LIGO Data Analysis and Computing

A. Lazzarini

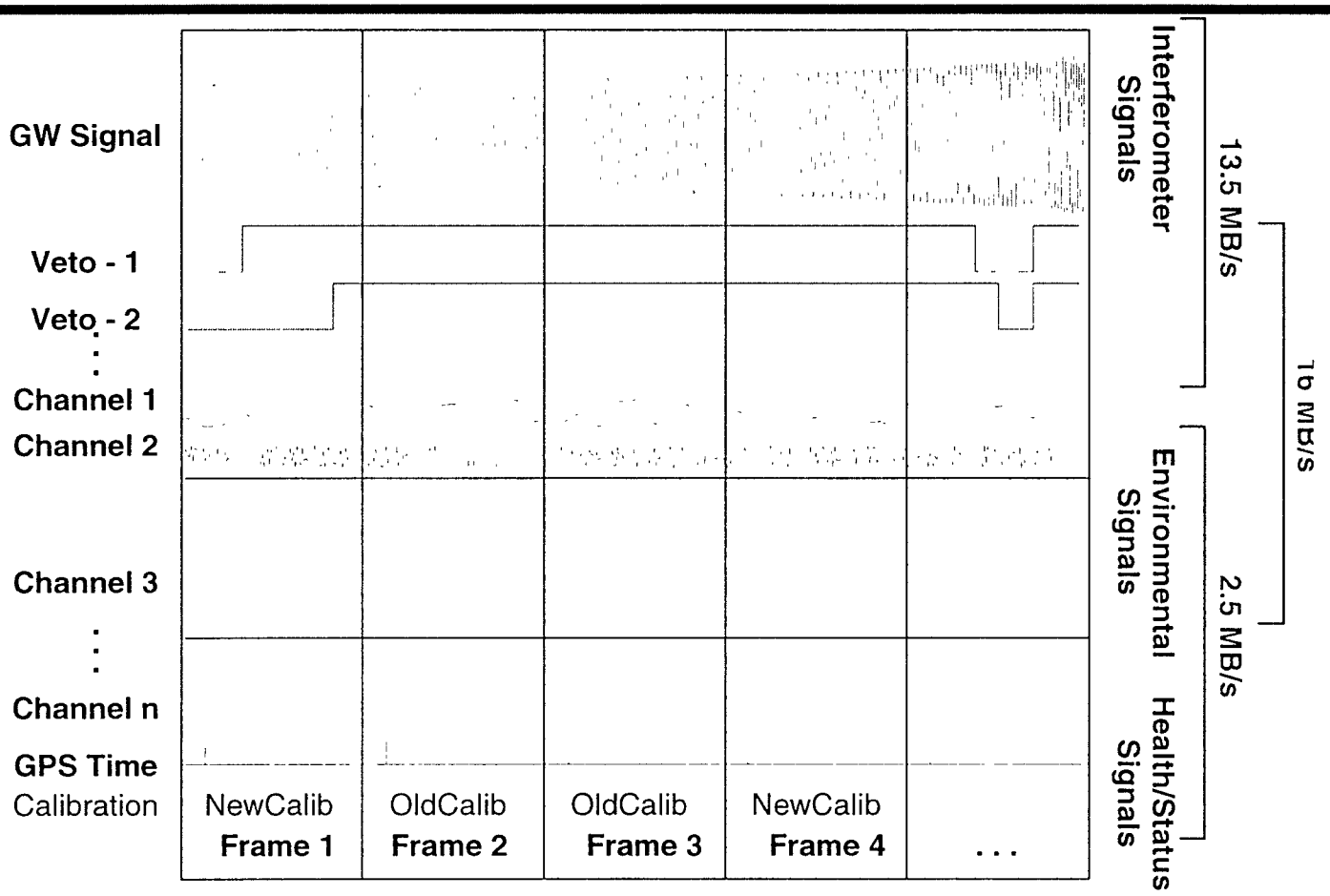
J. K. Blackburn

PAC Meeting

6 - 7 January 1997

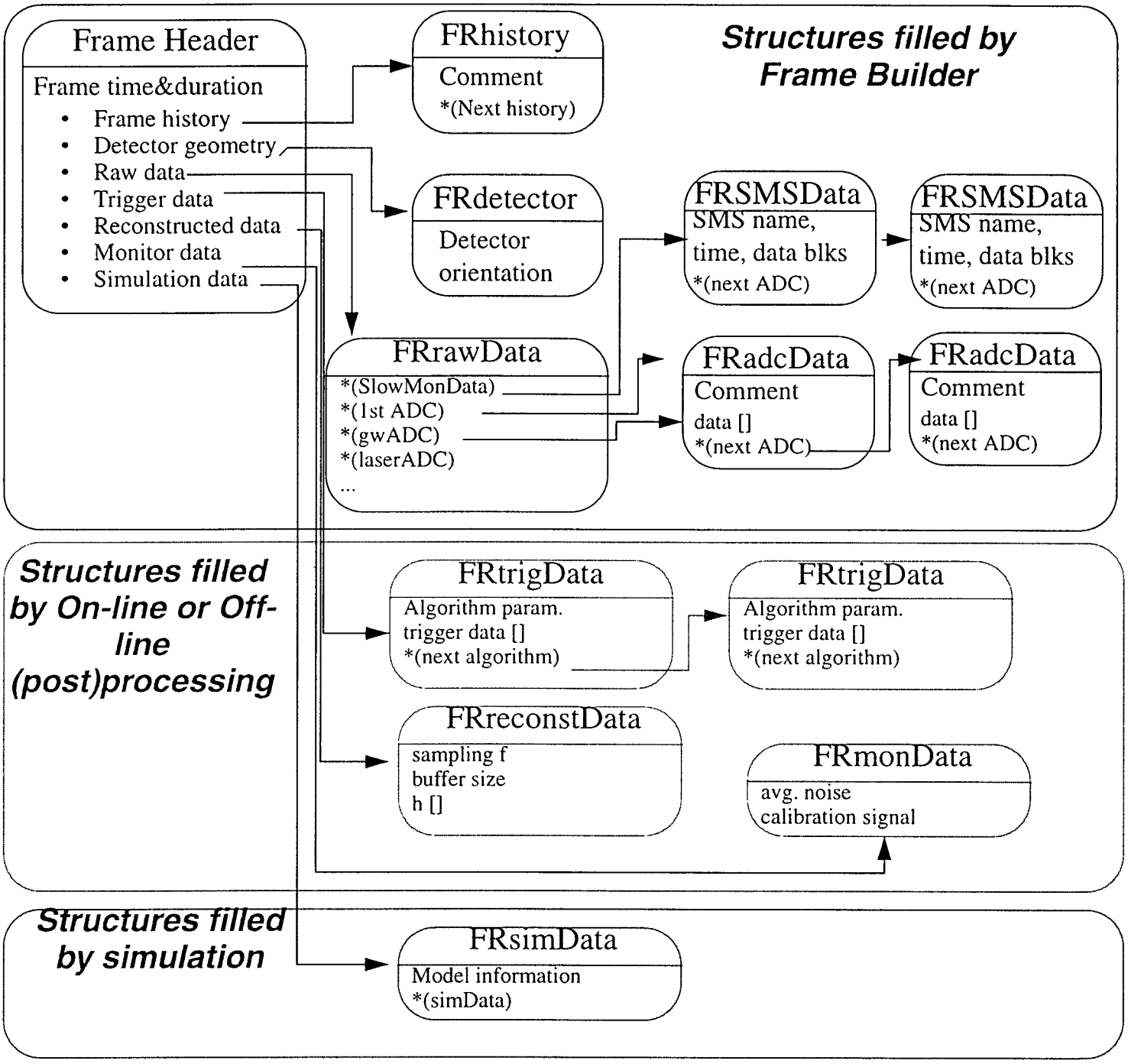
California Institute of Technology

LIGO Data Stream and Data Frame Design



