Use of LDAS in LIGO Science

NSF Review of LIGO Operations Proposal for FY2002 – 2006

LIGO Hanford Observatory
Hanford, Washington
February 26, 2001

James Kent Blackburn
LIGO Laboratory, Caltech

LIGO-G010050-00-E

LDAS Science Goals

- The principle purpose of LDAS is to support the scientific goals of LIGO data analysis by:
  - Providing data storage and archival with user interfaces allowing general and targeted access to data of interest to the scientific community, including reduced data sets.
  - Providing network interfaces used to communicate user requests to issue LDAS jobs and retrieve data products associated with those requests and to monitor these jobs as they execute on the LDAS system.
  - Providing a framework for conducting scientific “pipeline” analyses needed to characterize the LIGO detectors and to perform astrophysical searches.
  - Providing LIGO database for cataloging metadata and its relationships for the purpose of detector and event characterization.
  - Providing data-mining tools over the LIGO database useful in characterizing, classification, and validation of scientific results.

LDAS Science Scope

- Detector & Environment
  - Detector Diagnostics
    - Ingest triggers/vetoes and spectra generated in the Diagnostic Monitor Tool (DMT) into the LIGO database.
    - Ingest transfer functions and detector state information in LIGO database.
  - Detector Characterization
    - Measure narrow spectral features (power lines, violin modes, internal mirror modes, etc), calculate IFO noise and strain sensitivities.
      - LDAS's dataConditionAPI, equipped with significant signal processing functionality, will play a central role.
    - Store instrumental and terrestrial signatures and its associated metadata in the LIGO database.

LDAS Science Scope (cont)

- Relativistic Gravitation & Gravitational Wave Astronomy
  - Astrophysical Searches (Detection and Upper-Limits)
    - *Using simulation techniques, upper limit studies on event rates and signal strengths and likelyhood analysis will be carried for all searches – detection not a pre-requisite to conducting scientific studies.*
  - Binary Inspiral/Merger/Ringdown
    - Carry out these searches using the LDAS pipelines both on site and off using hierarchical templated optimal filtering and other methods such as FCT.
    - Conduct in-depth searches in archival data based on additional knowledge and insight such as neutrino or optical observations.
    - Perform coherent network analysis using data collected from multiple detectors.
  - Un-modeled Sources
    - Carry out searches for waveforms which can not be calculated (supernovae, binary merger phase, serendipidous sources) using time-frequency, power monitoring, time-domain thresholding, pulse-matching, and two site correlation.
    - Compare LIGO event lists, triggers and vetoes between the LIGO interferometers *and* GEO and VIRGO.
    - Perform event list comparisons with other astrophysical observations (gamma-ray burst, neutrino, optical, x-ray, etc).
  - Continuous Wave and Pulsar Signals
    - Carry out directed searches for continuous wave source using position and phase catalogs for known pulsars.
    - Searches for unknown pulsars at the limit of the interferometers' sensitivity is not feasible. However, such searches will be carried out constrained by available computational power. This will very likely involve non-LDAS components in an effort to gain access to the most powerful computers available.
  - Stochastic Background
    - Characterize to the limits of sensitivity, these weak incoherent background sources through inter-detector correlation in conjunction with environmental signals.

LDAS Science Scope (cont)

- Data Distribution and Archival
  - Data Archival
    - Data will be collected in frames at the sites and shipped to the main archive at CIT.
    - Data will be inserted into the archive where metadata about the data will simultaneously be generated and stored.
    - Four data sets will characterize the data acquired at the sites and later stored in the archive:
      - Level 0: Full IFO Data Streams (exists for ~16 hours at the sites),
      - Level 1: Archived Reduced Data Set (all channels deemed important for archive; ~10% of full data stream),
      - Level 2: IFO strain plus data quality channels (basic IFO strain plus a variety of quality channels),
      - Level 3: Whitened GW strain data (best estimate of whitened gravitational wave strain in an ~500 Hz band with metadata).
  - Data Mining
    - Construct coincidence observation between triggers, vetoes, instrumental/terrestrial/astrophysical events.
    - Produce routine summary reports on rates for triggers, vetoes, and all classes of events.
  - Data Distribution
    - Provide collaboration wide access to LIGO frames, reduced data sets, metadata, and other data products associated with data analysis.
- Computation Facilities
  - Data analysis and computation will take place at three distinct types of sites:
    - IFO Lab Sites (Hanford and Livingston),
    - Non-IFO Lab Sites (CIT and MIT),
    - Non-Lab LSC Sites (Currently ANU, PSU, UWM, UTB are running LDAS).

