Status of LIGO Data Analysis System

Albert Lazzarini
LIGO Laboratory at Caltech

LSC Meeting
14 August 2001
Hanford, WA

Outline

- LDAS software
  - » Systems preparation
  - » MDC plans, status
- LDAS Hardware
  - » Systems preparation
  - » Archive plans
- Issues, challenges for E6 and UL analysis
- Grid computing & LIGO Data Analysis opportunities, challenges

LDAS Update for LSC 2001/08

Software

- LDAS code base is ~90% complete.
- ~90% of work on 14 of 16 APIs has moved into bug-fixing, maintenance, and enhancement
  - Added functionality beyond originally specified baseline
- Final two APIs (disk cache, data ingestion) still pending a final decision on archive (*refer to HW status section*)
  - Will develop an LDAS interface to the APIs that come with chosen solution
- Focus is on upcoming MDCs and E6 run

LDAS Update for LSC 2001/08

Software

- Data Conditioning API status
  - Resampling of data by rational fraction complete, with anti-aliasing either built-in or user-specified
  - PSD estimator
    - Non-parametric - available since first MDC
    - Parametric using AR model is 50% complete -on hold due to other priorities
  - Whitening by AR implemented, partially tested
  - Line removal
    - Kalman filter for thermally excited lines implemented, tested on simulated data, undergoing further tests for MDC, E6.
    - Output Error time-domain estimation of cross-channel xfer functions implemented, tested on simulated data - can be used to remove power mains
  - Arbitrary linear time domain filtering complete
  - Wavelet decomposition
    - Simple PSU model nearly complete
    - More sophisticated UFI model being prepared

LDAS Update for LSC 2001/08

Software

- Data Conditioning API status
  - Database insertion/retrieval model functional requirements being identified, specified by Finn, Blackburn, Shawhan, Burst Group task leads, Riles
  - State preservation (for recurrent filtering between jobs)
    - Requirements, specification under development by Blackburn, Finn, Maros
  - Miscellaneous recent functionality:
    - Random number generation(uniform, normal deviates)
    - Limited signal generation capabilities for testing
    - Data type conversion (e.g., LIGO time series from data sequences)

- ***Sergei Klimenko (UFl), Michael Landry (LHO) have joined the development effort***

LDAS Update for LSC 2001/08

- Software releases
  - » Alpha - 0.0.XX (April 1999 - September 2001) - ONGOING
  - » Beta Release LDAS 0.1.0 (November 2001) - after last MDC inchpebble - in time for E6
  - » First Official LDAS Release 1.0.0 - after final MDC (Integrated archive test); in preparation for the LIGO I Science Run

LDAS Update for LSC 2001/08

Software:

- Completed MDCs
  - » Dataconditioning MDC (stand-alone testing, validation)
    - – August 2000
  - » MPI MDC (stand-alone testing, validation)
    - – January 2001
  - » MetadataAPI MDC (stand-alone testing, validation)
    - – April 2000
  - » Inspiral MPC (multiple shared objects running together, use of data from dataconditioning API, databases)
    - – May 2001

LDAS Update for LSC 2001/08

Software:

- Pending MDCs
  - » Joint Stochastic, Burst MDC
    - – 4-10 September @ MIT-LDAS
  - » CW MDC - ***LDAS-based*** approaches - ***originally proposed UL group scope***
    - – Long-stretch coherent GW search directed at known CW sources
      - Date, site TBD - (led by LIGO-LHO)
  - » CW MDC - ***non-LDAS-based*** approaches - ***added UL scope***
    - – Heterodyne pulsar search - 1st week of October, Glascow - (led by GEO)
    - – Hierarchical Hough transform - 1st week of October, AEI/Golm - (led by GEO)
    - – Incoherent stack - end of October, site TBD - *(led by U. Michigan)
  - » Archive MDC
    - – Pending archive decision (ref. hardware status section)

LDAS Update for LSC 2001/08

Hardware - Phase I

- *Phase I: COMPLETE*
  - *28 TB of RAID disk systems for Observatories, Caltech, MIT*
    - *Sites will have 1 month look-back capacity on spinning media @3 MB/s per interferometer*
    - *Caltech HPSS disk cache*
    - *MIT disk cache*
  - *6000 slot (500+TB capacity) robotic silo for HPSS at Caltech*
  - *PCs, servers for E6 run*

LDAS Update for LSC 2001/08

Hardware (Phase I, UL run):