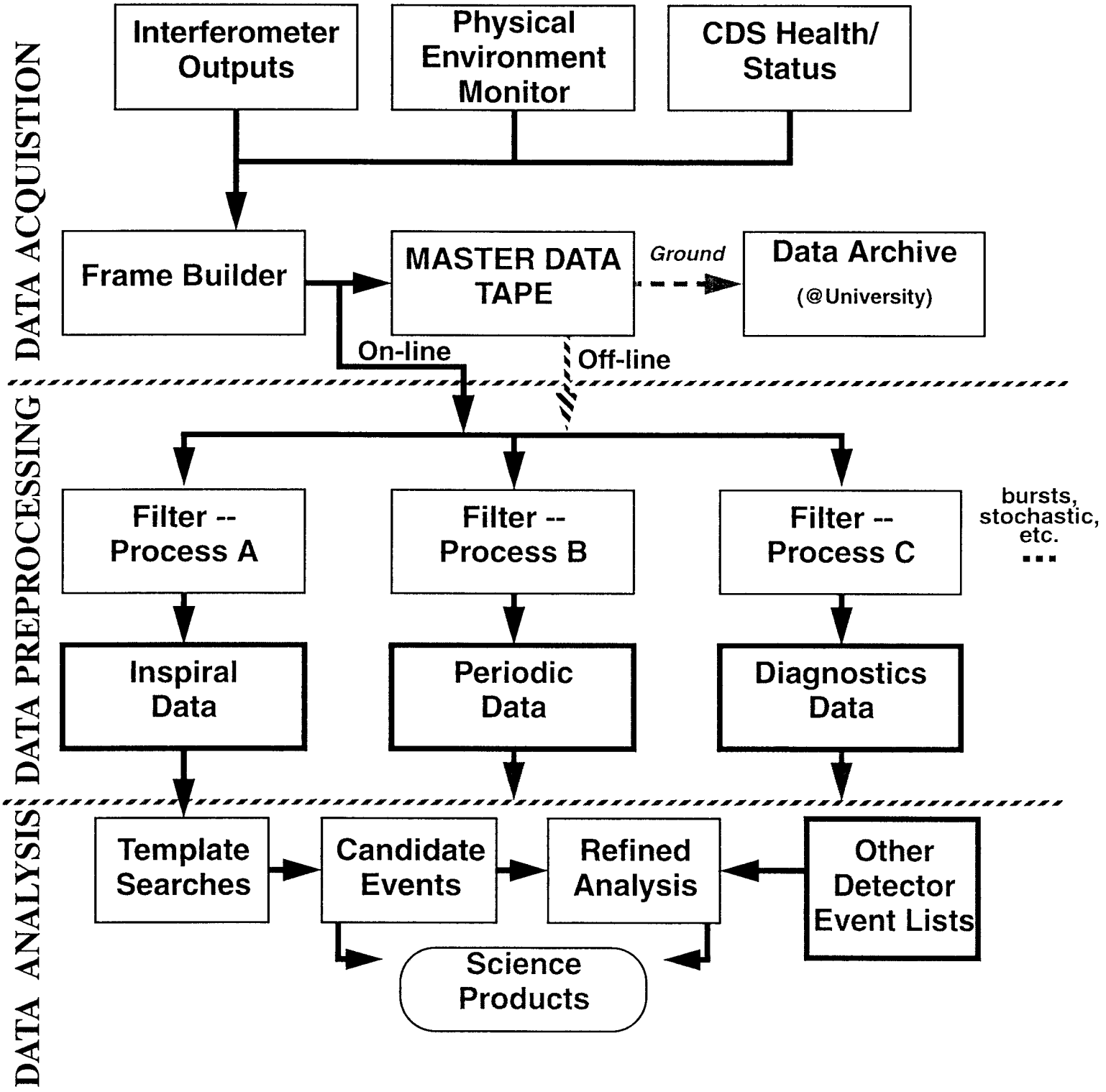
- Frame is (structured) self-contained snapshot of data for a period of time
  - GW channel & ancillary IFO channels
  - Environmental monitoring (veto) channels
  - Facilities/Vacuum health & status
  - Hierarchical organization of data reflects IFO subsystems for more efficient veto utilization

LIGO-VIRGO DATA FORMAT



- **Frame has tree structure:**
- **Individual blocks are C structures**
- **Extensible to arbitrary length with design evolution**
- **Utilized for both on-line & off-line analyses**

