I: Data from the LIGO I Science Run

II: GriPhyN: The Grid Physics Network

Relevance to LIGO Data Analysis

Drexel University Workshop on Astronomical Sources

Philadelphia,Pennsylvania
30 October 2000

Albert Lazzarini
LIGO Laboratory
California Institute of Technology
Pasadena, California 91125

LIGO-G000315-00-E

Data from the LIGO I Science Run

Drexel University Workshop on Astronomical Sources

Philadelphia,Pennsylvania
30 October 2000

Albert Lazzarini
LIGO Laboratory
California Institute of Technology
Pasadena, California 91125

LIGO-G000315-00-E

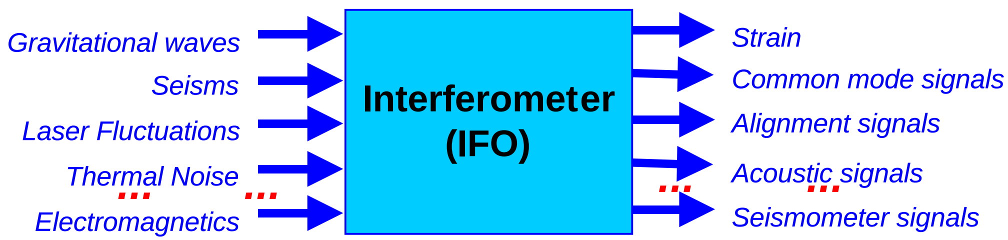
LIGO I Data
Overview

- Data stream from the interferometers
- Pre-processing, data conditioning
- Data analysis systems
  - » At the observatories (on-site, near real time)
  - » At the universities (off-site)
- Data products, volumes
- Data access
  - » Tools
  - » Policies

What are the Data?

- Continuous time series
  - » 2^N samples/second, 16 bit
- Data analysis: digital signal processing
- Analysis performed in both time/fourier domain
  - » Single channel, over a long time; many channels, over a short time
  - » How to cache, catalog, replicate, this ***virtual data***
- Results of analysis: events, spectra, N-D representations ("images")
  - » Environmental, instrumental "events": vetoes
  - » Astrophysical events
  - » Time stamp, Process ID generating event, Parameters associated with event, ...
  - » Stored in a relational database for later retrieval, reanalysis:
    - – tables, "blobs", links to data