LDAS Science Capabilities

- LDAS is a distributed computing environment mixing remote process control (RPC) with parallel message passing (MPI).
  - A TCL scripting layer is used to manage processing dynamics.
  - An underlying C++ extension to TCL provides the computational power.
- LDAS provides a framework for conducting scientific studies of LIGO data primarily though the concept of a “pipeline”.
  - Data flows through the pipeline from one LDAS API to the next, being down-selected, pre-conditioned, searched over for interesting signals and discriminated for likelihood, before finally being committed to the database and presented to the user as reduced data products.
- LDAS, with its data archive and network interfaces, allows users to perform “pipeline” analyses on any data set ever recorded on any system available on the network authorized for the types of studies being performed.

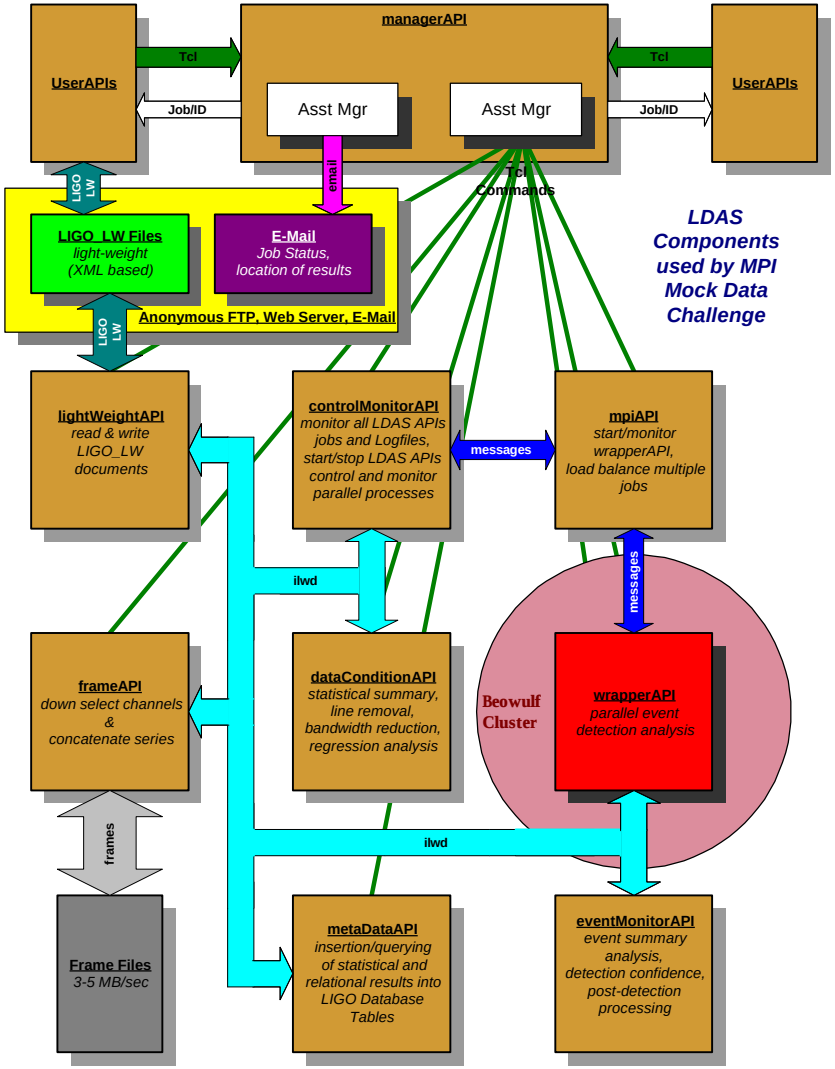
LDAS Science Capabilities (cont)

- dataConditionAPI:
  - The dataConditionAPI is the first component of the LDAS “pipeline” used to actually carry out pre-conditioning and calculations of data products
  - It uses a token language called “actions” which are interpreted by a highly efficient C++ parser, providing a signal processing environment much like matlab.
  - Currently the dataConditionAPI supports statistics, FFTs, PSDs, linear filtering, heterodyning, and basic mathematics. It will also include line removal, linear regression, time-frequency and other higher level “actions”.
  - The dataConditionAPI primarily prepares data for the wrapperAPI where extensive searches using LAL based algorithms are applied.
  - It also can produce intermediate results which can be stored in the database or distributed to users.
  - The dataConditionAPI is based on threaded function calls and is intended to run on SMP platforms such as a Sun 450 or 4/8 way Intel PIII Xeons servers.

