGO Laboratory's Tier 1 Center will likely not be able to provide all the computational resources that will be required to support the numerous research programs being undertaken by the LSC. For this reason, LIGO Laboratory and several LSC institutions have begun to develop resources that will eventually become Tier 2 Centers for the collaboration. At the time of this writing, the exact number and locations of the LSC Tier 2 institutions has not been defined. Much of This deployment will be carried out as part of the NSF's GriPhyN Project and the LSC Tier 2 Centers will constitute a portion of the US Grid that GriPhyN and other DOE and NSF funded programs are developing. The LSC Tier 2 Centers will number between 3 and 5 in addition to the effective Tier 2 center at MIT.

LIGO Laboratory and the LSC will work together to define and develop the prototypical Tier 2 Center for the collaboration. Subsequent centers will be replicated from the prototype, with allowances for specific configuration details as they may be warranted. The designated Tier 2 Center LSC host institutions will be selected according to a set of criteria that are aimed at maximizing the accessibility and utility of these centers for LIGO scientific research and data analysis across the entire collaboration. Although the specific criteria have not yet been set down, it is expected that they will include, as a minimum, the following elements. The home institutions must:

Be acceptable to the NSF;

Be active in Grid (GriPhyN) research;

Have a technically capable PI who can devote a significant fraction of her/his time to the Tier 2 development and operations effort;

Prototype facilities will need to have resources such as high speed LAN and WAN connections, support staff and possibly existing hardware such as Pentium Linux processors and disk storage.; Much of these basic resources should be preexisting and leverage existing institutional infrastructure, because the institution should have had some experience at managing a facility of this type.

Be located in a geographic location with suitable available network connections, which may affect the selection process.

### *Role for Tier2 Centers*

Tier 2 centers for the LSC shall provide mirrors for critical datasets that are useful to a broad segment of the LSC. These will include, e.g., Level 3 data and, to the greatest extent possible, Level 2 or subsets thereof. In order for these to be truly accessible to the LSC community, adequate bandwidth dedicated to LIGO (initial connectivity of OC3 is sufficient; ultimately, an OC 12 BW has been identified for the GriPhyN Tier 2 Center design) must be available at the Tier 2 Centers in order to ensure that no local bottlenecks exist in the distribution of LIGO data.

Tier 2 centers also represent significant computational capacity that is intended to augment the Tier 1 capacity. It is envisioned that the scale of each Tier 2 center will be comparable to the LIGO Tier 1 Observatory Site components. Unlike the Tier 1 Center, the Tier 2 center resources shall be available for exploratory analysis of datasets or analysis that is not in the mainstream of LSC searches.

In both these regards (data distribution and mirroring, and computation) the Tier 2 centers serve to offload the demand of resources at the Tier 1 Center in those instances where the analysis or data requests can be handled at the Tier 2 Center.

### *Access to Tier 2 Center resources*

Tier 2 Centers serve a role for the entire LSC. In this regard, the resources invested by NSF in these centers must be managed with a stewardship for all LSC member institutions. The operation of the Tier 2 Centers for LSC science will be the responsibility of the host institution. Access to the resources shall be provided by a mechanism within the LSC that ensures equitable distribution of bandwidth and CPU time to all LSC members. One mechanism that will be introduced is a computational resources allocation committee within the LSC that entertains proposals and requests and then distributes time and bandwidth for LSC common resources (Tier 1 and 2) according to the requests.

## Infrastructure requirements

~~The LIGO/LSC usage model determines the network, computing, storage and support personnel at each type of site.~~

~~LIGO Lab IFO and non-IFO sites~~ Tier 1 Center

### LIGO Laboratory - Observatories

The observatories will have the capability of providing a 28-day look back period for data acquired locally. These will be available on a disk farm so that there will be a minimum overhead to access data.

The observatories will be running continuous pipeline analyses on PC Linux clusters. It is estimate that each interferometer will require O[20 GFLOPS] of computational capacity in order to process data at the same rate at which it is acquired. ~~To support data pipeline activities at Laboratory IFO sites, LIGO/LA and LIGO/WA will each have a Beowulf cluster providing a minimum of 20 Gflop/s.~~

The metadata storage capacity for the observatories will accommodate approximately 500GB per interferometer.

These capabilities will be extended during the LIGO I Science Run as experience and resources will permit.

It is envisioned that, by the time of the LIGO I Science Run, the LIGO Laboratory WAN connecting the two observatories with Caltech and MIT shall be able to support OC3 bandwidth. This is sufficient to enable the data acquired at the observatories to be streamed to the archive.