Raw Data Stream Characteristics

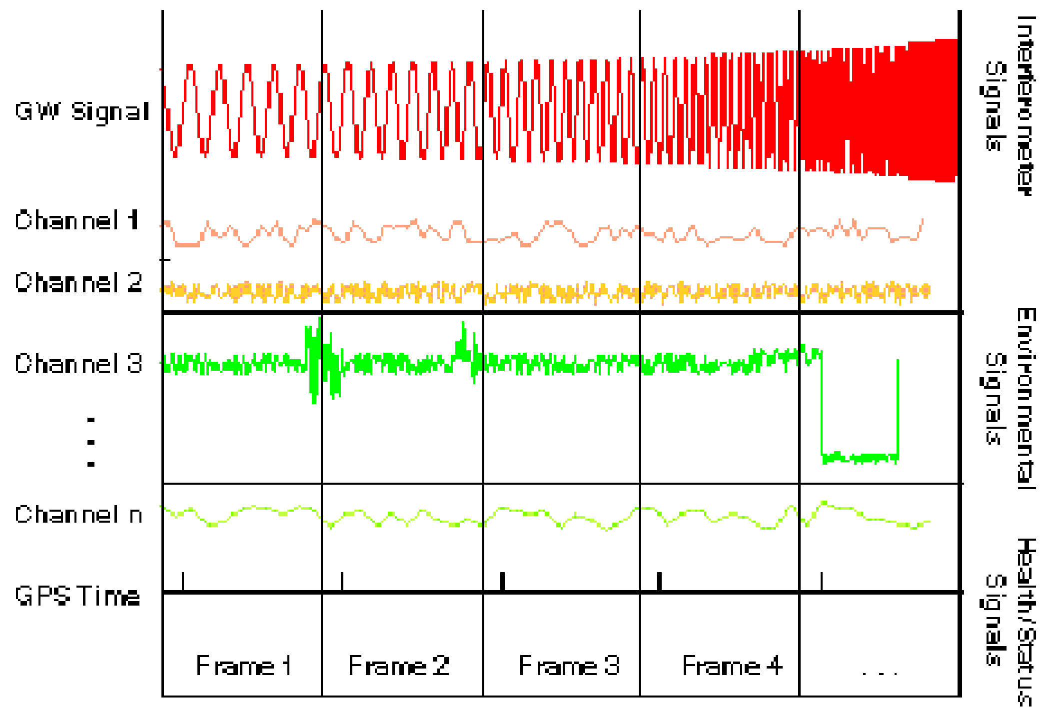


Science Channel (1.4%)

Interferometer Channels (81%)

Environmental Channels (17 %)

- **All interferometric detector projects have agreed on a standard data format**
- **Anticipates joint data analysis**
- **LIGO frames for 3 interferometer are ~ 7MB/s**
  - **96 kB/s strain (2 Bytes x 3 IFOs x 16 kSample/s)**
  - **~ 5.7 MB/s other interferometer signals**
  - **~ 1.2 MB/s environmental sensors**
  - ***Strain is ~1.4% of all data***