LIGO Data Analysis Flow -- Baseline



Ongoing Activities
Prototyping

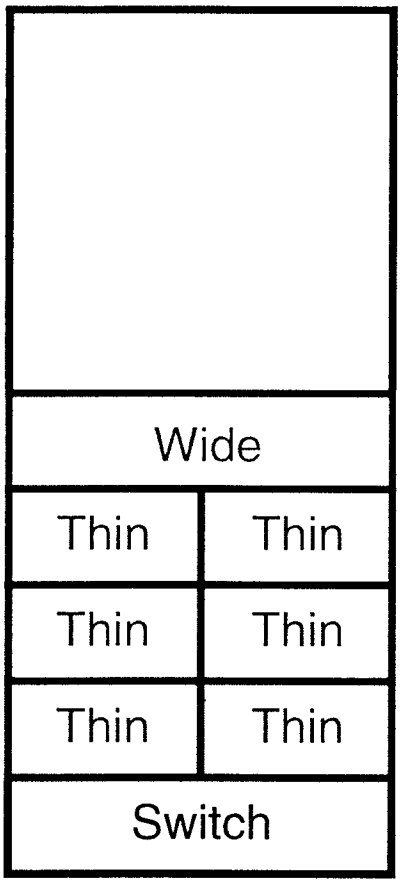
- Detector construction phase is developing a prototype DAQ system for the 40m facility
  - Utilize 40m to acquire datasets of substantial length (1/2 day) on a regular basis
  - Experimental use of ancillary channels for data qualification
- LIGO co-authored joint proposal for IBM Sponsored University Research (SUR) Grant funding - $800k of processor hardware will be awarded
  - LIGO will participate in hardware configuration definition; to be shared with other campus groups
  - Hardware to be installed at Center for Advanced Computing Research (CACR)
  - CACR already has similar NSF-funded hardware for astrophysics data analysis
- Use ongoing work to provide realistic scaling of parallel analysis algorithms for large data sets
- Establish data link from 40m to CACR

