LIGO Data Analysis
Data Formats
and
Modeling Activities

Albert Lazzarini
LIGO Integration Group

NSF Fall Review
22 - 24 October 1996

OUTLINE

1. Data Analysis System for the Initial LIGO Detector
   - Science requirements/computational requirements
   - Preliminary concept
     - Data analysis flow
     - Distribution of computing resources
     - Access to resources -- network options
   - Ongoing & planned activities; issues
2. Data Formats for LIGO Detector
   - Status of collaboration with VIRGO
   - Common format -- VIRGO model
   - Unresolved issues
3. Modeling & Simulation Activities in LIGO

Data Analysis for Initial LIGO

- LIGO Construction Phase includes Data Acquisition System (LIGO DAQ)
- Archival & Analysis Systems fall within scope of Operations Phase.
  - Need will grow gradually during detector commissioning
- McDaniel Panel Report to NSF identified need to develop analysis capability to support both Laboratory and Collaboration research:
  - Computing systems for LIGO; networks -- WAN; maintenance & management of resources
  - Greater computing power required for more complex searches
  - Data distribution and availability -- PAC consultation
- LIGO is developing a conceptual plan for initial data analysis system which will be accessible to both Laboratory and Collaboration:
  - Outline prepared for Fall 1996 NSF Review
  - Refinement of requirements and concept to be conducted in conjunction and in consultation with broader community (LRC).
  - White paper to be available Spring/Summer 1997.

Data Analysis Requirements

Science & Computational Requirements

Initial LIGO Sources and Estimated Analysis Capability Requirements

	Sources	Initial LIGO Performance Estimate	Data Analysis Requirements: CPU	Data Analysis Requirements: Storage	Data Analysis Requirements: Comments
Burst Signals $\Delta T < 1$ s	Supernovae	$\Re_0 \sim 2-3/\ \mathrm{yr}$ @ 15 Mpc If sufficiently asymmetric	Minimal for straightforward correlation; *if optimal filters are discovered, problem may increase in complexity.*	Minimal Need PEM/houskeep-ing data for veto	• *On-line analysis* desirable for correlation with other astrophysics: **Electroweak** • visible/radio/γ (HETE, GRO) • ν (Super-K/SNO) **Gravity** • VIRGO/GEO • Resonant bars • Waveforms unknown • 2x/3x IFO correlation • Off-line analysis to enhance SNR
	BH/BH Collisions	$\Re_0 \sim 1/\ \mathrm{yr}(?)$ @ 500 Mpc; $M_{BH} \sim 30$ - $200 M_{SUN}$			
Chirped Waveform $10s < \Delta T < 1000s$	NS/NS Inspirals	$\Re_0 \sim 3/\ \mathrm{yr}$ @ 23 Mpc; $\Delta T \sim 4 \times 60$ s $\quad M_{NS} \sim M_{SUN}$	~ 2 GFLOPS	Templates/Data ~20 GB /~1 GB	• *On-line analysis* for M_{NS}>M_{SUN} can be done; appears feasible down to ~ 0.3 M_{SUN} • 2x/3x correlations feasible depending on SNR. • Coalescence event may generate correlated (EW) signals as above. • PEM/housekeeping needed for vetoing • Template matching (Wiener filtering) or wavelet analysis in f-t domain. • Off-line analysis to enhance SNR
		$\Delta T \sim 4 \times 500$ s $\quad M_{NS} \sim 0.3 M_{SUN}$	~ 50 GFLOPS	~500 GB /~10 GB	
	BH/BH Inspirals	$\Re_0 \sim 1/\ \mathrm{yr}$ @ 150 Mpc; $\Delta T \sim 4 \times 10$ s $\quad M_{NS} \sim 10 M_{SUN}$;	~ 2 GFLOPS	~20 GB /~1 GB	