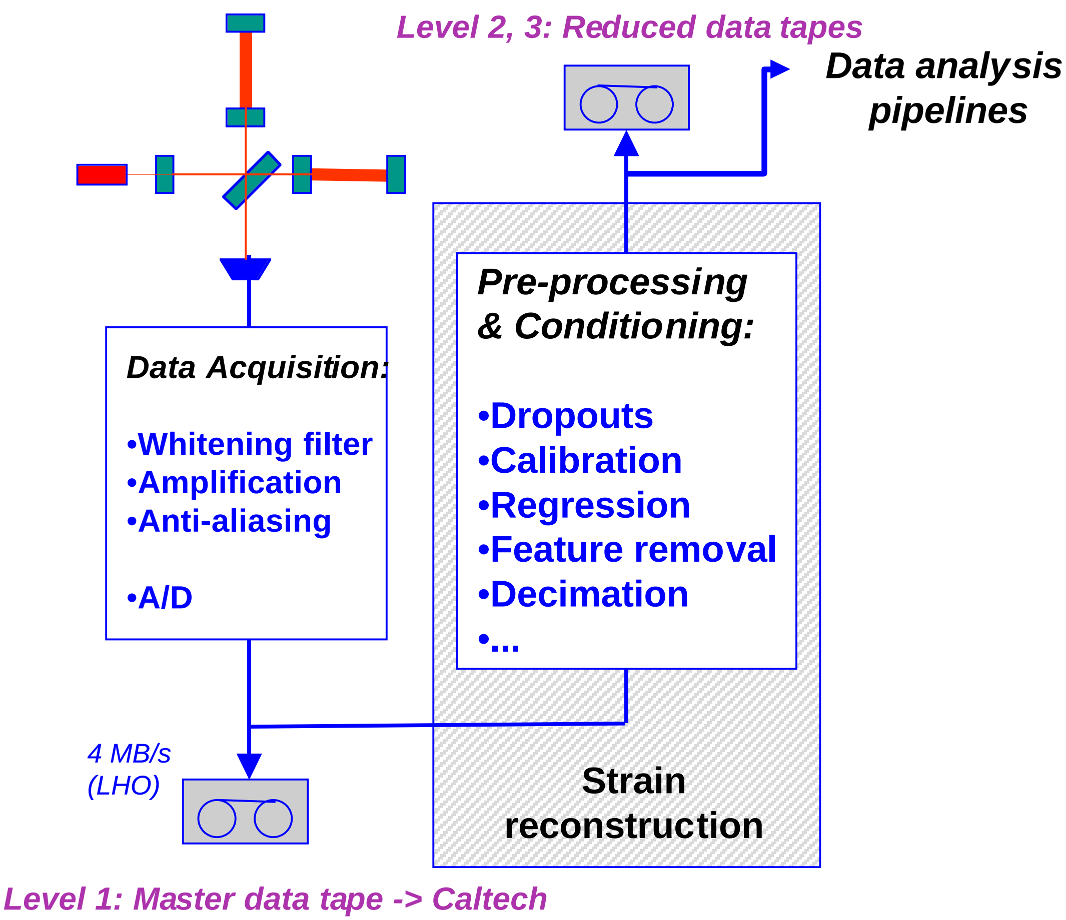


LIGO-G000315-00-E

LIGO I Data Channel Count
by Acquisition Rate

Acquisition Rate, samples/second (16 bit)	Number of Channels
16834	124
2048	532
256	109
64	205
16	208
Total No. of Channels:	1178

Data pre-processing at observatories



- Master data tapes transported to Caltech for deep archive (HPSS)
- Reduced data tapes provide reduced bandwidth sample of data stream; needed for search algorithms
- Whitening required due to dynamic range of signals
- Regression & feature removal reduces RMS, dynamic range from narrowband line features
- Resampling & decimation matches data rate to search bandwidth
- Calibration provides physical strain

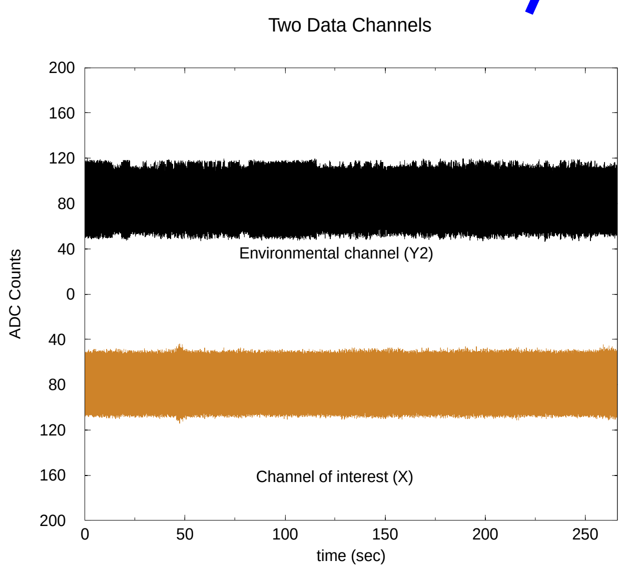
LIGO-G000315-00-E

Data Pre-processing: removing instrumental effects

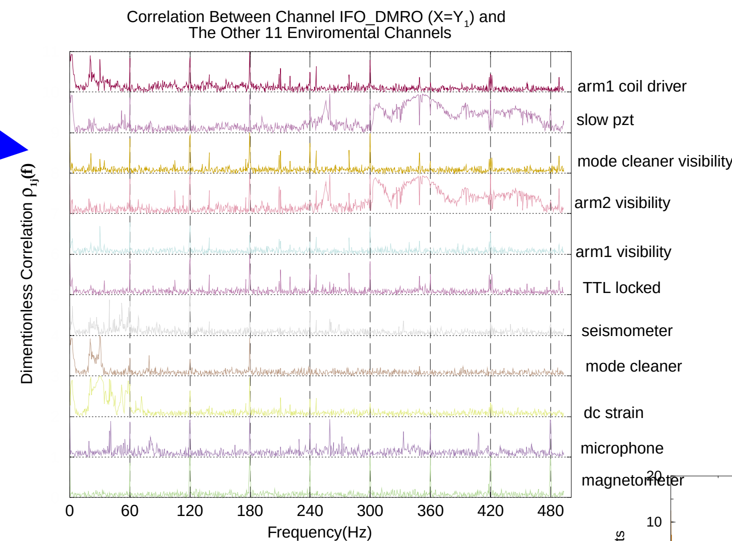
- Cross channel regression will be used to improve signal to noise ratios when possible (need adequate SNR)