IBM SP2 HARDWARE

Upgrade of Existing SP2 Frame/Node Hardware

Wide Node:
512MB RAM
2.2GB Disk
156 MFLOPS

Thin Node:
128 MB RAM
2.2 GB Disk
133 MFLOPS



Space for additional
8 Thin Nodes
or
4 Wide Nodes

Add 384MB RAM
bringing total to
512MB in Wide Node

Add 2MB of L2
Cache to each
of 6 Thin Nodes

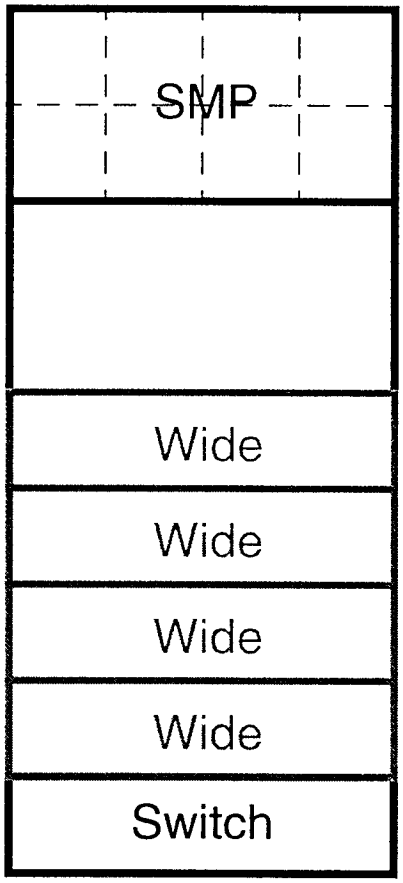
Upgrade Switch &
Switch adapter cards

IBM SP2 HARDWARE

Acquisition of New SP2 Frame/Node Hardware

SMP Node:
8x604CPUs
512MB RAM
4.4 GB Disk
8x23MFLOPS

2 Wide Node:
512MB RAM
2.2GB Disk
262 MFLOPS



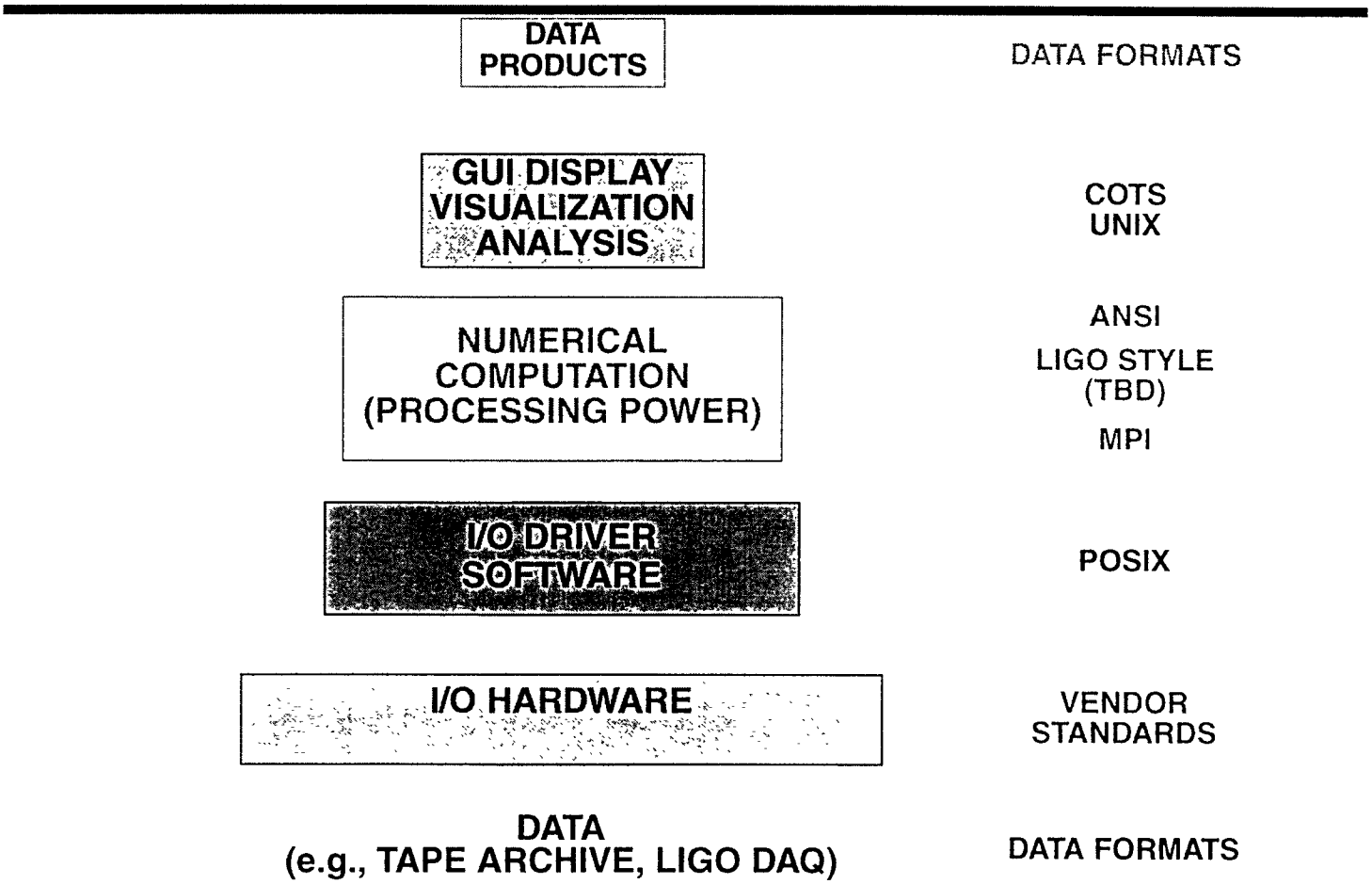
Space for additional
4 Thin Nodes
or
2 Wide Nodes

2 Wide Node:
256MB RAM
2.2GB Disk
262 MFLOPS

440 meter dprototyping activities

- Nov 1994 Data Run
  - ›› 46 hours (15GBytes) of data collected
  - ›› Analyzed by Gillespie & Lyons in Ph. D. Theses
- Recent Analysis (B. Allen)
  - ›› Developing package to read/analyze:
    - set up template family spacing (Owens, Sathyaprakash)
    - do optimal filtering (Wiseman)
    - rejection of "bumps", "drips", "scrapes", "pings", "howlers"
    - 169 triggers in 4 hour stretch of data studied
    - simulation: chirp signal injection and detection (to 15 kpc)
    - parallel processing being implemented using MPI
    - binary inspiral analysis of data complete by spring
  - ›› Also Complete:
    - pipeline for stochastic background analysis (Romano)
    - simple time/frequency domain diagnostic tool (real-time)
  - ›› To Be Added:
    - quasi-normal mode ringdown detection (black hole formation)