~~LIGO Laboratory -- Universities~~ LIGO Data Archive at CACR

The archive at LIGO/Caltech will have the capability of providing access to at least 360 TB of data (which exceeds the presently envisioned volume of data from the entire LIGO I Science Run). Depending on which data are requested, access time will vary from delays consistent with disk I/O and internet access to delays consistent with the retrieval of archived tape data from the large robotic silo.

The LIGO archive will be augmented with an extensible disk farm designed to accommodate the most commonly accessed LIGO data (e.g., Level 3 and Level 2). The farm will be extended throughout the LIGO I Science Run as data growth dictates and as resources permit.

LIGO/Caltech will have available a number of PC Linux clusters for different purposes. Software development and algorithm development will be supported on two small-scale clusters. Multiple interferometer pipeline analysis will be performed on a

dedicated cluster whose scale is approximately equal to the aggregate capacity of both observatories.

The metadata storage capacity for the Caltech will accommodate approximately 100GB of relational database storage.

These capabilities may be scaled during the LIGO I Science Run as experience and resources will permit.

It is envisioned that, by the time of the LIGO I Science Run, Caltech LIGO Laboratory will be connected through university infrastructure to the internet with an OC48 bandwidth.

~~To support science data analysis the LIGO/CIT site is allocated a Beowulf cluster providing a minimum of 40 Gflop/s. To support data archive activities, the LIGO/CIT site will be equipped with a 100 TB tape robot, 1 TB disk storage, and an OC 3 network connection.~~

LIGO/MIT ~~computing, storage and network connections are TBD~~ will have data storage resources sufficient to provide data mirroring for the Level 3 data set. There will also be a PC Linux cluster of sufficient size to enable algorithm development and data analysis on a sufficient scale to support the research program of the MIT LIGO science staff.

~~Processor improvements are expected to boost computing performance at all sites by a factor of 2-3 over the course of the first two year science run; additionally, disk storage at LIGO/CIT should be increased by a factor of at least 2 by the end of the LIGO-I science run.~~

#### *Non-Lab LSC sites* Tier 2 Centers

~~The LIGO/LSC usage model involves computing and data storage at Non Lab LSC sites. To support science analysis at Non Lab LSC sites we define an LSC minimum workstation configuration:~~ The configuration of the Tier 2 Centers will be defined through an extensive prototype R& D phase under the GriPhyN Program. The centers are expected to have the following characteristics:

- Between 64 and 128 of the latest generation Linux/Intel processors, each with 512-1024 MB RAM, 72 GB or greater disks, 100BT or faster connections to a switch;~~0.5 Gflop/s processor speed;~~
- 50 GB disk;
- TBD OC12 connection to the WAN~~WAN connection;~~
- At least 20 TB of disk storage and accompanying high throughput data servers;
- A small AIT-2 or equivalent technology robotic tape unit for creating datasets for distribution to the LSC;

This ~~workstation~~ system configuration is expected to support a standard software environment, consisting of

- The LDAS software environment, which is supported only on Linux systems;
- a DB2 client for database access; and
- other TBD software.

All computing and data storage ~~Computing~~ infrastructure ~~acquired~~ used for LIGO data analysis ~~substantially beyond the LSC minimum configuration~~ are to be accessible to the entire LSC researchers as ~~LSC communal~~ resources, ~~providing with~~ access modes described under remote-usage models described above. In addition, the Laboratory will provide resources for its science staff at MIT, Caltech, and the observatories, who are participating in analysis and detector-based R&D.

#### LSC-wide Support for Computational Resources

The infrastructure described above constitutes a formidable array of resources that will become available across the collaboration. In order to guarantee open access and efficient usage of these resources, it is necessary that the collaboration as a whole develop a mechanism of support in the form of a distributed help desk system. The burden of operating this system will be shared by all institutional members of the LSC and these terms will be defined in the MOU each institution has with LIGO Laboratory.

## **9. Long Range Program and Anticipated Needs**

The near-term research program ensures that within the limitations of the available manpower and computing resources, LIGO can carry out reasonably sensitive searches for the primary categories of expected sources. The most pressing need is to begin these activities early, so that during the commissioning phase of the LIGO detectors, the data analysis systems can be tested, debugged, and optimized.

In the longer term, LIGO's program will evolve toward increasing detection sensitivity and bandwidth and in the ability to widen the scope of the search. Eventually, when detections are made, the program will transform into a study of the nature of the signals and the properties of their sources.

Elements in a long range program are both in the intellectual development of improved understanding and software and in the exploitation of the improvements in the hardware.