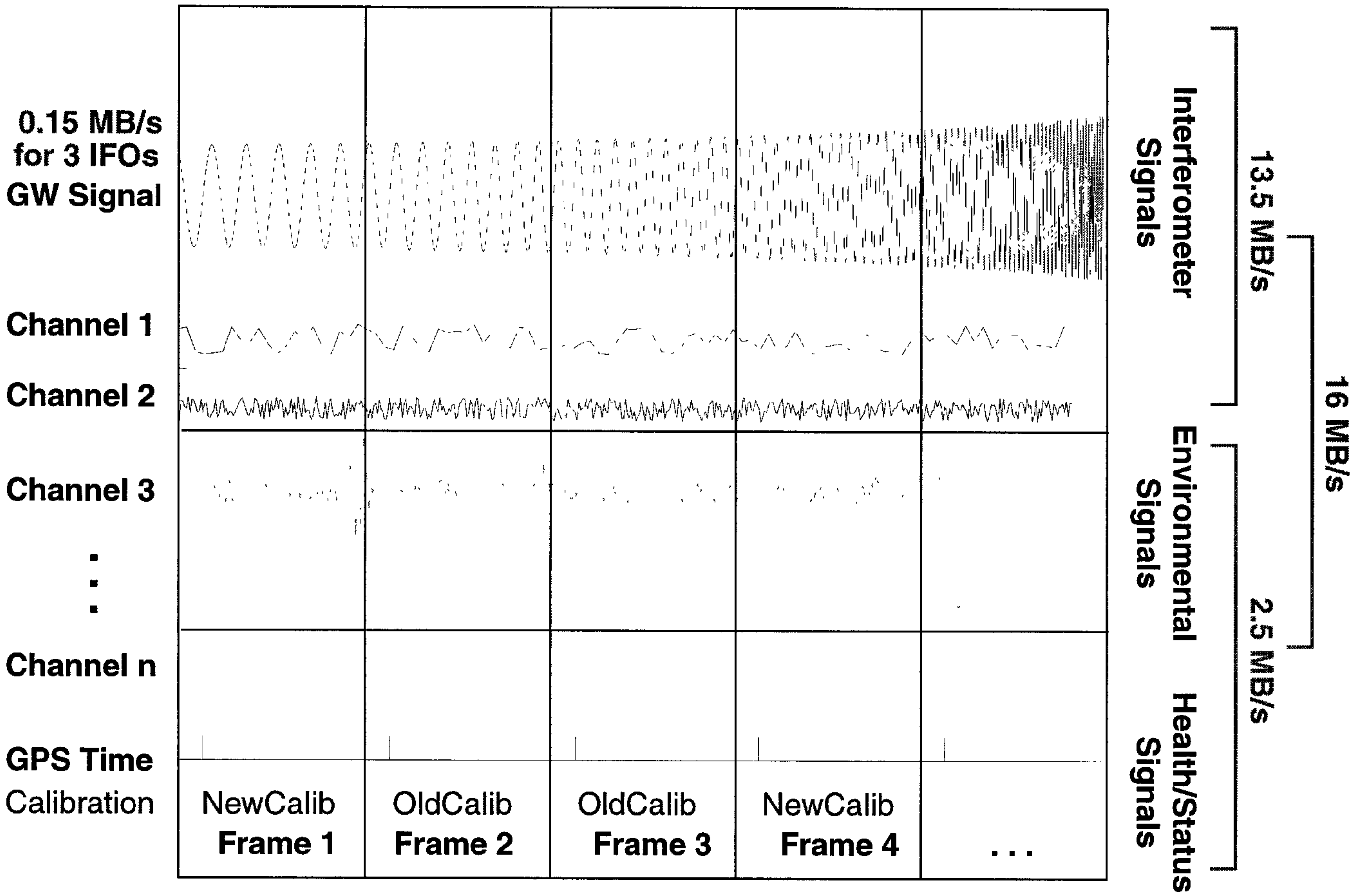
Data Analysis Requirements

Science & Computational Requirements

Initial LIGO Sources and Estimated Analysis Capability Requirements

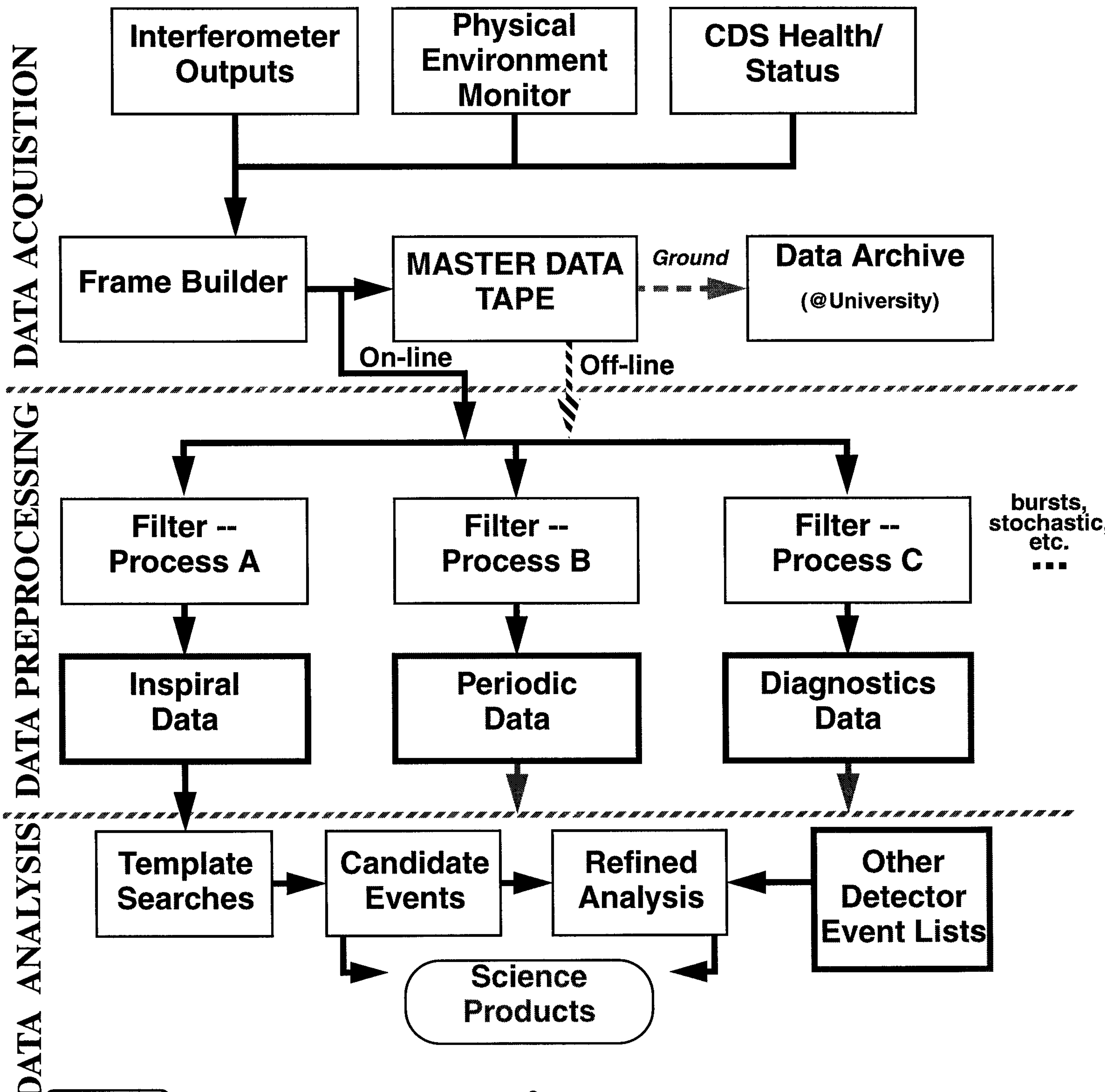
	Sources	Initial LIGO Performance Estimate	Data Analysis Requirements: CPU	Data Analysis Requirements: Storage	Data Analysis Requirements: Comments
Periodic Signal $\Delta T \sim 10^6 - 10^7$ s	Pulsars with mass asymmetry $h \propto \left(\frac{\varepsilon}{10^{-6}}\right)\left(\frac{10\text{kpc}}{r}\right)\left(\frac{1\text{ms}}{P}\right)^2$	$\varepsilon = 3\text{x}10^{-5}$; r=10kpc ; P=1ms $T_{int} = 10^6$s SNR ≈ 5	Directed searches (e.g., galactic center, known pulsars) require minimal resources All-sky searches require tens of TFLOPS -- beyond anticipated capabilities	10 GB for 10^6s (GW waveform)	• *Off-line analysis* • Detection less sensitive to non-Gaussian noise; more sensitive to calibration drifts&drop-outs • Detection techniques as for pulsars -- narrow line sources with modulated frequency. • Correlations among interferometers may be performed (if needed) after detection. • All-sky search requires decomposition of 4π sr into $>10^{10}$ pixels, each region requiring a different spectral transformation of same dataset.
Broadband Signals $\Delta T \sim 10^6 - 10^7$ s	Stochastic Background $\Omega \equiv \frac{\Omega_g}{\Omega_0}$	$\Omega \geq 3\times10^{-6}$ Δf, f ≈ 100Hz $T_{int} = 10^7$sec	Minimal requirements -- analysis maybe done on single workstations		• *Off-line analysis* • Requires multiple interferometers to be correlated; may use PEM to imprive SNR.