LIGO Standards



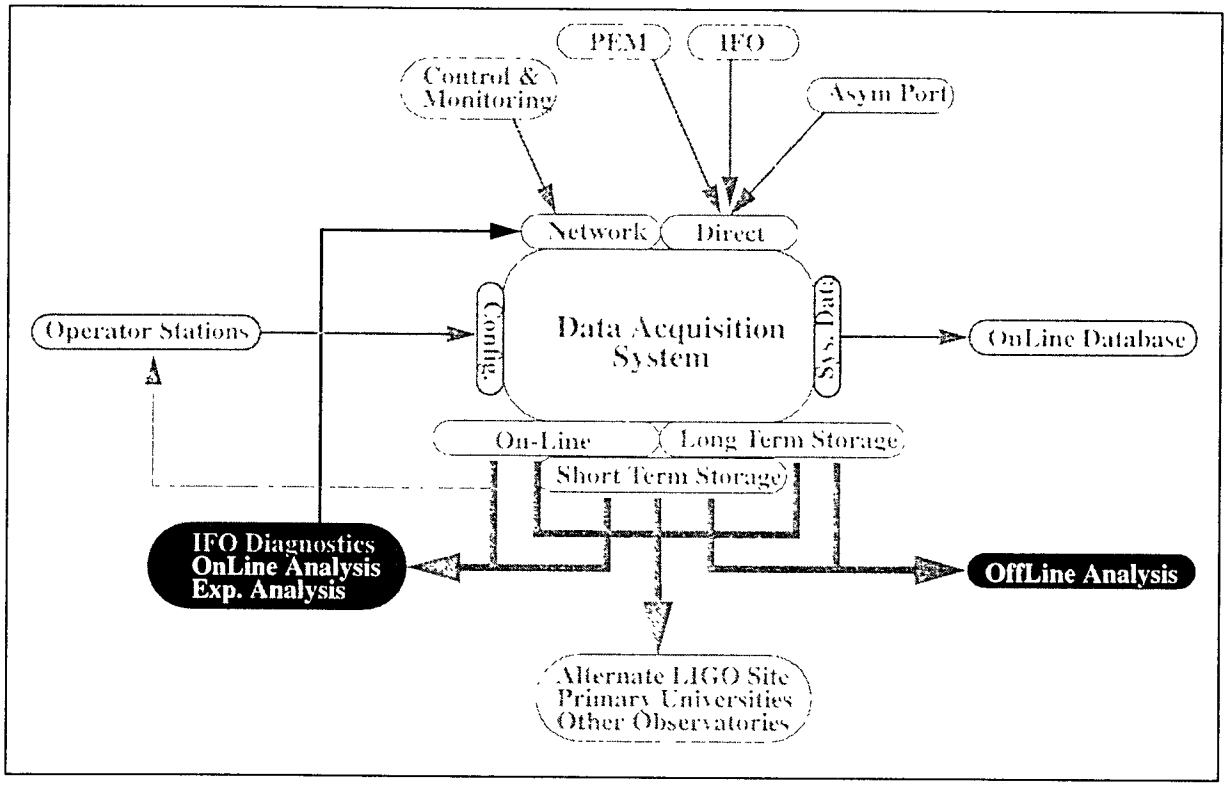
- Software libraries are modular,addressing various analysis functions: analysis/visualization; computation; I/O; etc.
- Limit the sources of potential platform-dependent occurrences of software routines to low-level (standardized) drivers

Data Analysis System for the Initial LIGO

- Software
  - Layered, standardized, stylized, documented (users' manuals,etc)
  - I/O libraries to access archived frames
  - concatenation of frames
  - extraction of specific channels
  - cross-correlation among channels
  - diagnostic software - time and frequency domain
  - analysis software - filtering algorithms
  - 2D/3D display & visualization

Data Analysis System for the Initial LIGO

- LIGO Data Acquisition System (DAQ) Design



DAQ Interfaces

- On line analysis & diagnostics are closely related
- Off line analysis supports "deep" searches and (multiple) searches not feasible on line

Data Analysis System for the Initial LIGO

- CDS DAQ provides for on-line short term storage
  - 63GB (commissioning) -> 400 GB (later)
  - 8 hr complete data framess (CDF); 10 days limited data frames (LDF)
  - 25MB/s I/O
- Online data analysis system will use same resources and augment them as needed

Data Analysis System for the Initial LIGO

- On-line analysis
  - ›› Diagnostics - ensure instrumental sensitivity at all times
    - performance metrics -- Gaussian noise
      - $h_{rms}[t]$
      - statistics of h in limited frequency bands
      - frequency-time analyses
        - Wigner-Ville distributions
        - moving window periodograms
    - performance metrics - non-Gaussian noise
      - templates (limited range: $m_{NS} \approx m_{Sun}$)
      - wire resonances -- $\delta A[t]$ - $\delta\phi\,[t]$ phasor diagram
      - frequency-time analyses
        - Wigner-Ville distributions
        - moving window periodograms
      - physical environment monitoring system (PEM)
    - calibration - lines/broadband - h[t] extraction, if possible
    - "quick-look" analyses - real time

