LIGO Data Analysis

Data Formats

and

Modeling Activities

Albert Lazzarini
LIGO Integration Group

NSF Fall Review
22 - 24 October 1996

LIGO-G960211-00-E

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	Storage	Comments
Burst Signals $\Delta T < 1s$	Supernovae	$\Re_0 \sim 2-3/$ yr @ 15 Mpc If sufficiently asymmetric	Minimal for straightforward correlation; *if optimal filters are discovered, problem may increase in complexity.*	Minimal Need PEM/houskeeping 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	$\Re_0 \sim 1/$ yr(?) @ 500 Mpc; $M_{BH} \sim 30 - 200 M_{SUN}$			
Chirped Waveform $10s < \Delta T < 1000s$	NS/NS Inspirals	$\Re_0 \sim 3/$ yr @ 23 Mpc; $\Delta T \sim 4 \times 60$ s $M_{NS} \sim M_{SUN}$ $\Delta T \sim 4 \times 500$ s $M_{NS} \sim 0.3 M_{SUN}$	Templates/Data ~ 2 GFLOPS ~ 50 GFLOPS	~20 GB /~1 GB ~500 GB /~10 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
	BH/BH Inspirals	$\Re_0 \sim 1/$ yr @ 150 Mpc; $\Delta T \sim 4 \times 10$ s $M_{NS} \sim 10 M_{SUN}$;	~ 2 GFLOPS	~20 GB /~1 GB	