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

LIGO Data Analysis Flow -- Baseline



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 ($\Delta T < 1\,s$) -- GW + superveto/QA
    - – Inspiral & coalescence waveforms ($10s < \Delta T < 1000s$) -- events
- System configuration (target: $\mathbf{M_{NS} > 0.3\ M_{SUN}}$)
  - ›› 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:300 GB
  - ›› ~ 2-50 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*

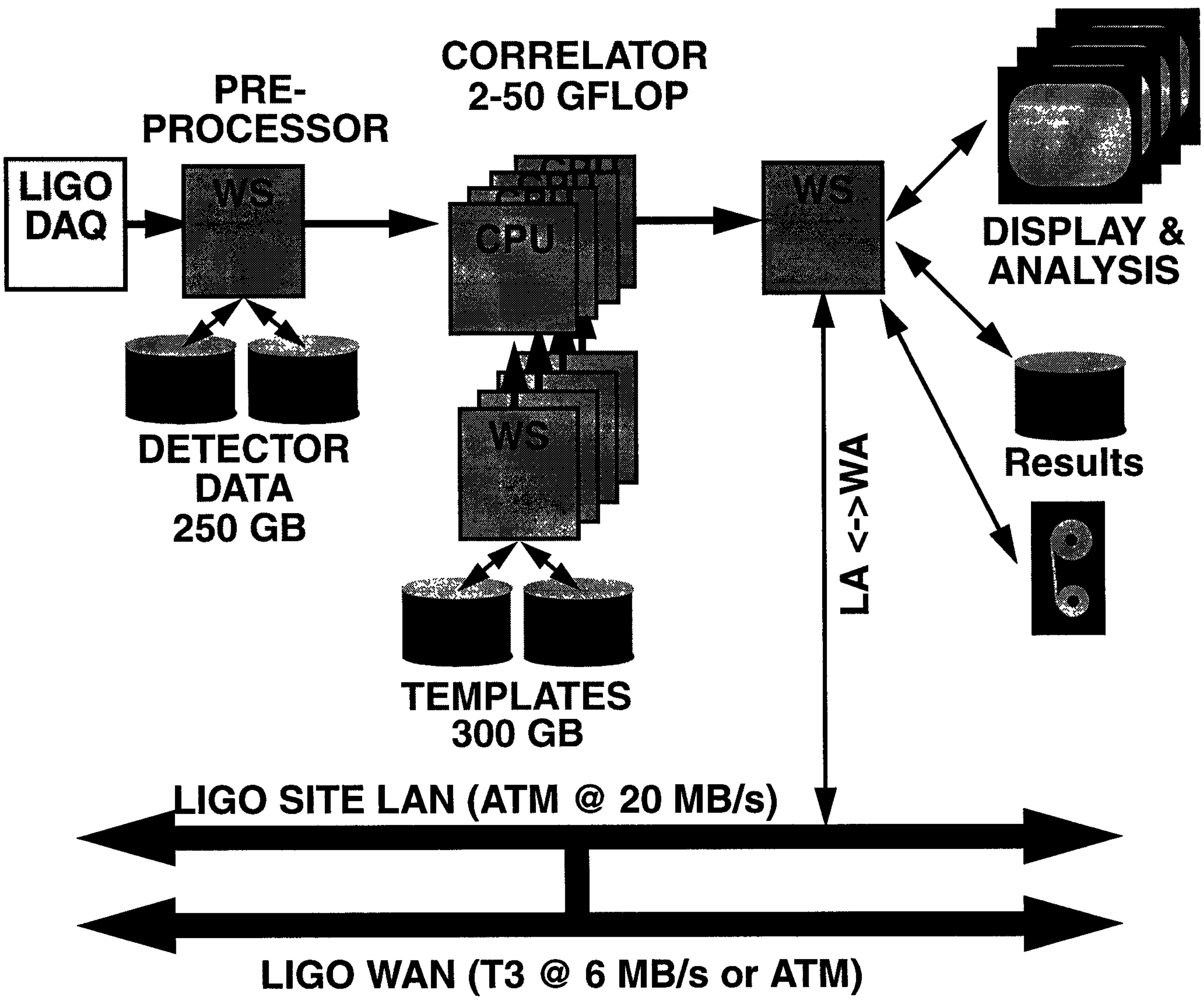
Data Analysis for Initial LIGO

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 + super-veto; event lists)
    - Two way: WA->LA & LA->WA
      - Can support independent algorithms
    - T1: 0.2 MB/s is barely sufficient for GW WA->LA
    - T3 (6 MB/s) or ATM (20 MB/s) will be available by time needed

Data Analysis for Initial LIGO

On-line Processing Computing Resources & Distribution



Data Analysis for Initial LIGO

Off-line Processing Computing Resources & Distribution

- Single system at a LIGO Laboratory University*
- 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
- Provides/manipulates data archive.
- Data access via WAN to other LIGO sites and users.
- Utilizes and is designed around existing University resources for maintenance, availability, communications & support.

