# LIGO Data Analysis

# Update

Albert Lazzarini

GWDAW
19-21 November 1998

Pennsylvania State University

State College, PA

# LDAS Status Summary

- Software design essentially complete
  - PDR scheduled for January.
  - Software components specification for Application Programmer Interfaces (APIs) complete
  - Lightweight data type specification released
  - Selection of database management system is pending
- Completed a GenericAPI template, from which specific APIs will be extended
- Prototyping activities under way in several important areas:
  - Software module development
  - Data distribution using web tools
  - Interprocess communications, data transmission
  - Data flow for (directed) pulsar searches
- Hardware conceptual design complete
  - On-line systems at observatories
  - Off-line system at data repository (CACR/Caltech)
  - Wide area network for inter-site connectivity

# LDAS Status Summary

- Recommendations from recent NSF panel review of LDAS:
  - LIGO and LSC should organize a series of (end-to-end) mock data runs using either prototype data or simulated data in the LDAS environment to validate software components.
  - LIGO Laboratory and LSC need to defien a procedure for LSC to contribute (on a schedule) validated software components (numerical filter library components) for use in LDAS.
  - Data use and resource use model needs to be developed in cooperation with LSC:
    - Data distribution
    - Resource allocation for specific analysis tasks
    - Proposal procedure
    - ...

# LIGO Data Types/Products

## Frame format data

- The full [raw] detector datastream will be acquired and recorded as data frames.
  - Format for data frames has been unified with VIRGO in anticipation of being able to share software (now) and data (at some future date)
  - Other major interferometer projects have adopted standard
    - GEO
    - TAMA

- Frame Class Library (C++ implementation)
  - Implements Frame Format Specification
  - Progress to date:
    - v1.01
    - Documentation available on web (http://docuserv.ligo.caltech.edu/~wmajid/fcl/index.html):
    - HowTo's, Sources
    - Compatible and interfaced with CERN's ROOT package
    - Fcl to LigoLW(XML) translation module completed
  - Planned work (next quarter):
    - Implement interface to Matlab
    - Provide additional UNIX shell tools
    - Develop Tcl level Frame API for incorporating Fcl into LDAS architecture

# LIGO Data Types/Products

## Lightweight Data Format -- XML

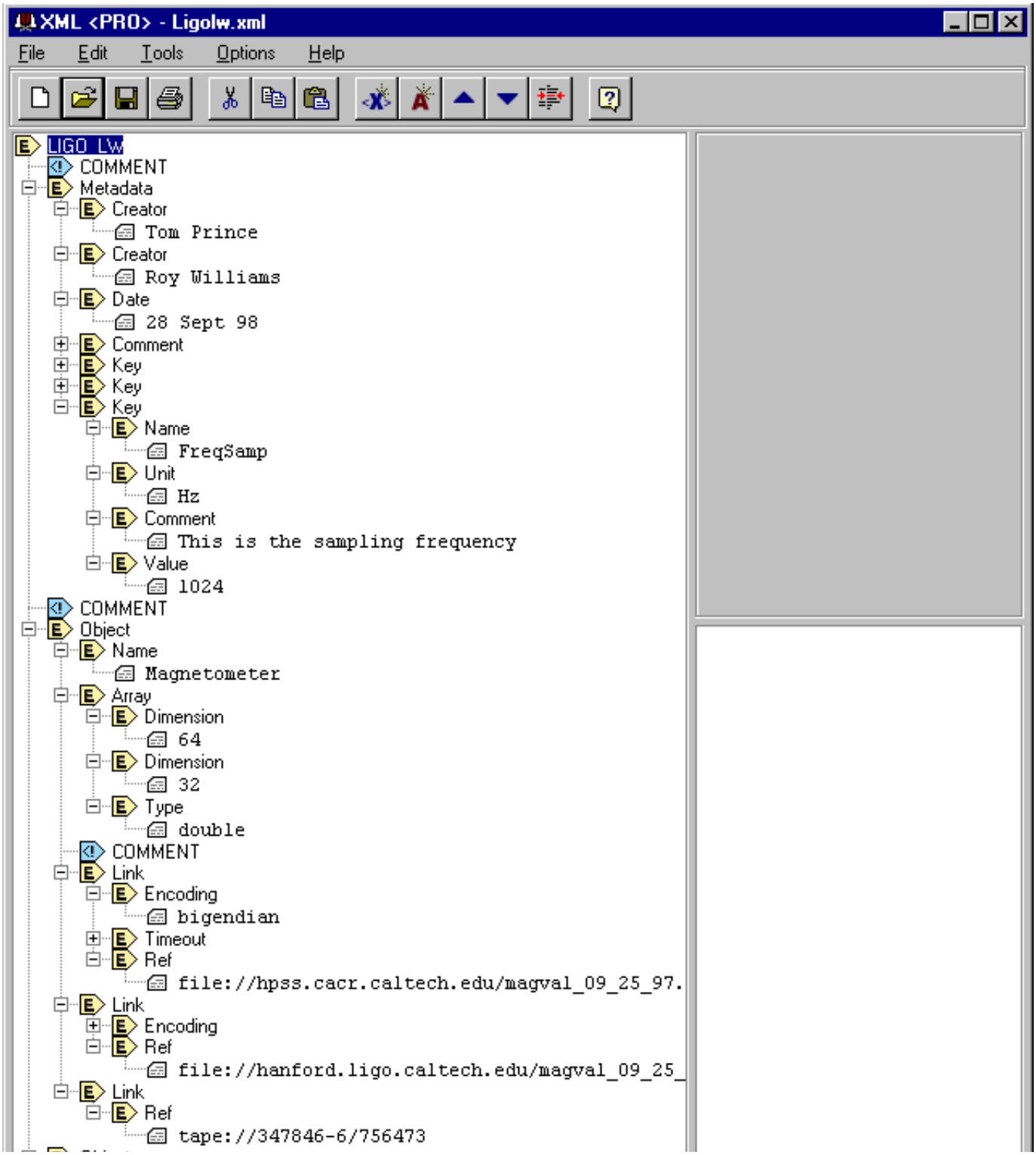
- LigoLW is based on XML to anticipate web-distribution, network distributed processing
  - Metadata: tags, keywords, elements, attributes
  - Data: encoded binary; ASCII; raw binary(?); other objects;...
- Lightweight format complements frames:
  - interprocess data communications (@ socket level)
  - easily readable/parsable format for end users
    - quick-look products, single channels
    - spectra
    - plots
    - events
    - metadata
  - estimated data volume: ~< 600 GB/yr reduced data;
    ~ 135 MB/yr metadata
- Reduced, processed, or otherwise non-frame data will be recorded in LIGO-standardized lightweight data format (LigoLW)
  - Metadata (data about data: frame catalog indices, operator logs, textual data, etc.)
  - Event data [event specification still TBD]
  - Spectra, time series snapshots, intermediate analyses performed with commercial/public-domain tools (MATLAB, Mathematica, ROOT, Triana, ...)