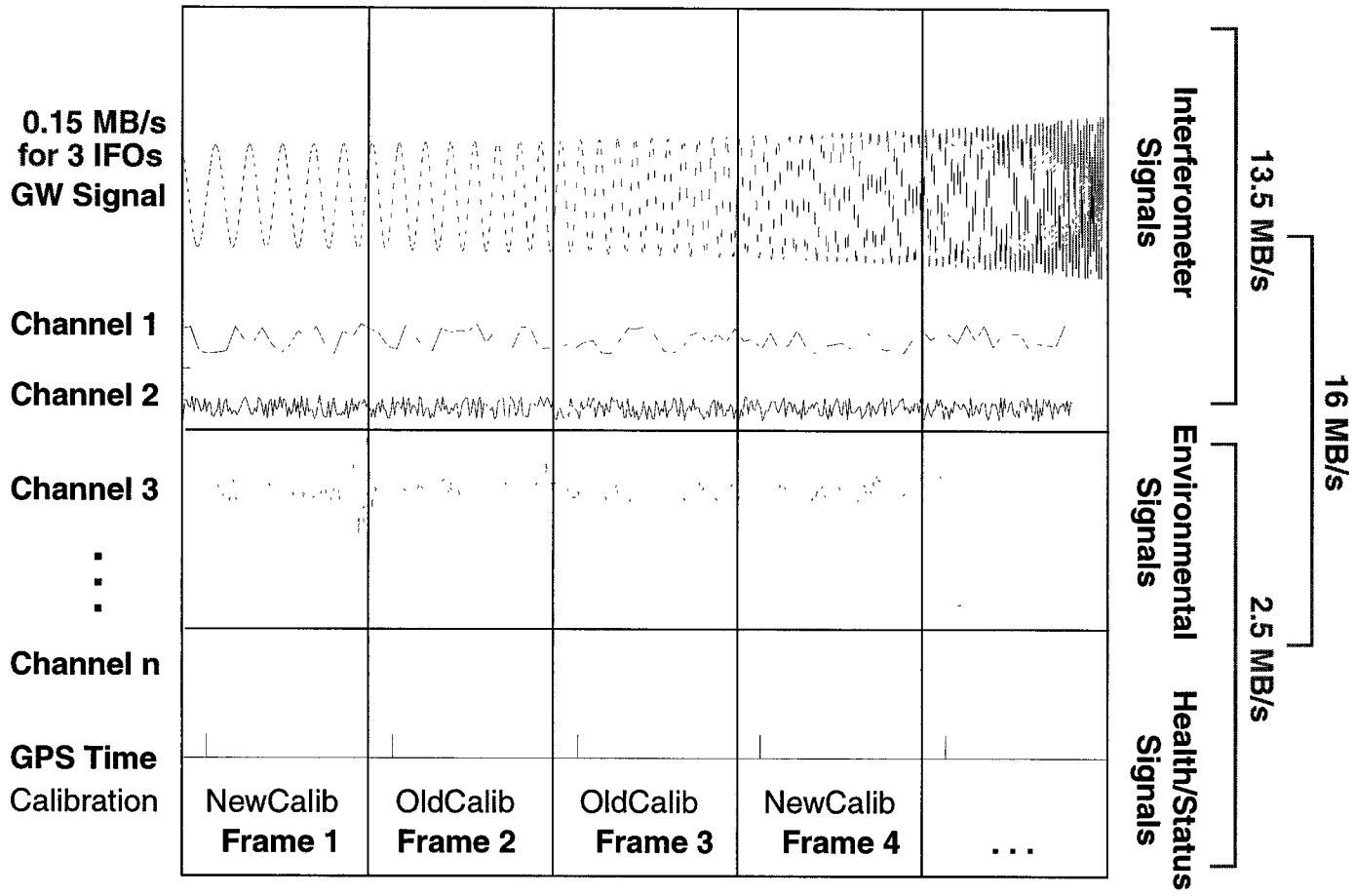
Data Analysis Requirements

Science & Computational Requirements

Initial LIGO Sources and Estimated Analysis Capability Requirements

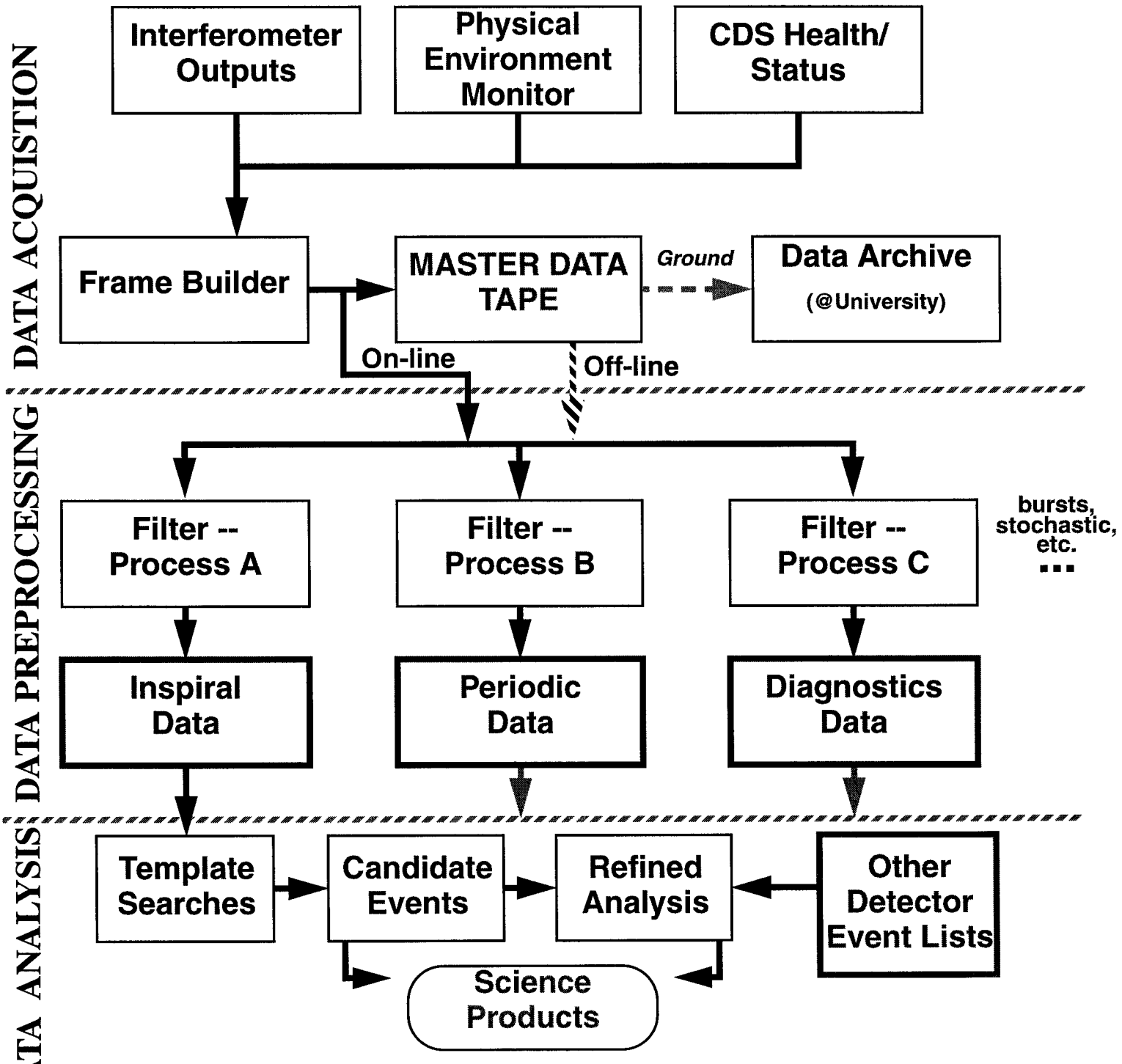
	Sources	Initial LIGO Performance Estimate	Data Analysis Requirements		
			CPU	Storage	Comments
Periodic Signal $\Delta T \sim 10^6 - 10^7$ s	Pulsars with mass asymmetry $h \propto \left(\frac{\varepsilon}{10^{-6}}\right)\left(\frac{10\text{kpc}}{r}\right)\left(\frac{1\text{ms}}{P}\right)^2$	$\varepsilon = 3\times10^{-5}$; 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 ($\Delta T < 1s$) -- 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*