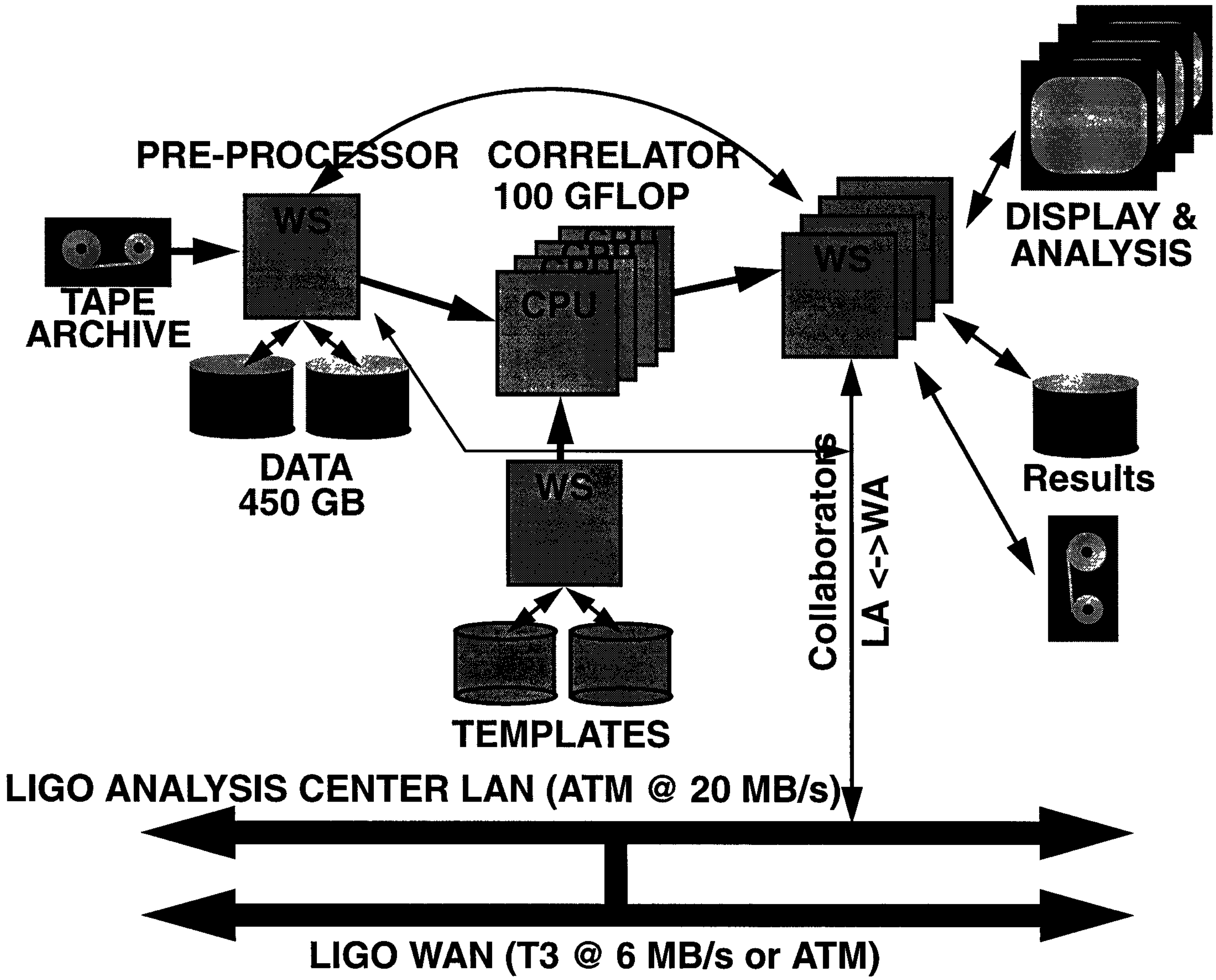
Data Analysis for Initial LIGO

Off-line Processing Computing Resources & Distribution

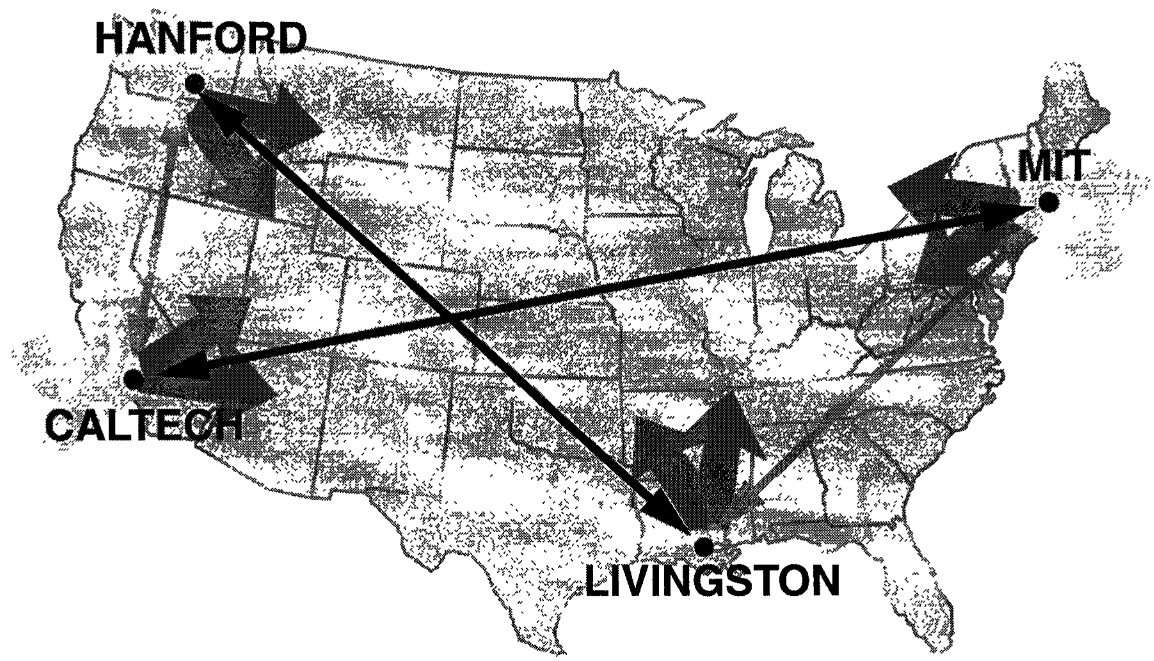
- System configuration (target: max. capability for multiple users)
  - Large data archive
    (~ 500 TB/yr => 10k tapes/yr @ 50 GB/tape => $0.5M/yr @ $50/tape)
  - Robotic tape access -- size TBD
  - Disc cache system capable of storing 450GB of data
    - 8 hours of 100% data ~ 450 GB
    - ~ 5 weeks of GW data (suitably filtered to not require ancillary channels)
  - Processors for computationally intense analyses (100+ GFLOPS)
    - Support multiple, independent analyses (4 - 6)
    - Parallel processor(s) -- *monolithic/efficient/more expensive*
    - Workstation cluster -- *versatile/less efficient/less expensive*
    - Distinctions will fade with time
  - High bandwidth communication to other LIGO sites & collaborating institutions
    - T3 (6 MB/s) or ATM (20 MB/s)

Data Analysis for Initial LIGO

Off-line Processing Computing Resources & Distribution



LIGO Site-to-site Communications



- 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

Site Communications

- Options for utilizing existing resources -- these are being explored:
  - Caltech:
    - HEP link to MIT/CERN (DOE:ESNET; plan: OC12@70+MB/s)
    - IPAC/JPL link to NASA backbone (NASA)
    - CACR link(s) to SC centers (NSF: VBNS->OC12@70+MB/s))
  - MIT:
    - HEP link to Caltech/CERN (DOE - ESNET)
    - NASA backbone (NASA)
    - Link(s) to SC centers (NSF - VBNS)
  - Livingston:
    - LSU link to MSFC/NASA backbone (NASA)
    - LSU link to SC centers (NSF - VBNS)
  - Hanford:
    - HNR/BNWL (DOE - ESNET)

Planned Activities

Timeline for Development

Milestone or Event	Date	Communications	Hardware	Software
Begin Coincidence Operations	7/00	Common		
On-Line System Available	1/00	Common		Verification
	3/99-12/99	Implementation	Procurement & Integration	Development
	11/98	Agreements	Specifications	
System FDR	11/98	Definition	Design & Prototyping	Specifications
System PDR	11/97			Design & Prototyping
System DRR	5/97			

LIGO-G960211-00-E

/home/lazz/Presentations/NSF_Reviews/NSF_96_10/NSFReview961022-p1-v4.fm5

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

Issues

- LIGO Analysis System design must contend with two conflicting needs...
  - Rate of technology growth argues for delaying investment in hardware to the latest possible moment...
  - Need to develop/debug analysis software on specific platform(s) to support detector commissioning. COTS & strict adherence to standards.
- Efficient utilization of 40m prototype DAQ system and CACR is key to developing an extensible, modular system which is capable of providing LIGO Laboratory & Collaboration adequate analysis tools for the first generation detectors:
  - Validation of software
  - Identification of best hardware approaches
  - Benchmarks for on-line processing

Issues
(cont.)