LDAS Science Capabilities (cont)

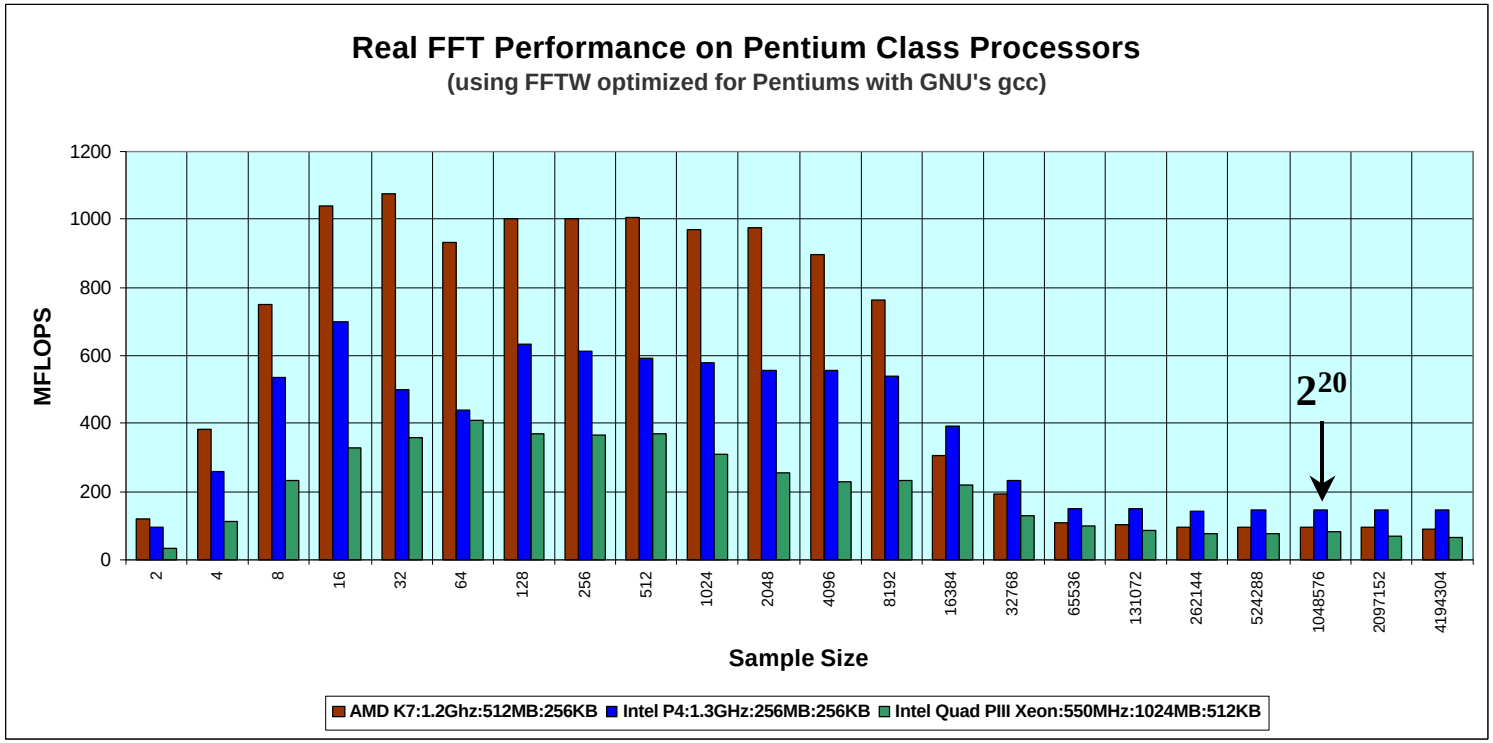
- wrapperAPI:
  - The wrapperAPI performs parallel/distributed computation based on the MPI message passing interface standard.
  - It is the only component of LDAS to run on the Beowulf Cluster.
  - LSC developed search algorithms are incorporated into the wrapperAPI using the concept of dynamically loaded shared object libraries. Each search algorithm developed by the LSC uses a standard interface called the LALwrapper to guarantee communications between LDAS and LAL code.
  - The wrapperAPI is in constant communications with LDAS through a communications port with the mpiAPI. This allows for status and health monitoring, as well as possible load balancing with competing parallel jobs on the LDAS Beowulf Cluster.
  - Three types of data products are generated by the combination of the wrapperAPI and the LAL search codes:
    - Event data destine for the LIGO database
    - State data which will be reused by a later wrapperAPI/LAL search code
    - User data which will assist the developer's of the search code in understanding the behavior of their codes when challenged by LIGO data sets.
  - Current plans are to have the LIGO LDAS Beowulf Cluster configurations with [48, 96, 144] nodes on which to support multiple wrapperAPI/LAL search codes at [LLO, LHO, CIT].