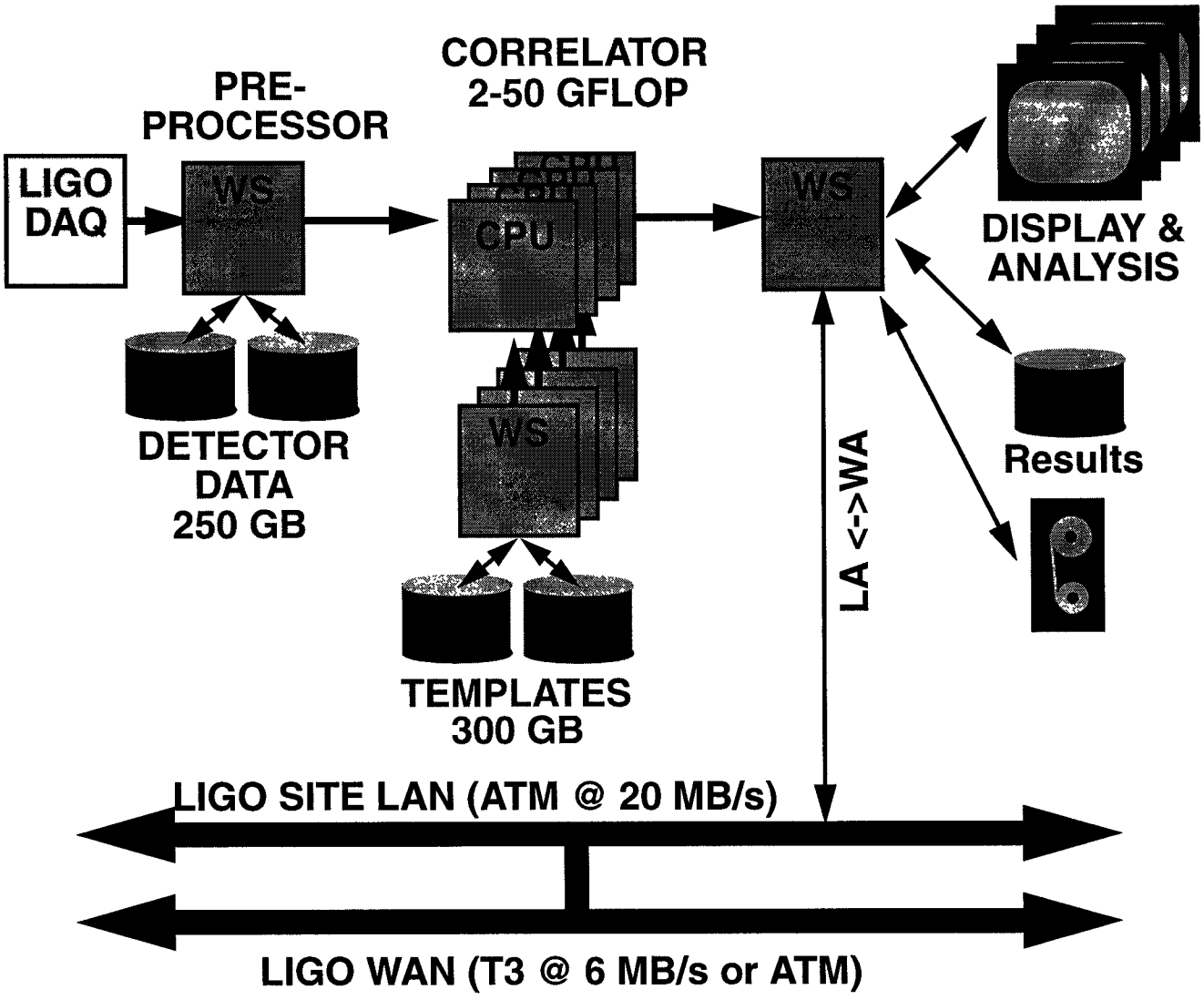
Data Analysis for Initial LIGO

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.