- Efficient use of detector ancillary data channels is key to containing archive growth
  - 100% data stream corresponds to $>10^4$ tapes/year;
  - GW channels correspond to $<10^2$ tapes/year
- Actual cost of archival is bounded...
- Two approaches possible...
  - ›› Start with minimum channel count and add channels as experience dictates through commissioning phase
  - ›› Start with 100% channel count and pare back as experience dictates
  - ›› First option more reasonable and less costly.
- During definition phase, LIGO will actively seek LRC representation in design inputs.
  - ›› This is the first presentation by LIGO
  - ›› Process will take a year or more

LIGO-VIRGO DATA FORMAT

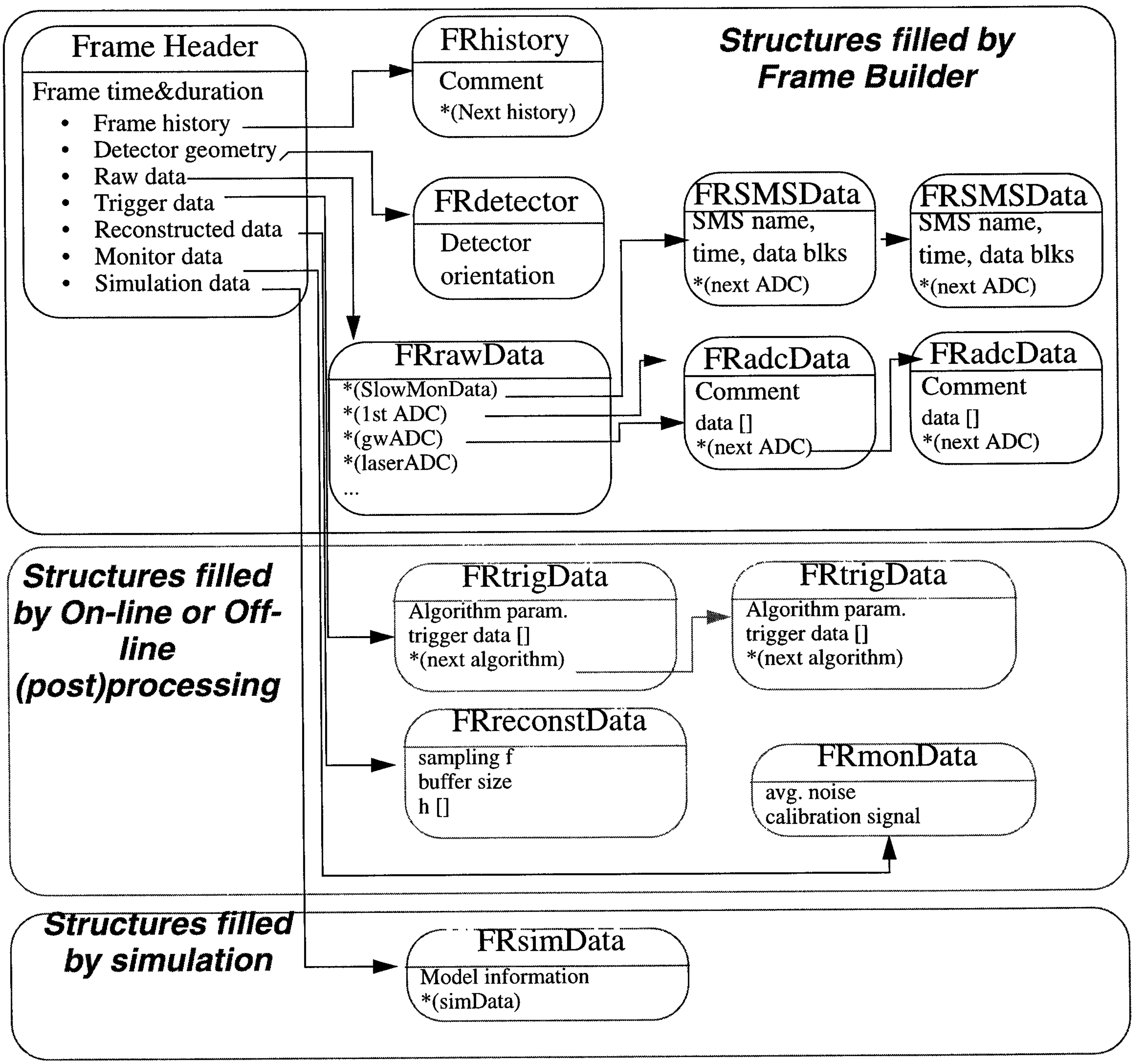
Status

- Initial meeting with VIRGO in April hosted by LIGO
  - VIRGO format presented, compared with LIGO needs
  - Attractive (to LIGO) because of maturity & availability of existing I/O libraries
  - Tuned for time-series data stream (vs. events or images)
- Alternatives explored by LIGO
  - Public domain standards - CDF/HDF
  - Used for image frame data distribution (NASA)
  - Greater overhead per frame than VIRGO
  - Well suited for eventual data distribution
- Continued interaction with VIRGO
  - Format evolving under collaborative effort
  - Software availability: commited to public domain access
  - Joint approach to be presented at TAMA Fall Meeting in Japan