# LIGO Data Types/Products

## Lightweight Data Format -- XML

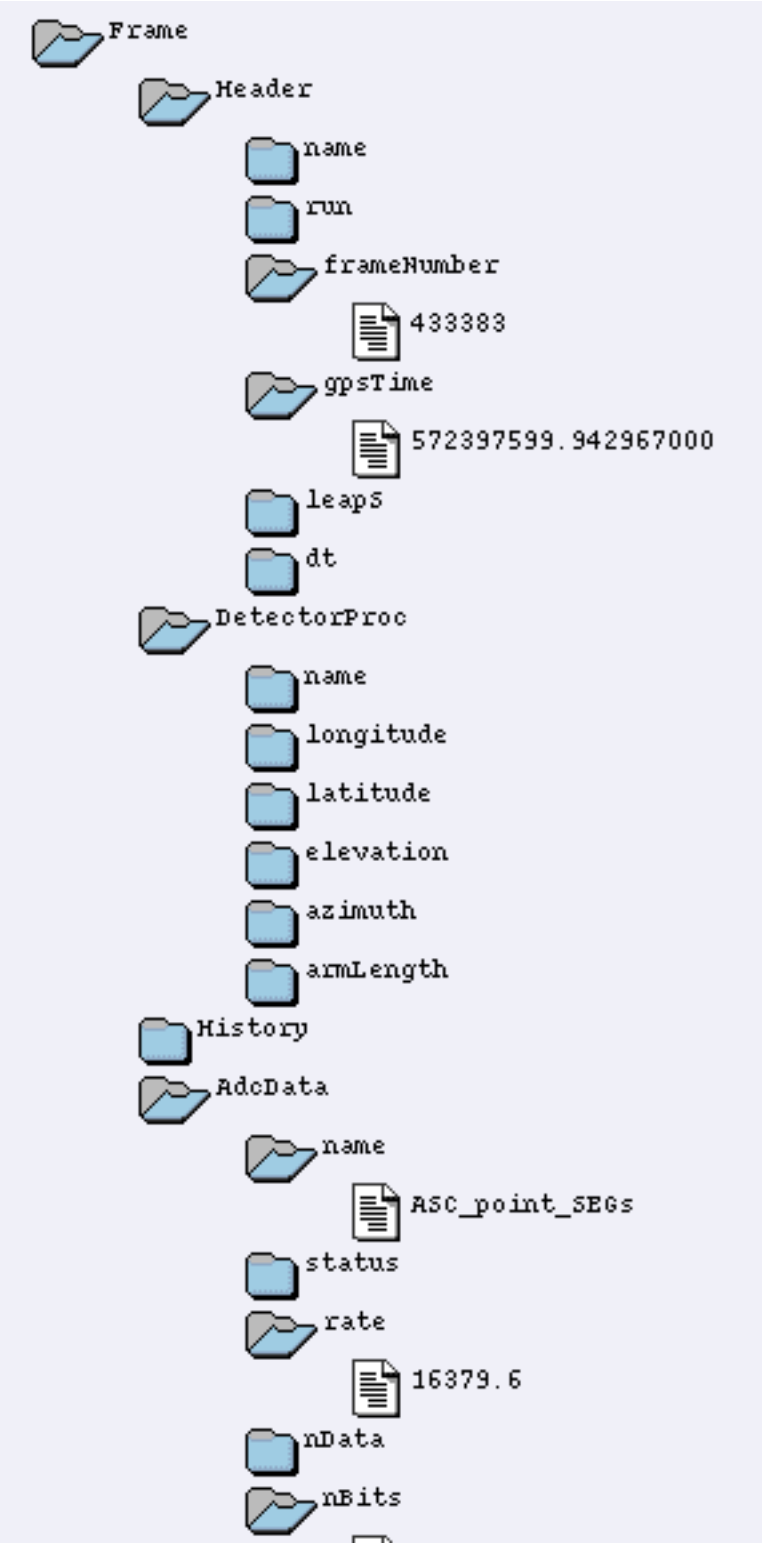
- Status:
  - ›› Specification released - draft DTD
    - – LIGO-defined data objects with defaults enable simple utilization *[http://www.cacr.caltech.edu/ligo/ligolw]*
      - – tables (ntuplets: points in a hyperspace)
      - – arrays (indexed elements of data)
        - • matrix
        - • vector, time-series, power-spectrum, ...
  - ›› First implementations
    - – Directed pulsar search results from 40m dataset - summer student project
    - – Socket-to-socket and Tcl-C++ LDAS interprocess communications prototyped
  - ›› Parser built to extract metadata from frames and to create LigoLW metadata catalog
  - ›› Revise specification over next quarter as experience indicates
    - ✓ Inputs from detector team (diagnostics) received, being incorporated