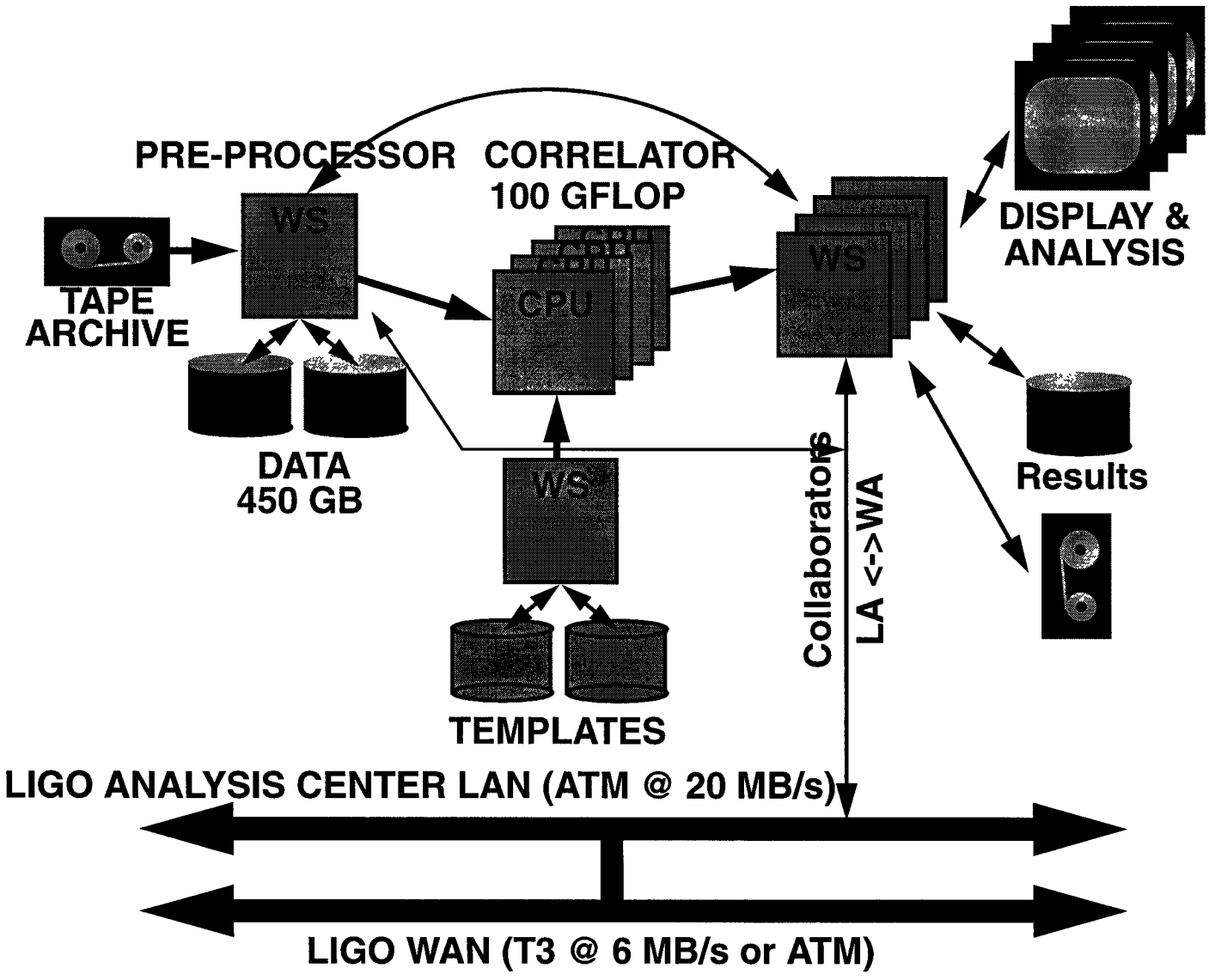
Data Analysis for Initial LIGO

Off-line Processing Computing Resources & Distribution

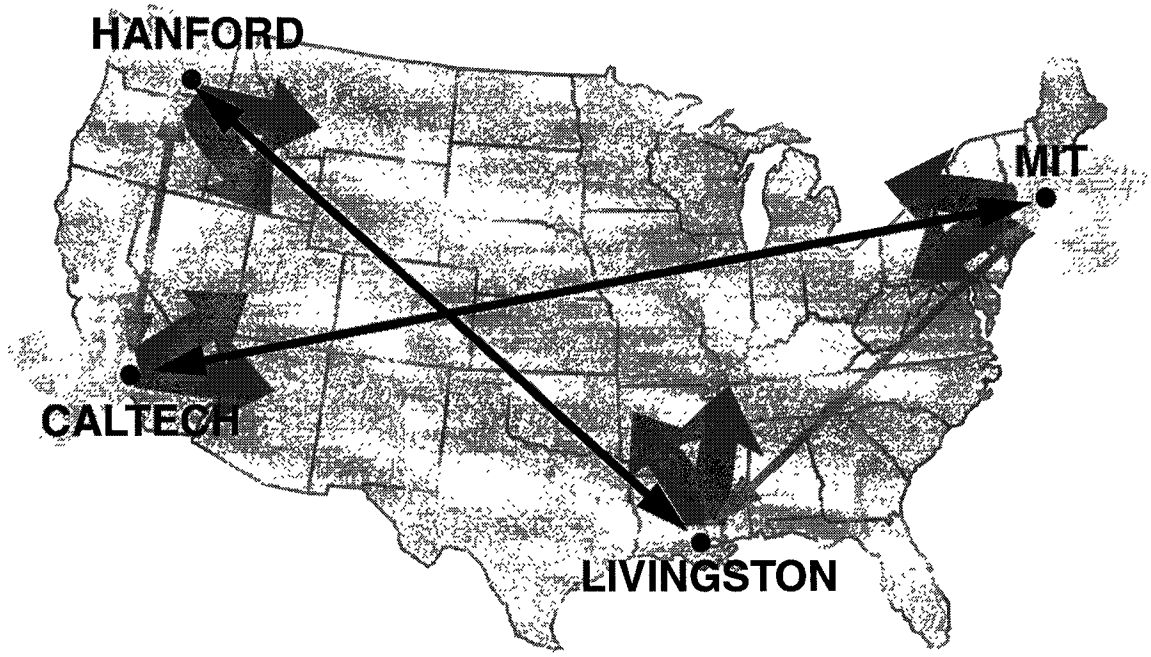
- System configuration (target: max. capability for multiple users)
  - Large data archive
    (~ 500 TB/yr => 10k tapes/yr @ 50 GB/tape => $0.5M/yr @ $50/tape)
  - Robotic tape access -- size TBD
  - Disc cache system capable of storing 450GB of data
    - 8 hours of 100% data ~ 450 GB
    - ~ 5 weeks 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		
	3/99-12/99	Definition, Agreements, Implementation	Procurement & Integration	Development, Verification
	11/98		Specifications	
System FDR	11/98		Design & Prototyping	Specifications
System PDR	11/97			Design & Prototyping
System DRR	5/97			

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