LIGO-VIRGO DATA FORMATS

- PROPOSED FORMAT (Adopted from VIRGO)
  - ›› FRAMES (unit of information containing all information needed to understand the interferometer behavior over a finite time interval)
  - ›› C STRUCTURES (frames are organized as a set of C structures)
  - ›› FRAME HEADER (holds pointers to additional structures that contain all information)
  - ›› LINK LISTS (used to collect generic data types, PEM, ADC, etc.)
  - ›› HEADER HOOKS (pointing to frame elements used by on-line processing or by off-line reprocessing)
  - ›› 2^N DATA POINTS (allowing faster FFT analysis on individual frames)
  - ›› DICTIONARY (acts as a catalog of C structures and pointer offsets)

LIGO-VIRGO DATA FORMAT



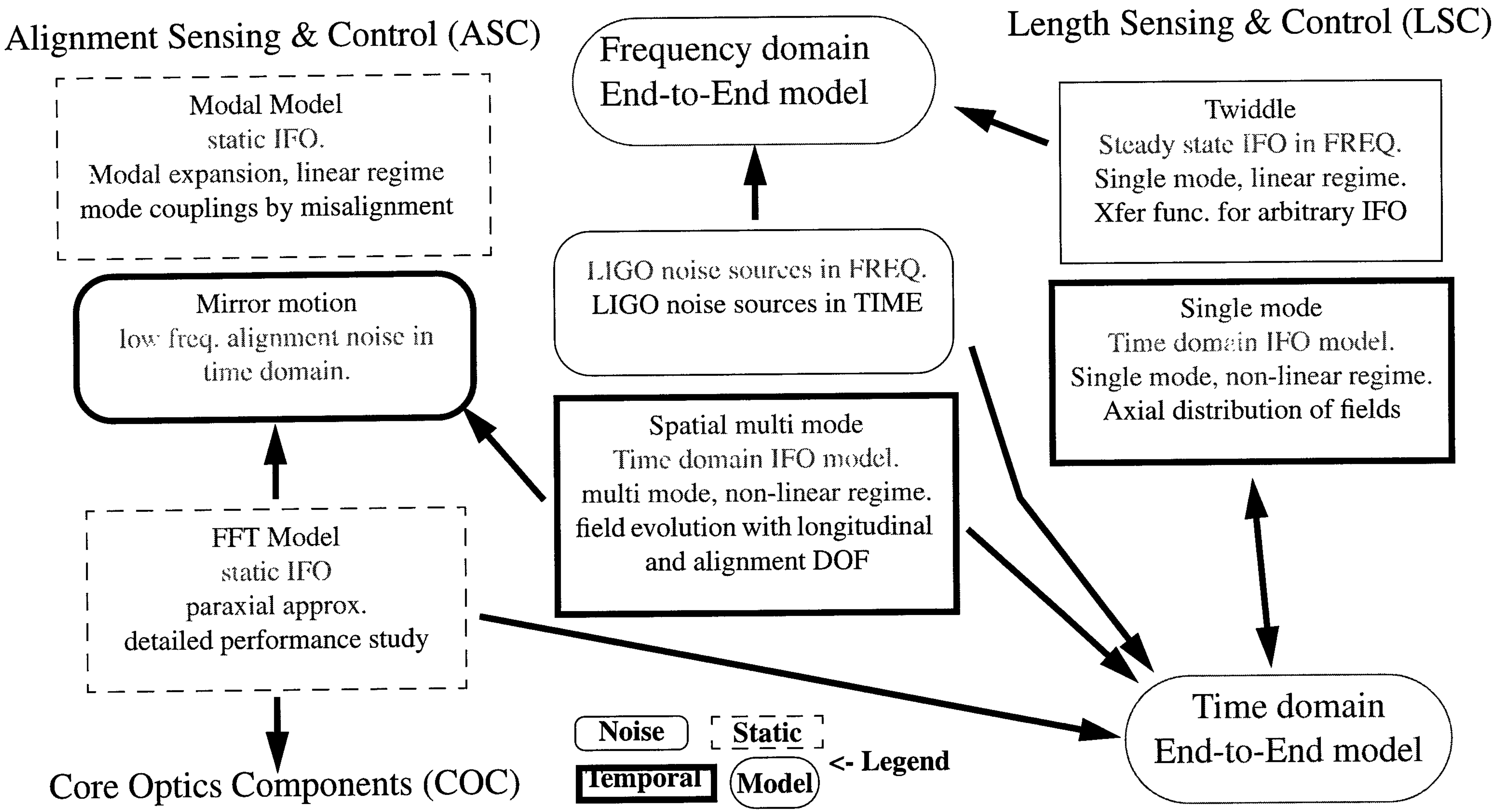
- **Frame has tree structure:**
- **Individual blocks are C structures**
- **Extensible to arbitrary length with design evolution**
- **Utilized for both on-line & off-line analyses**