# LigoLW
## Parsed LW data object



# Frame - LigoLW Conversion

## Parsed frame object [XML <-> Fcl]



# LigoLW
## Example -- Metadata

```
<?xml version="1.0"?>
<!DOCTYPE LIGO_LW SYSTEM "Ligolw.dtd">
<LIGO_LW>
<!-- First the Metadata ------------------------------ -->
 <Metadata>
  <Creator>Tom Prince</Creator>
  <Creator>Roy Williams</Creator>
  <Date>28 Sept 98</Date>
  <Comment>LIGO power spectrum of 32 magnetometers at 64 frequencies</Comment>
  <Key>
   <Name>LIGOType</Name>
   <Comment>The Ligo data type is defined here...</Comment>
   <Value>Power Spectrum</Value>
  </Key>
  <Key>
   <Name>StartDate</Name>
   <Comment>Can't remember exactly but this date is close!</Comment>
   <Value>03/21/97</Value>
  </Key>
  <Key>
   <Name>FreqSamp</Name>
   <Unit>Hz</Unit>
   <Comment>This is the sampling frequency</Comment>
   <Value>1024</Value>
  </Key>
 </Metadata>
```

# LigoLW
## Example -- Data

```
<!-- Now for the Data objects ------------------------ -->
  <Object>
    <Name>Magnetometer</Name>
    <Array>
      <Dimension>64</Dimension>
      <Dimension>32</Dimension>
      <Type>double</Type>
    </Array>
<!-- This Array is at Cacr, Hanford, and on a tape -->
    <Link>
      <Encoding>bigendian</Encoding>
      <Timeout>600</Timeout>
      <Ref>file://hpss.cacr.caltech.edu/magval_09_25_97.bin</Ref>
    </Link>
    <Link>
      <Encoding>base64</Encoding>
      <Ref>file://hanford.ligo.caltech.edu/magval_09_25_97.bin</Ref>
    </Link>
    <Link>
      <Ref>tape://347846-6/756473</Ref>
    </Link>
  </Object>
   <Object>
    <Name>Magscale</Name>
    <Array><Dimension>32</Dimension></Array>
<!-- Embedded data -->
    <Data>
      1.28374 1.23453 1.94847 2.148474 2.39484 2.84746 3.10928 4.92827
      5.28374 5.23453 5.94847 6.148474 6.39484 6.84746 7.10928 8.92827
      9.28374 9.23453 9.94847 10.18474 10.3984 10.8446 11.1928 12.9827
      13.2874 13.2453 13.9847 14.18474 14.3984 14.8446 15.1928 16.9827
    </Data>
  </Object>
  <Object>
    <Name>Magoffset</Name>
    <Comment>This is the magnetic offset</Comment>
    <Array><Dimension>32</Dimension></Array>
<!-- Data follows from the end of the previous Object in the same stream -->
    <Follows/>
  </Object>
</LIGO_LW>
```