### **1. Development of improved detection algorithms**

- *Improved sensitivity.* Because the LIGO measures the amplitude of the gravitational wave, even small increases in sensitivity result in significant changes in event rate. For example, a 25\% improvement in sensitivity through improved algorithms can increase the event rate by a factor of 2 or make a corresponding change in an upper limit.

- Extended searches. Development of advanced algorithms for binary inspiral and periodic sources will open more of the gravitational wave sky in this branch of the research which is both software and hardware limited. A relevant study is the influence on the data analysis of the improvements at low frequencies being projected for LIGO II which will extend the search at low frequencies by about a decade, to approximately 10Hz.
- Modeling of astrophysical sources. Research into predicting gravitational waveforms of astrophysical sources will continue to play a critical role in the design of search filters. Two examples are: a program to bound the waveforms of the recently hypothesized r-mode sources and NS-BH and BH-BH systems and the completion of the program to determine the waveforms from colliding black holes with orbital and spin angular momentum.
- The inverse problem. Research is required in the development of the computational techniques to fully utilize the dynamical information in the gravitational wave time series in a high signal to noise observation. The detected gravitational waves signals are field amplitudes rather than intensities and retain detailed information of the dynamics at the source. The full inversion will most likely require both position and polarization information determined from detections at multiple sites.
- Improved visualization techniques Automated pattern recognition as has been developed for speech recognition and oceanographic research may provide new methods to diagnose the detectors as well as to search for unmodeled gravitational wave sources.

## 2. Improved hardware

- Broader band inspiral binary systems. Searches for inspiraling binary systems over a wider range of system masses and spins would be enabled by faster computation. The amount of computation power required grows as a rapid power of the lower-mass limit of the search: currently LIGO's data analysis facilities are scoped to carry out a search down to 1 solar mass (10 Gflops). A search for objects to a lower mass limit (0.1 solar mass) would require ~1Tflops.
- Unprejudiced search for periodic sources ~1Tflops computer could carry out an all sky searches for CW/pulsar signals to within about a factor of three of the limit of instrument sensitivity. Additional computational power would make it possible to approach the instrument sensitivity, and also consider larger ranges of spin-down parameters.

These longer-term activities should develop naturally out of the LSC's near-term research program but will require a greater concentration of effort in software and theoretical development. A well placed investment is in the support of additional scientists interested in the software and data analysis of gravitational wave astrophysics.

Improvements in computer hardware and the bandwidth of communications networks will enhance the effectiveness of the LSC data analysis activities. The rapidly-decreasing price of commodity computer hardware and the concurrent development of very cost-effective parallel computing architectures such as Beowulf systems should make it feasible for different LSC groups to make timely and effective contributions to the overall computing infrastructure needed to analyze LIGO data. These efforts will benefit from development efforts in other fields to create software and hardware configurations that can handle these enormous data sets. In common with some of the data from other fields, (most notably, high-energy physics) much of LIGO data has an *event independence* which allows the data to be efficiently analyzed in parallel. This suggests that the databases and tools which are used or might be developed for these fields have substantial overlap with GW detection.

Because LIGO's data rates are fixed at around 15 Mbytes/sec, and the speed of the national and international networking infrastructure continues to improve exponentially, easy access to LIGO data should become available in the long term. But the next five years are crucial ones, and during this time the LSC needs to make a continued effort to improve access to the data and resources. For example, by the end of the first science run it may be possible to put all the LIGO data onto spinning media, and make it available anywhere within the US, at reasonable cost.

These improvements in networking and facilities will enable another critical step in the field's evolution by the full use of the international network of gravitational wave detectors (GEO, VIRGO, TAMA, ACIGA, bar detectors) to gain position and polarization information on the observed sources. Improved networks will also enhance the ability of the gravitational wave detectors to provide a trigger to other astrophysical observations after an impulsive event has been detected. A model for this is the Supernova Neutrino Network (SNNET) which has been set up to provide alerts if neutrino bursts associated with supernovae are detected.

We strongly endorse the LIGO visitor's program. This has proved to be an effective way of reaching out for expertise and assistance from the scientific and engineering community. Data analysis problems comparable to those encountered in gravitational wave detection occur in several research areas such as speech analysis, oceanography and other branches of observational astrophysics. The visitor's program is an effective way to bring individuals who have developed particular methods and abilities into close contact with the LIGO detectors and data.

It is our expectation that gravitational wave observations will become a standard part of astrophysical measurements in the next decade and add new and complementary insight into the nature of the universe. The most promising direction in which the field

will develop is not easy to predict. It is, however well known, that those best prepared will be most likely to discover something new and enduring.