Raw channel data (40m prototype)

$$s_a(t) \Rightarrow \hat{s}_a(f)$$
$$s_b(t) \Rightarrow \hat{s}_b(f)$$



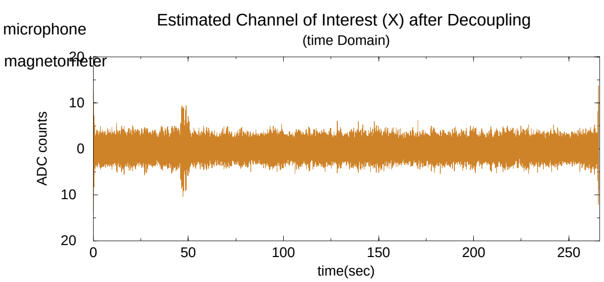
LIGO-G000315-00-E



ref: Allen, Hua, Ottewill (gr-qc/9909083)

Cross channel spectral correlation

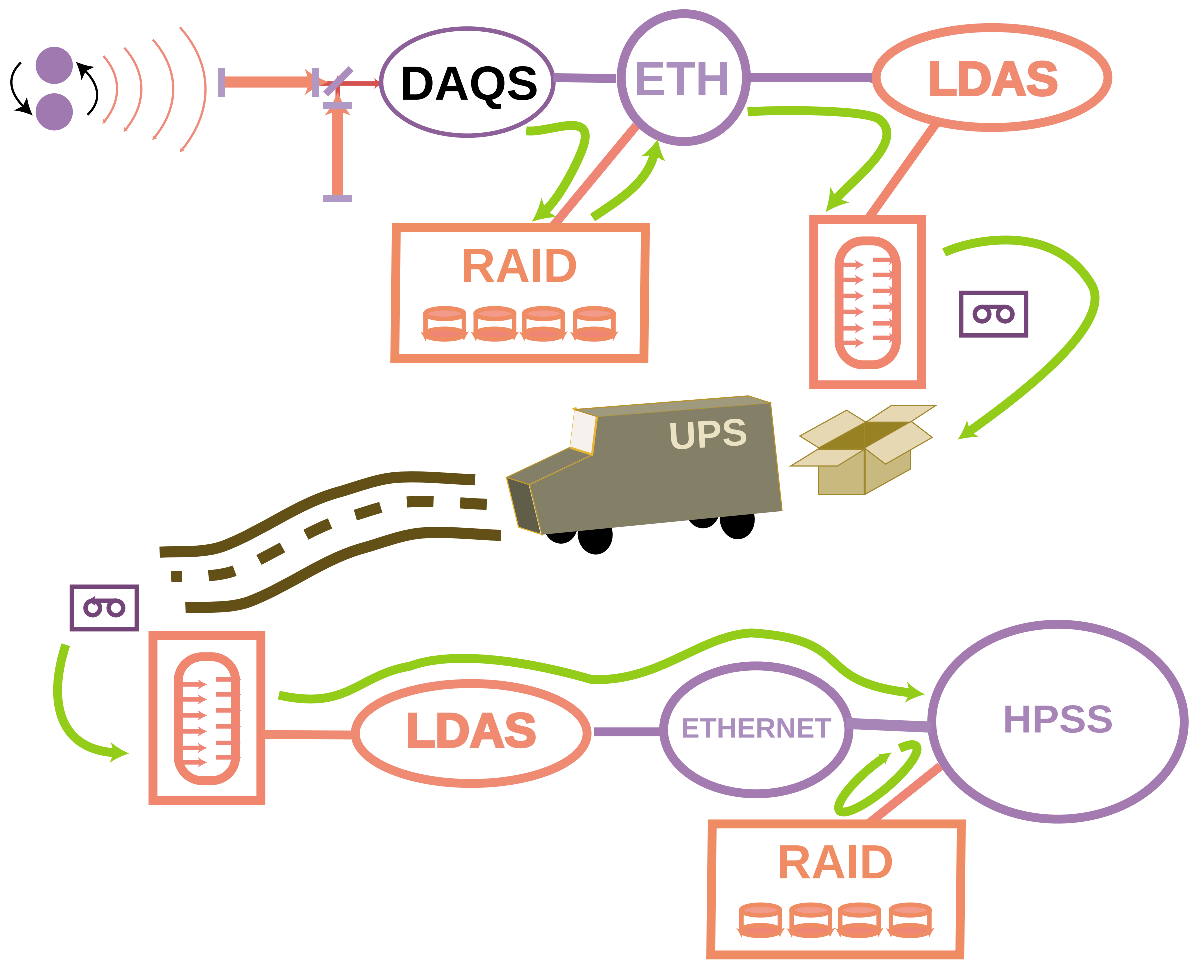
$$\Gamma_{ab}(f) \equiv \frac{\left|\hat{s}_{ab}(f)\right|^2}{\hat{s}_a(f)\ \hat{s}_b(f)}$$