Data Analysis System for the Initial LIGO

- On-line analysis
  - ›› Astrophysics
    - detection of transient phenomena for which coincident operation with other (highly) directional instruments is feasible/ desirable
    - Supernovae
      - Lmited to Milky Way -- rare (.025/yr)
      - ν detectors are omnidirectional => can be done with post processing correlation
      - γ /visible light curves => want to track curve from onset of SN explosion
        - GW signal would provide trigger => directional instruments slew to position;
        - Requires cooperation among GW detector projects
          - LIGO localizes event to a cone -- insufficient information;
          - 3 - 4 instruments needed;
        - Requires site-to-site communication correlation in real time
        - Maximum few hour latency allowed - (< 2 hours)

Data Analysis System for the Initial LIGO

- On-line analysis
  - ›› Astrophysics (cont.)
    - — Supernovae
      - Supernova signature is of short duration
        - wavelet characterization -- astrophysical templates not presently known
        - data stream cross-correlated to obtain confirmation/direction
        - using on-line event detection triggers, echange short stretches of data windowed on event
    - — NS/NS coalescence
      - LIGO detection would not be able to affect other directional detectors
        - end point may produce fast/short burst of EM radiation, but detection does not allow slewing of directional instruments
        - long stretches of data analyzed together for computational efficiency
      - EM detection (γ burst) can always be correlated off-line since LIGO is omnidirectional
      - On-line template filter bank is planned to aid in characterization non-Gaussian noise performance

Data Analysis for Initial LIGO

On-line Processing Computing Resources & Distribution

- Redundant systems at LA & WA Observatories
- Support for 1x, 2x, 3x operations independently
  - Diagnostics -- especially during commissioning
  - 2x/3x operations between sites feasible with reduced datastreams
    - Transient/burst signals ($\Delta T < 1s$) -- short stretches of GW + veto/QA
    - Inspiral & coalescence waveforms ($360s\ (m \approx m_{sun}) \leq \Delta T \leq 5000\ s\ (m \approx 0.2 m_{sun})$) -- events
- System configuration
  - Volatile data storage for 3 hours of data + 3 hours of analysis (FIFO) for 2 IFOs (WA) @ 100% data stream: 125GB+125GB
  - Template storage for : 5 - 500 GB ($M_{NS} \approx 1 - 0.2\ M_{SUN}$)
  - 2 - 50 ($M_{NS} \approx 1 - 0.2\ M_{SUN}$) GFLOP CPU system -- intrinsically parallel computational requirements:
    - Parallel processor(s) -- *monolithic/efficient/more expensive*
    - Workstation cluster -- *versatile/less efficient/less expensive*
    - Specialized (DSP) system -- *less versatile/efficient/least expensive/upgrade difficult*

Data Analysis for Initial LIGO

On-line Processing Computing Resources & Distribution

- System configuration (cont.)
  - Site-to-site communication link to provide 2x and 3x real-time cross-correlation
    - Selected (pre-processed) data subsets (GW + veto; event lists)
    - Two way: WA->LA & LA->WA
      - Can support independent algorithms
    - T1: 0.2 MB/s is sufficient for peak rates required for GW WA->LA
    - T3 (6 MB/s) or ATM (20 MB/s) will be available by time needed

Data Analysis System for the Initial LIGO

- Off-line analysis -- must keep up with data stream
  - "Production" of data -- deal with 500 TB/yr problem
    - consolidation/refinement of data for permanent archive
    - data product generation -> h[t] for scientific analysis
    - "quick-look": 1 day -> 1 month
    - data QA assessment: % of time vs noise floor/lock/availability
  - Scientific analysis
    - Supports analyses either not feasible or not required on-line.
      - Stochastic background
      - Pulsar searches (directed/partial sky)
      - Inspiral with combined IFOs (vector data for max. SNR)
      - Research on algorithm development & signal processing
      - Refined analyses
      - Novel searches
    - Manipulates data archive