# Database Management Systems
## DBMS

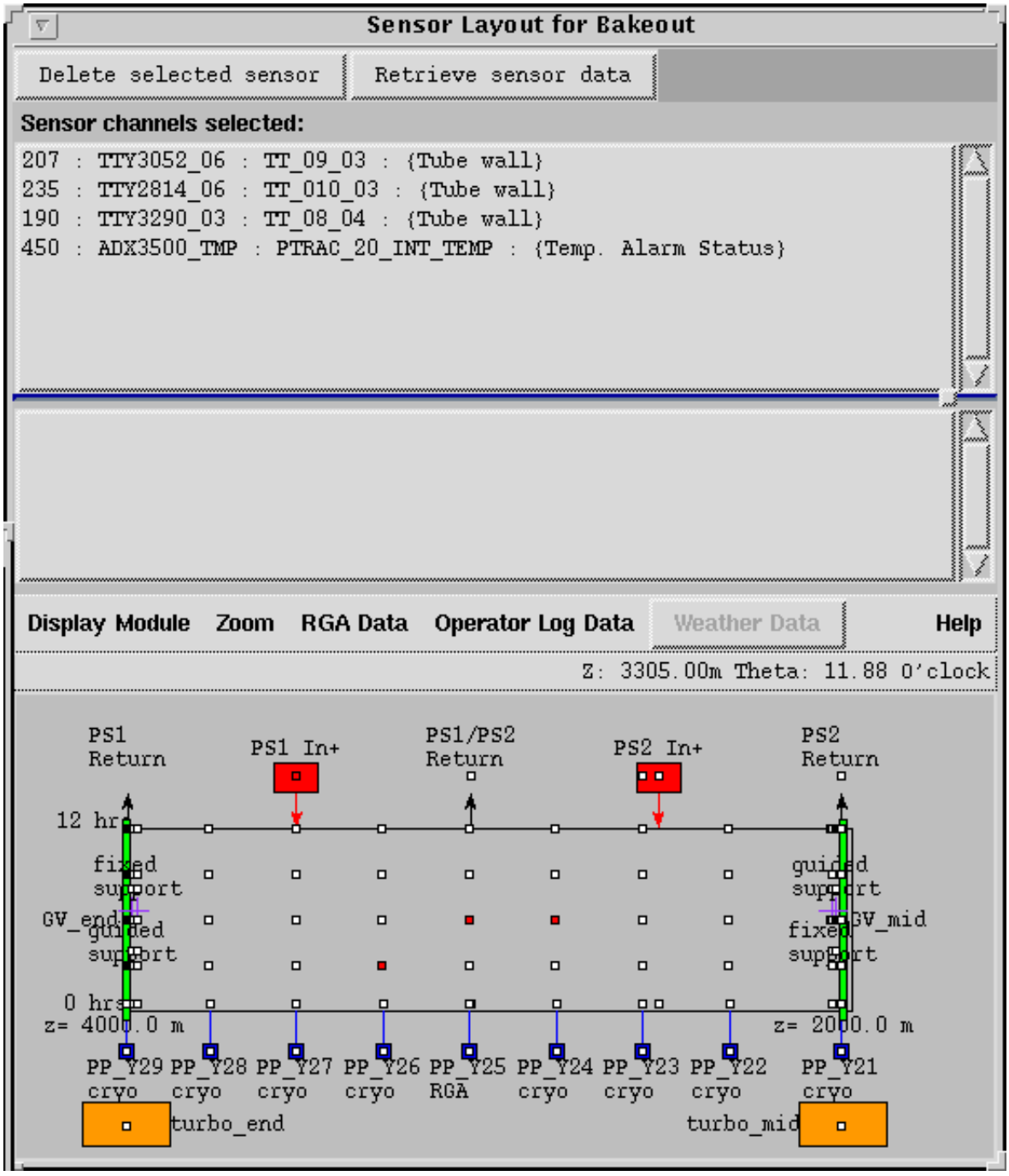
- LIGO has four data types that need to be managed:
  - raw, framed data -- HPSS or equivalent network file system
  - lightweight data -- HPSS or database management system (DBMS)
  - events (as they are generated, cataloged) -- DBMS
  - metadata -- DBMS
    - catalogs & indices
    - operator logs
    - trends and high-level descriptions of detector performance
- Still in process of deciding DBMS for LIGO
  - Held workshop 22,23 October with consultants from CERN, SDSC, CACR, Astronomy(IPAC/CIT) to review LIGO needs
  - Choices being considered:
    - relational [deemed sufficient for LIGO needs]
      - ORACLE (CIT license for campus MIS)
      - DB2 (CACR HESC license for Caltech and related site activities -- platform specific)
    - object-oriented DBMS [deemed unneeded for LIGO needs]
      - Objectivity
  - Issues: Buy-in costs; operational costs; upgrades if we start too low; metadata only vs (metadata+data); ...
  - Need to have a selection by January

# Database Management Systems
## DBMS prototyping activities

- Ongoing BT Bakeout activity is generating 4 disparate DBs
  - temperature, current, pressure data along BT
    - Microsoft ACCESS DB
  - residual gas data from RGA
    - Proprietary SW from RGA vendor: spreadsheet compatible
    - partial pressure vs time scans
    - mass spectra
    - calibrations
  - weather station/environmental data: proprietary SW with station vendor: spreadsheet compatible
  - operator logs
    - text (ASCII) files
- Data arrive weekly as (~80MB files)
  - 700 channels x 10000 rows
  - data are ingested (transformed) and metadata produced for indexing into archive -- 3 hour ingestion process on NT server @ CACR
- Need to make data available at future dates for intercomparisons as bakeout progresses
  - DB will eventually grow to ~ 2GB, indexed by timestamp
  - Metadata+data co-located in DB
  - http://www.cacr.caltech.edu/ligo/bakeout/index.html

# Database distribution tools

## BT Bakeout Data Distribution Prototype



# Data Analysis System for LIGO I
## Software Design

### *LDAS SOFTWARE DESIGN FEATURES -- LAYERED DESIGN*

**Languages:**

- ANSI C++
- ANSI C for wrappers to C, FORTRAN and TCL
- TCL (Tool Control Language) for control of resources.processes
- TK for Graphical User Interfaces
- Tclets (TCL/TK plug-ins) for web browser connectivity
- *TBD* database for data/metadata

**Communications**:

- TCL layer sockets to communicate commands and messages between processes
- C++ socket class library to communicate data between processes
- MPI (Message Passing Interface) for numerically intense parallel [scientific] computing.