Simple Inspiral Pipeline Example



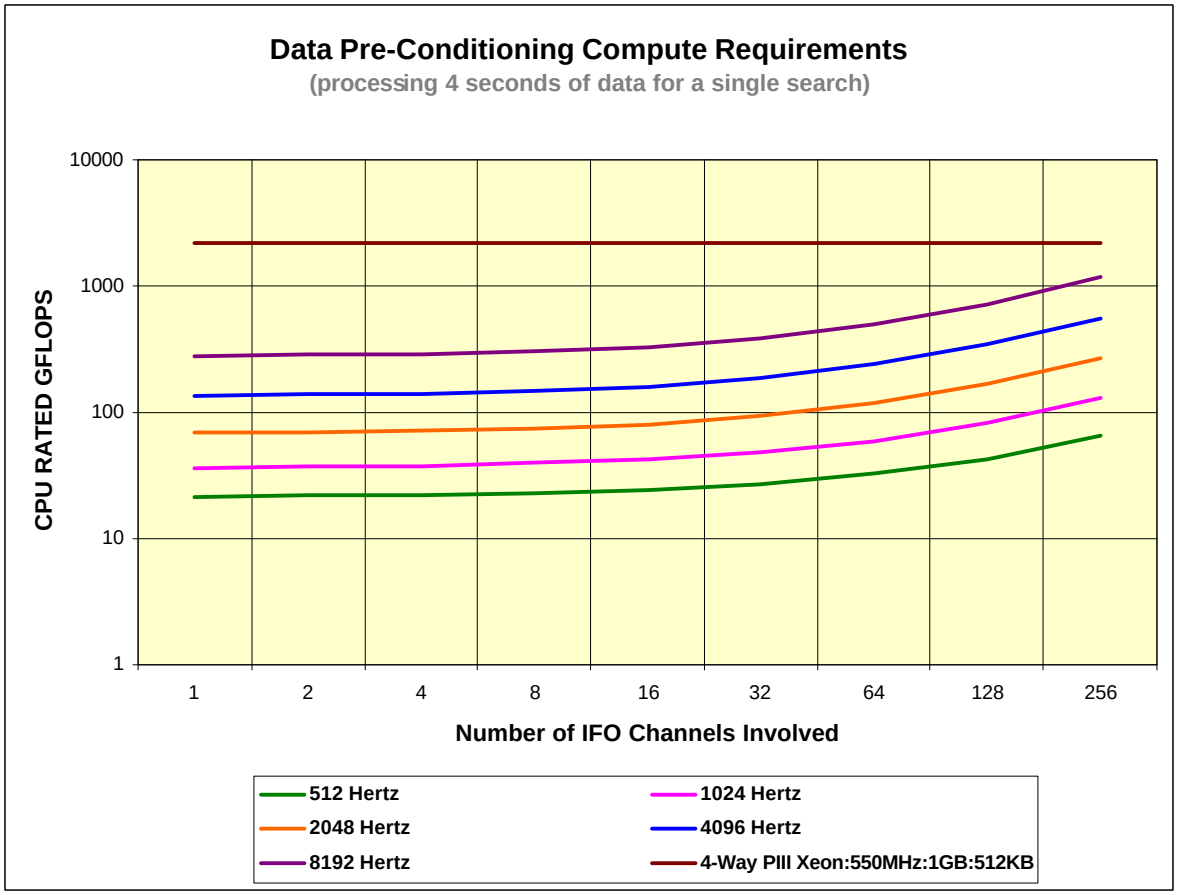
- **UserAPI** connects to **managerAPI** and issues a job request to start a pipeline analysis (e.g. inspiral search).
- **managerAPI** assigns an *assistant manager* process to control process and data flow through system (*green lines*).
- **frameAPI** reads *frames* into the LDAS system, down-selects to GW channel, quality channels, small set of PEM/control channels, and response function; concatenates data into requested segment size; and sends data to the **dataConditionAPI**.
- **dataConditionAPI** pre-conditions data by removing lines, decimating to interesting bandwidth (e.g. 1 KHz), computes average power spectrum, downsamples and bandpasses auxillary channels, etc.; calculates statistics on channels and sends results to **metaDataAPI** for insertion into database; constructs the composite dataset used by the search and forwards this to the **wrapperAPI** under the supervision of the **mpiAPI**.
- **wrapperAPI** performs hierarchical template based search using dynamically loaded shared object provided by LSC/LAL; results from the search are passed to the **eventMonitorAPI** for post processing and redirecting to the database via the **metaDataAPI**, storage of state information for use in further pipeline jobs, and posting of ancillary information under the *anonymous FTP/Webserver* for users to pick up. Concurrent with these activities the **wrapperAPI** and **mpiAPI** are negotiating requests and status information used to trace and possibly load balance the search in the shared environment of the LDAS *Beowulf Cluster*.