LIGO-VIRGO DATA FORMAT

Issues

- Testing & verification at LIGO using I/O libraries for tape uncovered problems with C function calls between platforms (DEC vs. SUN)
  - LIGO wants to adhere to established software-hardware interface standards (i.e., POSIX) to minimize cost of code transportability/maintainability/upgrade/compatibility
- LIGO is discussing concerns with VIRGO; depending on outcome, LIGO may adopt VIRGO paradigm but implement its own code.
- Issue to be resolved by Spring 1997.
- ***QA is the key to code extensibility, adaptability& maintainability***

Modeling Activities
Overview



Modeling activities

- Time domain interferometer model with length and alignment D.O.F.
  - Objective
    - Demonstration of lock acquisition
    - dynamic stability of coupled alignment and length controllers
    - transfer functions between Length and Alignment DOF
    - Pseudo-data for noise analysis
  - Parallel efforts by D. Redding/JPL and R. Beausoleil/Cygnus
    - different approaches
    - model cross-validation
    - different application - speed vs. accuracy
  - D. Redding - time difference equations; iterative solution
    - Length (single D.O.F.) part complete
    - Used for the design of LSC
  - R. Beausoleil - forward time propagator kernel
    - single Fabry-Perot cavity with length and alignment DOF complete -- being validated

Modeling activities

- FFT model
  - ›› Detailed study of interferometer performance
    - e.g., sensitivity study of mirror phase/reflectivity error due to coating & polishing
  - ›› Code is parallelized
    - running on PARAGON in CACR - 10 x faster than SS20
  - ›› Interface improved:
    - GUI interface for input data
    - Remote scripting
    - Database for maintaining the run summaries

End-to-End model

- Frequency domain (steady state model) version
  - Interferometer: *Twiddle* by M.Regeher/H. Yamamoto
  - Noise models: K. Blackburn (& R. Weiss et al.)
- Transition to time domain
  - Interferometer and noise in freq. domain are essentially done
  - Control system for LIGO is still in design stage
  - Time domain model will be developed and is more suited when modeling control system
    - Time domain IFO model with length and angular DOF - JPL/Cygnus models
    - Time domain noise models need to be developed
- Time domain version - just started
  - First target is 40 m testbed - serve as a prototype for the full version
  - Inclusion of control system - use design for 40m recycling
  - Include fundamental building blocks for LIGO

List of acronyms in the order they appear in presentation

VIRGO	Franco-Italian Laser Interferometer Collaboration
CDS/DAQ	Computer & Data Systems (part of Detector) Data Acquisition System
NSF	National Science Foundation
WAN/LAN	Wide/Local Area (Computer) Network
PAC	LIGO Program Advisory Committee (yet to be formed)
LRC	LIGO Research Community
NS	Neutron Star; BHBlack Hole
s	Second (T= time)
kB/MB /GB/TB	kilo-/mega-/giga/terabyte: 10^3/10^6/10^9/10^12 bytes
kFLOPS /MFLOP /GFLOPS	kilo/mega/giga Floating Point Operations per Second
PEM	LIGO Physical Environment Monitoring system
SNR	Signal to Noise Ratio
IFO	InterFerOmeter
kpc	3 x 10^3 lightyear (kiloparsec)

GW	Gravitational Wave
LA,WA	Louisiana, Washington sites (also, Hanford, Livingston)
1x/2x/3x	notation for single, double, and three-fold coincidence operational modes of the LIGO detector comprising of 3 interferometers (IFOs)
FIFO	First In First Out method of reading data written to dynamic memory
CPU	Central Processing Unit (a generic computer processor)
DSP	Digital Signal Processor-specialized CPU efficient at particular algorithmic calculations (FFTs)
FFT	Fast (Discrete) Fourier Transform
ATM	Asynchronous Transfer Mode; a protocol for inter-processor communications
HEP	High Energy Physics
DOE/ESNET	Dept. of Energy/ Energy Sciences Network
T1/T3/ ATM/OC12	Various telecommunications channels and bandwidths, approximately: T1: 200 kB/s; T3: 6 MB/s; ATM: 20MB/s; OC12: 70 MB/s.
IPAC/JPL	Image Processing and Analysis Center (at Caltech) /NASA Jet Propulsion laboratory
CACR	Center for Advanced Computing Research @ Caltech

VBNS	NSF counterpart to DOE's ESNET.
LSU	Louisiana State University
SC	Supercomputer(ing) Center(s)
HNR/BNWL	Hanford Nuclear Reservation (LIGO Site)/ Battelle Northwest Laboratories.
DRR	Design Readiness Review
PDR	Preliminary Design Review
FDR	Final Design Review
SUR	IBM's Sponsored University Research Grants Program
COTS	"Commercial, Off-the-shelf Software" and acronym for "buy versus make" when deciding how to develop software.
CDF/HDF	Common/Hierarchical Data Format-Image distribution data formats available in public domain.
TAMA	Japanese Interferometric Gravitational Wave Detector Project
ADC	Analog-to-digital convertor
DEC/SUN	Computer manufacturers: Digital Equipment Corp/ Sun Microsystems, Inc.
POSIX	Established industry standard for software/hardware interfaces.

DOF	Degree(s) Of Freedom
TWIDDLE	Name of a particular modeling code within LIGO.
ISC /ASC/LSC	Interferometer/Alignment/Length Sensing & Control Systems
SS20	SunSparc 20 Workstation