**Libraries**:

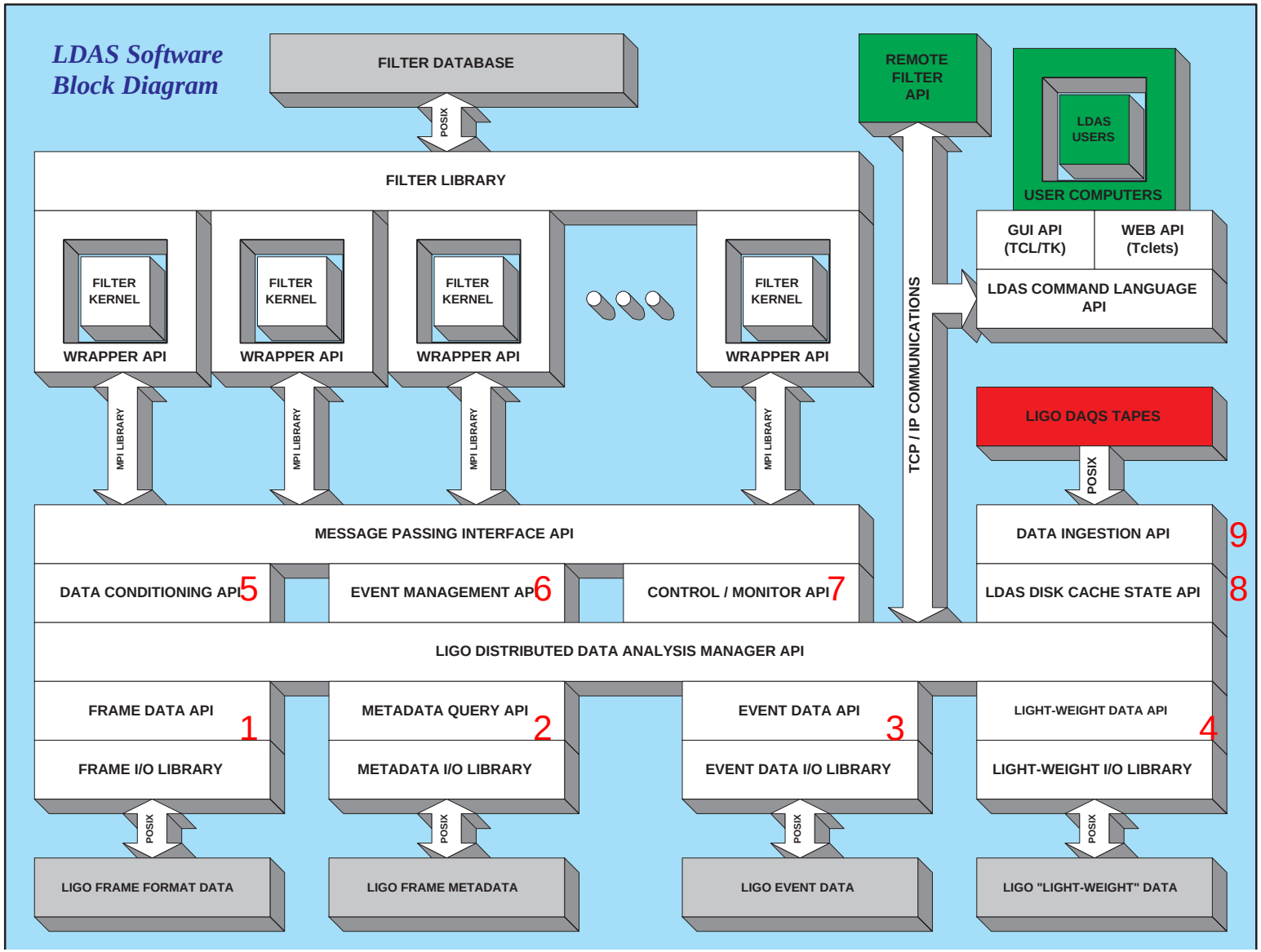
- Shared C++ Class Libraries, numerical libraries and I/O libraries on supporting platforms for efficient use of hardware resources