LDAS Compute Node Performance



- **Pipeline analysis of LIGO data is computationally dominated by the cost of Discrete Fourier Transforms.**
  - ***Example: Non-Hierarchical Binary Inspiral Search spends an average of 92% of CPU cycles performing FFT.***
- **Most practical/efficient data segment size centered around 2^{20} points for Binary Inspiral Search.**
- **LDAS plans to use SMP Pentium III Xeon processors for pre-conditioning of data and single CPU Pentium 4 or AMD K7 processors for Beowulf Cluster Nodes (*this may change - naturally*).**
- **A 1.3 GHz Pentium 4 with its 400 MHz bus delivers best performance near 2^{20} point data segment sizes: (95, 150, 75) GFLOPS.**

LDAS Data Pre-Conditioning



Normalized GFLOPS to PIII Xeon Clock, (e.g. 20% FFT efficiency).

Computational costs measured using complex spreadsheet.

Pre-Conditioning steps here involved:

- **Data Drop-Out Correction on 10% of data.**
- **Line Removal of 64 lines using multi-taper method with 7 slepians in GW channel.**
- **Calibration of GW channel.**
- **Decimation of all channel data to desired Nyquist Frequency using 8 taps in linear filter.**
- **Linear Regression to remove channel cross talk using all involved signals.**