- Prototype clusters will be up and running at all LIGO Lab. Sites by end of August
  - LHO - 16 P IV nodes, 512 MB RAM
  - LLO - 16 AMD K-7 nodes, 512 MB RAM
  - MIT - 16 AMD K-& nodes, 1024 MB RAM
  - CIT- dev 16 P III nodes, 512 MB RAM
  - CIT-test 8 dual CPU PIII nodes, 256 MB RAM - (very old)
  - Reference linux platform: RH7.1;
  - Providing reference rsync /ldcg LDAS software environment for LSC sites to mirror after gcc-3.x.x
- 28 TB of RAID (26 T3s@LHO, 14 T3s@LLO, 1T3@MIT, 8 T3s@CIT) - ***exported to CDS/DMT/Dataviewer at sites***
- Database, frame servers at each site (E450 SUN servers)
- Data conditioning servers at each site (4xCPU Xeon PIII)
- All LDAS networks converted to gigabit ethernet (1000 Mbps)

LDAS Update for LSC 2001/08

Hardware - Phase II

- After final beowulf, archive software benchmark tests
  - Data movers for archive (support 5 continuous streams of data at full bandwidth)
    - Plan: SUN Daktari line of ultra III processor 4x or 8X SMP server(s) to connect to tape drives from Silo
  - HPSS is baseline
  - Recent Sun/LSC SamQFS alternative being evaluated
    - SamQFS, has many desirable features, not as scalable as HPSS
    - Native unix file system; tar format for data archive
    - Uses unix inode disk structure to store metadata - FAST
    - Less robust against HW failure than HPSS* -> reason why it is FAST
    - Can be used AT ***ALL LIGO SITES TO FEDERATE/INDEX DATA***
  - PCs for main clusters at Caltech, MIT, LLO, LHO (total of ~400 units)

LDAS Update for LSC 2001/08

Hardware - Phase III

- Complete archive for Science run
  - STK model 9940A (60 GB), 9940B (200GB) tape drives for silo
  - Data tapes for Science Run
  - Start build-up of large disk farm in front of tape archive for data analysis
    - Continue to grow farm throughout Science run
    - Keep up with data growth
    - Target keeping all commonly used, needed data on disk
    - Use tape archive for backup, large data dumps to disk caches

- ***Phases II, III will occur after UL run and in time for the LIGO I Science Run***

Issues

- People are task-saturated
- Hardware resources are needed for simultaneous code development & validation by various groups:
  - » LIGO has 2 systems at Caltech that are busy ALL the TIME
  - » MIT will soon be available as 2nd Lab. resource (for MDC)
  - » Observatories are dedicated for data taking and not accessible for code development
  - » PSU, UTB, UWM developing LSC-based LDAS installations
- In order to be useful, all configurations must be mirrored in order to make installations, code development tractable

LIGO/LSC Longer Term Computing Opportunites

- LIGO laboratory:
  - » ***$53M Distributed Terascale Facility (DTF) award announced tat NSF his past week***
    - – Consortium composed of NCSA, ANL, SDSC, Caltech/CACR
  - » **Caltech /CACR component $3M**
    - – **0.4 teraflops McKinley linux clustrer at Caltech, HPSS silo upgrade at Caltech**
    - – **Resources available to LSC through LIGO-CACR MOU**
  - » Similar compute resources at ANL
  - » Several-fold greater resources at SDSC and NCSA
  - » **DTF use model includes redundant LIGO main archive (re-striped) data mirroring at other DTF sites - 100 gbps access across the DTF fabric**

LIGO/LSC Longer Term Computing Opportunites

- LSC Resources
  - » GriPhyN -- ITR2000 -- SW/CS component
    - – Caltech/UWM/UTB(dedicated to GriphyN-wide outreach) participation (0.5FTE/1FTE/1FTE)
    - – Use of grid software tools, infrastructure to support data replication, access in a robust, secure environment
  - » international Virtual Data Grid Laboratory (iVDGL) -- ITR2001
    - – ***Unofficial notification of pending award $2.110M for LSC***
      - ***New PSU Tier 2 center for LIGO data analysis***
      - ***Upgrade/maintenance for UWM Tier 2 center (existing MRI resource)***
      - ***Personnel for Tier 2 ctrs.***
    - – EU involvement in distributed computing (Italy/UK)
      - iVDGL can serve as a model for Virgo-LIGO data exchange, joint data analysis

LIGO-GriPhyN Priorities

1. Distribution/Replication of Computing/Data with Security

   wrap LDAS with GriPhyN

   pilot at CACR-CIT, LIGO-CIT, ISI-USC, UWM

2. International Cooperation

   Coincidence Studies LIGO (US) & Virgo (FR/IT)

   pilot at CACR-CIT, LIGO-LHO, LIGO-LLO, Virgo-Pisa

Grid LDAS Architecture