# Data Analysis System for LIGO I
## Software Design

### ›› Application Programmer Interfaces (APIs)

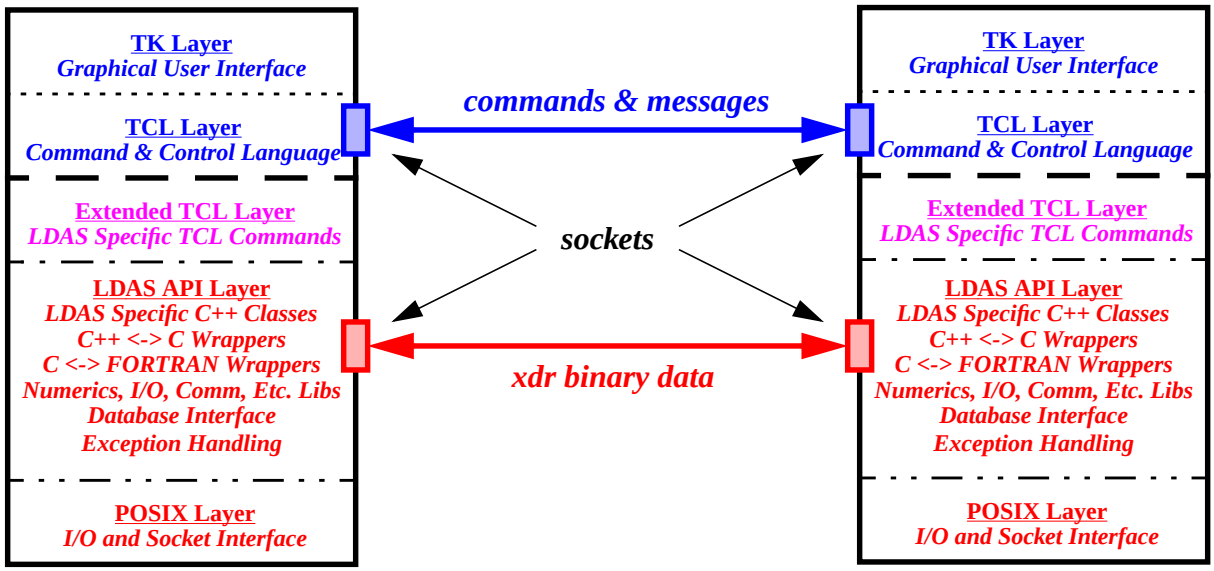
1. Frame Data
   - Manipulates framed data; I/O; channel extraction/ insertion; concatenation; ....
2. Metadata API
   - Interacts with the DB environment; data entry/extraction; data searches/sorts/queries.
3. Event Data API
   - Updates event lists; classifies events; searches on events;
4. LigoLW Data
   - Frame->LigoLW translator; data object extraction/ insertion.
5. Data Conditioning API
   - Data pre-processing; calibration; filtering; regression; computation either done using filter kernels or within this API (depends on complexity);
6. Event Management API
   - Receives output from the MPI based filter kernels; reports events; displays; ...
7. Control & Monitoring API
   - LDAS configuration, monitoring, exception handling, resource allocation; user interaction;
8. Disk Cache API
   - Stages data from archive/large disk farm to intermediate cache for efficient retrieval; queues data requests.
9. Data Ingestion API
   - Incorporates new data into archive; filter; reduce; compress.

# LIGO Data Analysis System
## Software Design



# LIGO Data Analysis System
## Software Design

### APIs "*TWO-LEVEL*" SOCKET COMMUNICATIONS




*Yellow boxes below indicate option to use SCSI*

| ***API*** | **FW** | **MPI** | **DC** | **EM** | **CM** | **FD** | **MD** | **ED** | **SDF** | **FK** | **COM** | **DI** | **DCS** | **MAN** | **RF** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **FW** | | mpi | | | | | | | | inherit | | | | | |
| **MPI** | mpi | mpi | socket | socket | socket | | | | | inherit | | | | | |
| **DC** | | socket | | | socket | socket | socket | | socket | | | | | socket | socket |
| **EM** | | socket | | socket | socket | | | socket | socket | | | | | socket | socket |
| **CM** | | socket | socket | socket | | socket | socket | socket | socket | | | | | socket | socket |
| **FD** | | | socket | | socket | | | | socket | | | | socket | socket | |
| **MD** | | | socket | | socket | | socket | socket | socket | | | | socket | socket | |
| **ED** | | | | socket | socket | | socket | socket | socket | | | | | socket | |
| **SDF** | | | socket | socket | socket | socket | socket | socket | | | | | socket | socket | socket |
| **FK** | inherit | inherit | | | | | | | | | | | | | |
| **COM** | | | | | | | | | | | | | | socket | |
| **DI** | | | | | | | | | | | | | socket | | |
| **DCS** | | | | | | socket | socket | | socket | | | socket | | socket | |
| **MAN** | | | socket | socket | socket | socket | socket | socket | socket | | socket | | socket | | socket |
| **RF** | | | socket | socket | socket | | | | socket | | | | | socket | |

# Data Analysis System for LIGO I
## Software Development - Status

- GenericAPI:
  - Basis for all other APIs. Initial investment in prototyping and design will allow rapid diversification into specific APIs by extension of the generic class.
  - Complete

*The first APIs to be developed will support Detector Installation milestones for the first interferometer*