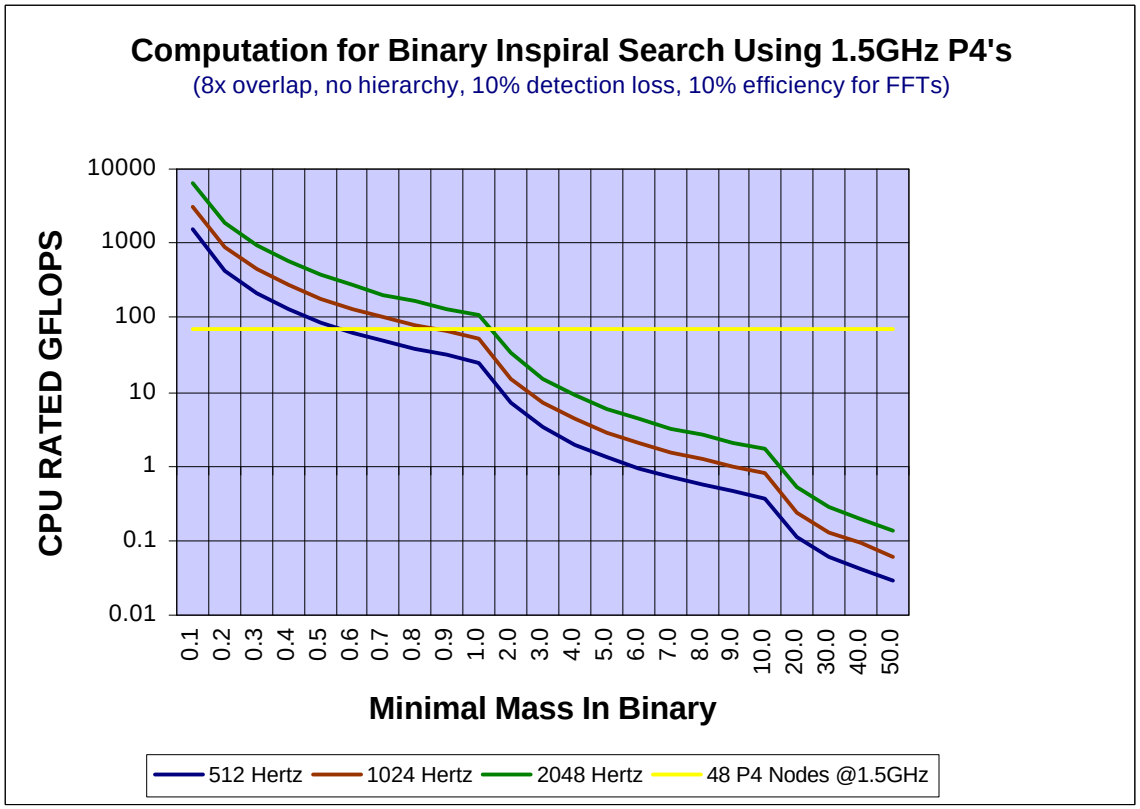
This is estimated on a 4 second stretches of data.

Longer data segments may be necessary (more data samples per pre-conditioning cycle can be estimated from curves) .

The roughly 8 unique searches expected to be active in LDAS could require unique data conditioning for a total of 8x this computational requirement.

4 or 8 way 1 GHz Pentium III Xeons with 1-2MB L2 cache should suffice for all searches.

Computation of Templated Binary Inspiral Search



- **Normalized GFLOPS to P4 Clock, (e.g. 10% FFT efficiency).**
- **Computational costs measured using complex spreadsheet developed in collaboration with Ben Owen.**
- **Non-Hierarchical Search for NS-NS Binary Inspiral ($1.4M_S$, $1.4M_S$) requires {15, 32, 67} GFLOPS.**
- **Calculations by Sam Finn indicate that 512 Hertz bandwidth is adequate for detections process (blue curve).**
- **Hierarchical search strategies expected to decrease shown computational cost by factors of 5x to 30x over these curves.**
- **Same spreadsheet estimates the cost of Templated Normal Mode Ringdown Search to be roughly 10% of this computational costly.**
- **All other searches (excluding so called all-sky pulsar search) are expected to be single node compute problems.**

Growth of LDAS Community

- LDAS presently installed on LIGO hardware systems at CIT(2), Hanford, and Livingston; MIT and a CIT production LDAS will be developed within the year.
- In addition LDAS is installed on hardware at PSU, UTB, UWM, ANU, and UF; Others within the LSC will certainly follow in these leads.
- LDAS has been successfully installed on low-end laptops, though no database functionality was enabled. Full installations of LDAS on high-end laptops should be possible.
- LDAS collaborating with designers of GriPhyn Project to assure the necessary interfaces are available to integrate these two computational worlds together.
- MDC’s have been an extremely useful environment for distributing the knowledge and experience for using LDAS among the LSC.

Closing Remarks

- LDAS has participated in all Engineering Runs, having successfully recorded DMT triggers in both runs (over 70,000 in E2).
  - Plans for E3 include the statistical and cross spectral density characterizations of IFO channels.
- LDAS has seen two Mock Data Challenges which have exercised much of the core scientific components for LIGO data analysis:
  - dataConditionAPI MDC; August 2000
  - wrapperAPI MDC; January, 2001
- Participation in the development of LDAS has significantly increased through the MDC process.
- Future MDCs will test the database, pipeline analysis, and the archive.
- These are strong indicators that LDAS is on the threshold of providing scientific results for LIGO data.