Data Analysis System for the Initial LIGO

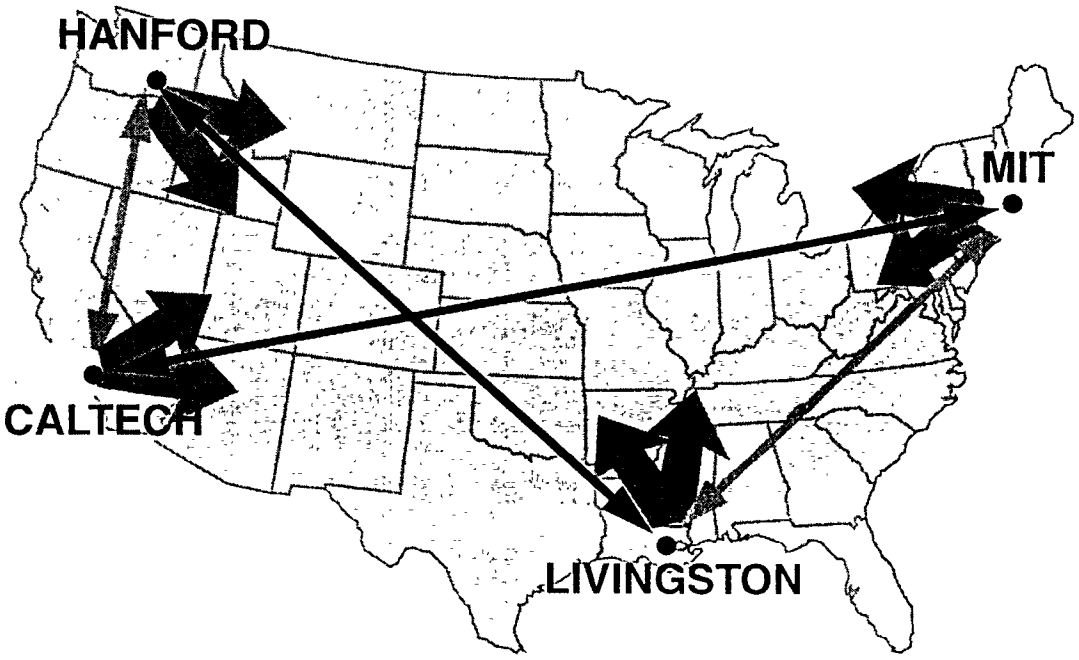
- Scientific analysis
  - Data access via WAN to other LIGO sites and users.
  - Utilizes and is designed around existing University resources for maintenance, availability, communications & support.
    - LIGO computational resources resident with archive
    - 10 - 50 GFLOPS
    - 450+ GB disk farm(s) for tape dumps (per analysis type)
    - Supercomputer center-class systems
  - Analyses likely to be distributed among several national resources (SDSC, NCSA, NPAC, etc.) e.g.: pulsar searches

Data Analysis System for the Initial LIGO

- Data "QA" Reviews
  - regular (frequent) meetings to review detector performance and statistics of data from previous period
    - % locked/up time
    - $h_{rms}[t]$ statistics
    - event lists (template filter outputs)
  - responsibility is to decide on archived data
    - what to keep
    - what to recycle (keep limited data frames -- always keep GW channel) -- annouce to collaborations: if someone wants 100% data, he/she can retrieve it, store it, etc.

Data Analysis System for the Initial LIGO

- Networks



- ›› Hanford-Livingston link permits real-time cross-correlations among instruments
- ›› Caltech-MIT link provides high speed link to data archives; data tapes to be archived at university.
- ›› Site-University links provides site scientific staff access to archived data
- ›› University gateways provide broader access to database
- ›› Data tapes transported to University repository

Data Analysis System for the Initial LIGO

- WAN options being explored:
  - Caltech
    - ESNET (DOE: planned upgrade to OC12@622MB/s)
    - JPL/NASA link to NASA backbone
    - CACR link(s) to other SC centers (vBNS:planned upgrade to OC12)
  - MIT
    - LNS/ESNET (B. Bruen/P. Dreher; DOE: planned upgrade to OC12@622MB/s)
    - CSR/NASA link to NASA backbone
    - vBNS
  - Hanford, WA
    - ESNET (DOE: planned upgrade to OC12@622MB/s)
    - Initial discussions with DOE (M. Plahuta) and PNWL (G. Johnson)
  - Livingston, LA
    - Initial discussions with LSU (Chip Dodson)
    - LSU part of a 6-university SE Consortium to access vBNS -- proposal pending @ NSF (vBNS/OC3)
    - NASA link to MSFC (T1)

Data Analysis System for the Initial LIGO

- Ongoing efforts:
  - Data Analysis System Requirements
    - Hardware - location/storage/performance
    - Communications - WAN requirements; bandwidth; up time; connectivity
    - Software - functional requirements/specifications/standards
  - Conceptual Design
    - Hardware architecture
    - WAN architecture
    - Software module specification; user environment; implementation approach
    - Prototyping activities for scaling studies and verification