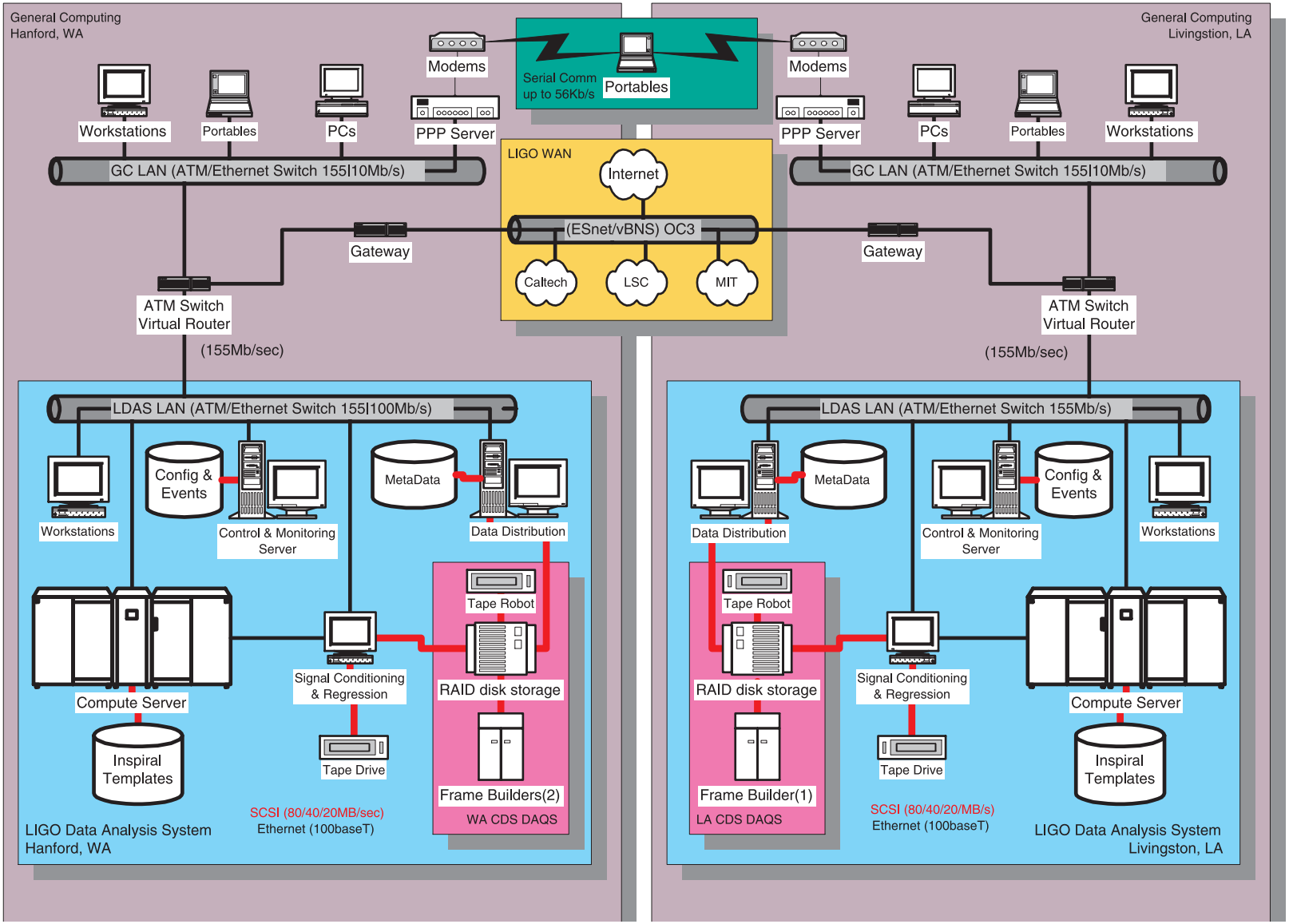
- Under development:
  - Fcl I/O Library: complete
  - Fcl Specification: complete
  - FrameAPI
  - ManagerAPI
  - DataConditioningAPI
  - UserAPI

- Reviews:
  - Develop prototype UserAPI by end of Dec.
  - Hold Preliminary Design Review in January to support first deliveries to Hanford.

