

The LIGO Program Advisory Committee

The Eleventh Program Advisory Committee
(PAC11) Meeting
LIGO Hanford Observatory, Hanford,
Washington
November 29th - 30th, 2001

REPORT OF THE LIGO PROGRAM ADVISORY COMMITTEE

Data Analysis Section of Report
January 29, 2002

REPORT OF THE LIGO PROGRAM ADVISORY COMMITTEE
Meeting of November 29-30, 2001

LSC data analysis and LIGO Laboratory Data Analysis

These efforts continue to make good progress and appear to be on track to provide the needed analysis as the interferometers come on line. We address some potential problems below. The program Advisory Committee heard reports by Wiseman for the LSC Data Analysis activities, and by Blackburn on the LIGO Lab effort.

The LIGO Science Community (LSC) effort reported by Wiseman is a search-oriented software system that accepts LIGO "frames" and produces data analysis results. This LSC effort has established, and is to a substantial extent achieving compliance with, a series of software structure and validation requirements.

Definitional requirements: The scientific search engine developed by the LSC must be LSC Algorithm Library Compliant (LAL Compliant). The standard describes:

- Reusable data structures
- Fault tracking
- Document requirements
- A prescribed interface to the LIGO Data Analysis System (LDAS); LIGO data products in become LIGO data products out
- Validation Requirements: The entire analysis pipeline for each functionality must be tested in a Mock Data Challenge, with blind injection of signals.

The LSC effort supports the code development through a defined infrastructure, utilizing CVS (the UNIX Control Version System) with a specified software change control board, and by maintaining bug tracking within the code.

The LSC computing capability contributes substantially to the overall LIGO capability. The LSC computing model is a distributed one that is not tightly coupled to the LIGO Laboratory computational infrastructure.

This situation has arisen because of the opportunity to combine research thrusts in The NSF, for instance the interest in Grid computing, with an application area; in this case with the data analysis for LIGO. This has resulted in a substantially greater resource for LIGO analysis than would otherwise be the case. It is unlike the "classic" high-energy physics model in which case most of the analysis resources are concentrated at the laboratory, there is a tight control of configuration and software among the different tiers of data analysis, and the total second tier resource does not exceed that of the laboratory itself. In the contrasting LIGO case, there is substantial computing power represented by the two Tier II Centers (The University of Wisconsin at Milwaukee (UWM) and Penn State (PSU)), each roughly equivalent to the LIGO Laboratory resource. The data handling and hardware approaches at the Tier II Centers are, however, not tightly coupled to the LIGO Laboratory computing centers.

In fact, for similar reasons the new generation of high-energy physics experiments, as well as other fields like astronomy, are also moving to this "grid" model of distributed wide area computing. It is important to note that LIGO will be breaking ground in encountering the management and sociology problems involved, because LIGO may well be the first to implement this new approach with real data and real physicists trying to produce a scientific result expeditiously. The key issues that must be addressed are:

- Allocation of the distributed resources in the most scientifically effective manner;
- Coordinated management of the distributed hardware and personnel resources so they can be most efficiently used; and
- Providing sufficient system management personnel throughout the distributed environment so that problems can be addressed responsively.

We discuss these points in detail below.

The future model at the Tier II centers is to use the Grid computing functionality to closely couple these resources. However, currently UWM and PSU maintain versions of the LIGO data analysis software that tracks the versions maintained in the LIGO laboratory, but with some version lag. In contrast to the LIGO Laboratory software that is synchronized across the Laboratory sites (Livingston, Hanford, Caltech, MIT) on a one-to-two day cycle, the Tier II centers cannot devote the manpower to such a synchronization cycle. In substantial part the difficulty in maintaining synchronization arises from the heterogeneity of the hardware at the Tier II Centers, and because of the staffing/funding model at those centers, which makes them "hardware heavy and manpower light".

The PAC considers it essential to maintain compatibility among the LIGO Laboratory and the Tier II efforts. To this end, we recommend the formation of an Analysis Resource Coordinating Group, with membership from the LIGO lab, from the LSC Tier II centers, and from the LSC "users". The charge to this small committee will be to maintain interface standards and to certify compliance to these standards, to maintain compatibility among sites. This Coordinating Group should provide insight and direction on resource allocation and priorities, manage a bug response effort, and advise on the development and performance of software "immediate response teams" for essential software maintenance. The coordinating committee would also be concerned with allocation of resources: for instance, computer usage and priorities, and would advise on the allocation of resources in the Grid computing environment. It would also advise on the overall computational resource development strategies.

The LIGO Laboratory effort is directed toward a tightly coordinated development product. The in-house LIGO Laboratory effort is making good progress in reacting to and correcting deficiencies as they are found. A large fraction of these software deficiencies have been found as the result of the Mock Data Challenges. There currently appears to be an appropriate balance between coding and problem fixes. Both this effort and the LSC effort appear to be very responsive to user requirements. However, as data begin to enter the stream, there will be a dramatic increase in help requests and problem reports. The current staffing will not be able to

cover this demand. The PAC notes that additional "system manager" staffing appears essential at the LSC Tier II sites, and at the LIGO Laboratory sites.

The LSC plans a system of rotating help desk assignments among LSC sites, to help LSC users access and process data. The PAC recommends appointment of a specific person in the LIGO Lab, and in each of the LSC Tier II development teams to be responsible to back up the help desk functionality, with immediate action on code bugs and deficiencies. The oversight of this function may naturally devolve to the Analysis Resource Coordinating Group mentioned above.

As code has been developed, Mock Data Challenges have been carried out, with accomplishment of inchpebbles (rather than milestones). One example is the Inspiral Mock Data Challenge, carried out in May 2001. LSC and LIGO Laboratory recognize the need for an end-to-end data handling demonstration. This will presumably be incorporated in the E7 engineering run, or the S1 science run. We note for instance that the Inspiral Search Mock Data challenge could not have achieved the physical (scientific) result of finding an upper limit to the inspiral rate, without the ability to inject software signals into the data stream. The PAC reiterates that complete physical verification of the interferometer-detector-analysis system is essential. This requires that physical signals be inserted at the interferometer, by moving mirrors in a controlled way, and their extraction as waveforms be demonstrated. G. Sanders indicated that this capability is available. Such physical signals inserted while the detectors are locked should be detected and will form the real basis for determining the signal sensitivity of the detector. For instance, a plan could consist of inserting 1000-second chirps at a range of signal amplitudes from very large to very small; initially the test analysis would be carried out with known amplitude and time of the signal; later runs should be done blind, with the detectors locked and in coincidence.

In order to avoid delays in scientific output, it is important to identify problems in the scientific code, and in the physical detector as early in the process as possible. In addition to the "inch pebbles" noted above, which focus on the software infrastructure rather than the scientific code, we suggest that coupled "milestone" level data challenges should be carried out in data analysis prior to, and during the Science and the Engineering (Sx and Ex) runs. Just prior to each early major scientific (or engineering) run, the first of these milestones would use simulated data to try to catch as many bugs as possible in the scientific code before those bugs can cause delays in analysis of real data. The second milestone would consist of detecting real physical signals that are inserted (moving the mirrors, as noted above) and then detected during the Ex and Sx runs, as final proof of the performance of the detector.