$$\begin{pmatrix} \hat{s}'_a(t) \\ \hat{s}'_b(t) \end{pmatrix} \Leftarrow \begin{pmatrix} \hat{s}'_a(f) \\ \hat{s}'_b(f) \end{pmatrix} = M(f) \cdot \begin{pmatrix} \hat{s}_a(f) \\ \hat{s}_b(f) \end{pmatrix}$$

Reduced data channel

LIGO Data Flow



**** LIGO plans a future upgrade to OC3 links from both observatories to archive to obviate shipping of tapes***

LIGO-G000315-00-E

LIGO Data Products - time series data

Mode	Raw and Derived Data for On-line Diagnostics	**Level 1** Full (100%) frame data for archiving	**Level 2** Strain and data summary, QA channels	**Level 3** Strain best estimate
Uncompressed Rate (MB/s)	LHO: 9.479 LLO: 4.676 Total: 14.155	LHO: 4.698 LLO: 2.278 Total: 6.975	Total: 0.300	Total: 0.006
w / 50% Hardware Compression MB/s onto tape media	-	LHO: 2.349 LLO: 1.139 Total:3.488	Total: 0.150	-
Data growth rate, per year of integrated running, ***TB/yr***.	-	LHO: 74 LLO: 36 Total:110	Total:9.5	***Total: 0.200***
Total including redundant 100% backup, ***TB/yr***.	-	LHO: 148 LLO: 72 ***Total:220***	***Total:19***	-
Purpose	For on-line monitoring of interferometers	Deep permanent archive	Science analysis, data exchange	Science analysis, data exchange
On-site look-back time	Must use real-time control and monitoring system (CDS) disk caches	LHO Disk cache: 60 hr LHO Tape robot: 49 d LLO Disk cache: 60 hr LLO Tape robot: 100 d	-	-
Off-site look-back time	-	As long as required	In perpetuity	In perpetuity

LIGO

Times series data uses

- Collaboration-wide searches (Lab resources)
  - » On-site at observatories: LIGO Data Analysis System (LDAS)
    - – 7x24 pipeline analysis to provide first pass through data
    - – Events (both instrumental vetoes from on-line monitors and astrophysical events from pipeline) registered in database
    - – Single-interferometer detections
    - – Near-real time information (e.g., SNe bursts, …)
  - » Off-site at Caltech: LIGO Data Analysis System (LDAS)
    - – Data ingestion into deep archive; mirroring of site event databases
    - – Pipeline analysis to provide second pass through data
      - Follow-up to on-site first passes
      - Multiple interferometers
    - – Events (both instrumental vetoes from on-line monitors and astrophysical events from pipeline) registered in database