# Hardware Status

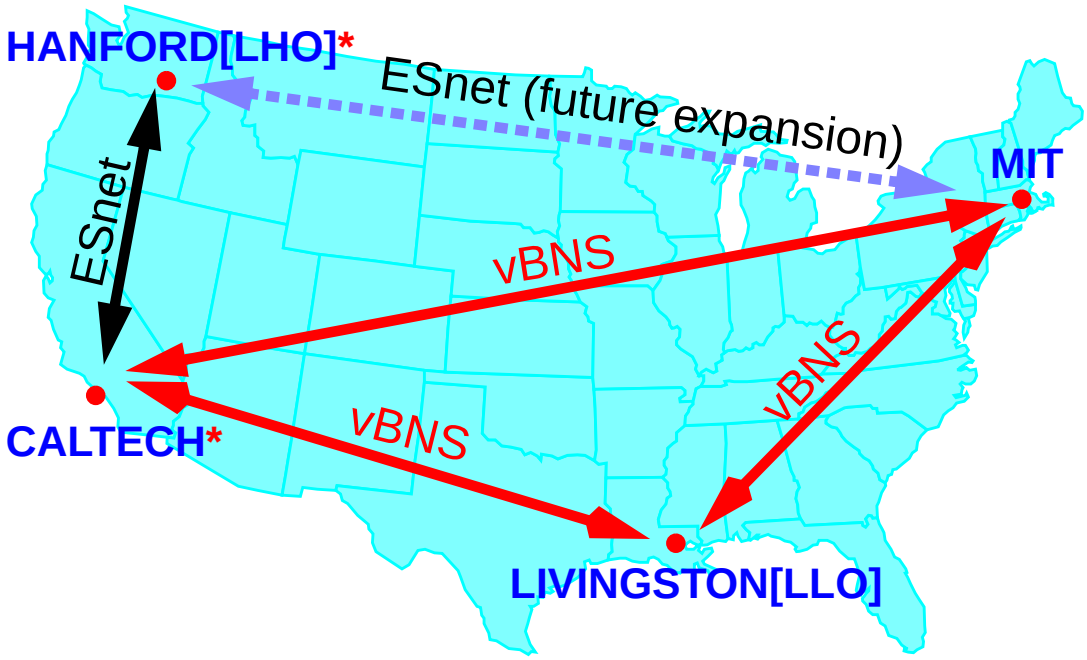
- Implementing prototype data server for Hanford at LIGO Caltech
  - LAN implementation
  - Transmission of data from 40m prototype
  - Data server
  - Data conditioning machine
  - Software installed on corresponding hardware platform per on-line LDAS design.
  - Linux cluster installation at LIGO
- Data analysis:
  - Beowulf and MPI has been demonstrated on LIGO-scalable data flows for inspiral detection
    - Joint effort with CACR (Paragon) & Univ. Wisc. (PC/ linux)
    - 8 node/16CPU integrated cluster in house at LIGO - being set up.
  - Directed pulsar search prototype code has been implemented using 40m data and CACR machines.
- Data archival technology choice will be deferred as late as possible (2001) => working through CACR
  - Optical tape technology replacement for magnetic media in HPSS (LOTS)
  - 1TB/cassette (same form factor as present IBM robot cassettes)
  - ~$250/cassette ($0.25/GB)
  - Optical heads replace magnetic tape heads in same cabinetry.

# LIGO Data Analysis System
# On-line architecture



# LIGO Wide Area Network

## WAN Topology




***LHO will be accessible to LSC member institutions via Caltech**

**WAN/LAN Connectivity among LIGO Laboratory Sites**

| *Site* | *Livingston, LA* | *Hanford, WA* | *MIT* | *Caltech* |
|---|---|---|---|---|
| *Caltech* | vBNS(OC3) | ESnet (4 X T1) <-> vBNS(OC3) | vBNS(OC3) | OC3/ATM 100BT |
| *MIT* | vBNS(OC3) | MIT<->Caltech<->Hanford<br>ESnet (4 X T1) <-> vBNS(OC3) | 100BT OC3/ATM(?) | |
| *Hanford, WA* | ESnet (4 X T1) <-> vBNS(OC3) | OC3/ATM 100BT | | |
| *Livingston, LA* | OC3/ATM 100BT | | | |

# LIGO Wide Area Network

## Status

- LIGO proposed & drafted an MOU between NSF/DOE to provide access to ESnet at Hanford
- Authorized in October 1998
- Implementing initial (T1) capability; requested up to 4 x T1 BW (cost is an issue).
  - Routing: LHO-PNNL-SDSC-CACR-LIGO/Caltech
  - In Progress:
    - DNS in process of being turned over to Caltech ligo-wa.caltech.edu
    - T1 connectivity tested and working
    - Move workstations over to new IP addresses
    - Setup E-mail and Web services
  - Planned:
    - MOU covers 4 T1 connections -- may take advantage of contingency.
    - WSU/Pullman (~ 100km NE) awarded an NSF grant to establish a vBNS hook-up
    - UW/Seattle (~350 km W) has vBNS at present
    - PNNL is investigating future high speed connections via Seattle -- LIGO will participate if costs are acceptable.
- MIT may be added later as a separate addendum to MOU
- Access to LHO will be via Caltech as a gateway to ESnet

# LIGO Wide Area Network

## Status

- T1 link to Livingston Observatory is in place
  - LSU awarded vBNS access in latest round of NSF awards -- includes LIGO access at Livingston
  - LSU provides gateway service
    - Caltech providing DNS services ligo-la.caltech.edu
    - E-mail and Web services in process of being setup (last week)
  - Planned:
    - Finalize hardware logistics with LSC
    - Install main server
    - Establish modem services and contingency plan
    - Establish OC3 Connectivity in the next 1-2 years depending on fiber availability (present connection is Cu)
      - LIGO will have to install FO lines from Livingston to the Observatory
      - Upgrade the routing equipment to accommodate new connectivity
- LLO is directly accessible by all LSC member insititutions via internet

# LDAS Development
## Timeline

- Highest priority: staged implementation of on-line systems to support detector testing:

| Detector Milestone: | Date | LDAS Need |
|---|---|---|
| PSL/Input Optics | 4/99 | Min. data dist. |
| Vertex Michelson, “first light” | 9/99 | Full data dist. |
| 2km operational | 8/00 | On-line system |

- Staged installation of off-line system at Caltech in period 9/99 - 12/01