Times series data uses

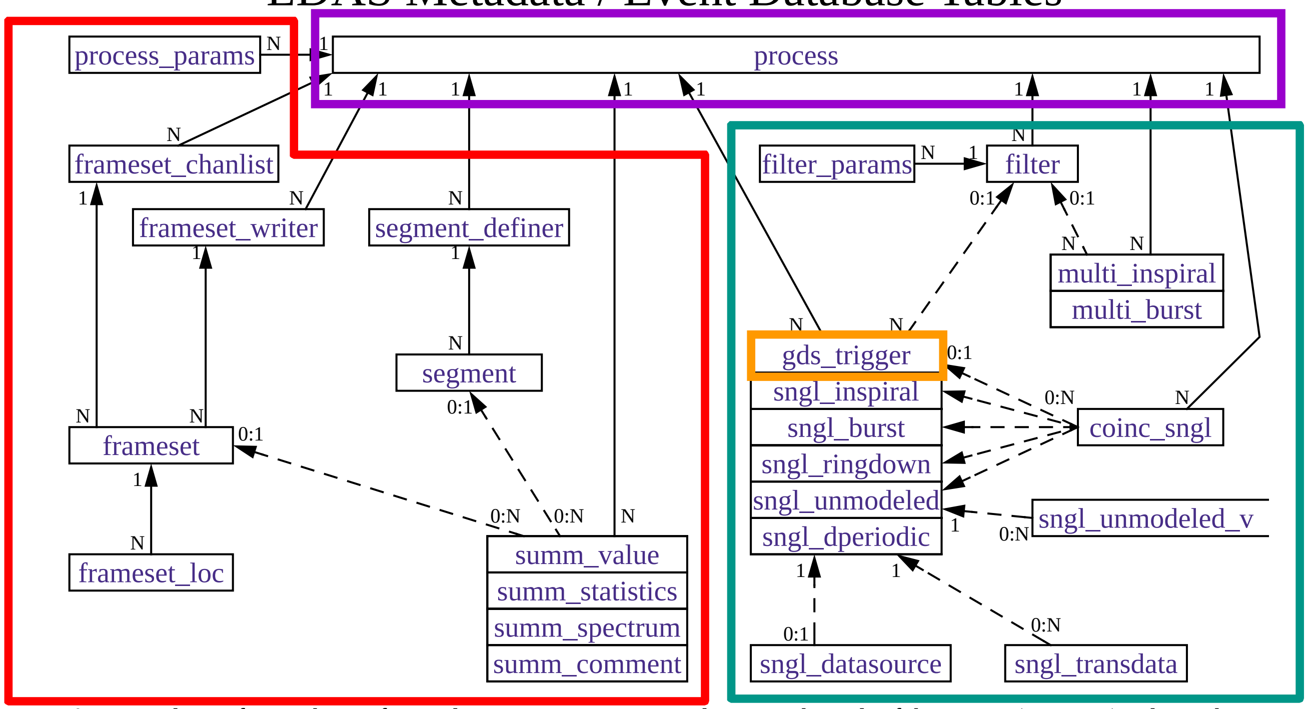
- Individual exploratory research (institutional resources)
  - » Reduced data sets available from Caltech archive
    - – Binary frame format
    - – LIGO-Lightweight data format (XML)
  - » Download to locally owned,managed resources for exploratory research
    - – Internet (small data sets), ftp, pftp
    - – Tapes (larger data sets)
  - » Analysis environments:
    - – Commercial tools
      - Matlab, Mathematica, IDL, …
    - – Replica installation of LDAS tools, APIs
      - Other LSC institutions
      - *Off-lline LDAS Development & Test systems at Caltech*
    - – Prototype tools, public domain code -- ROOT, GRASP, ...

Data products: Event database

- Event source (processes or filters)
- Raw data characteristics/location
- Raw data statistics
- Instrumental triggers (vetoes)
- Astrophysical search triggers
  - Single interferometer
  - Multiple interfertometers

LDAS Metadata / Event Database Tables

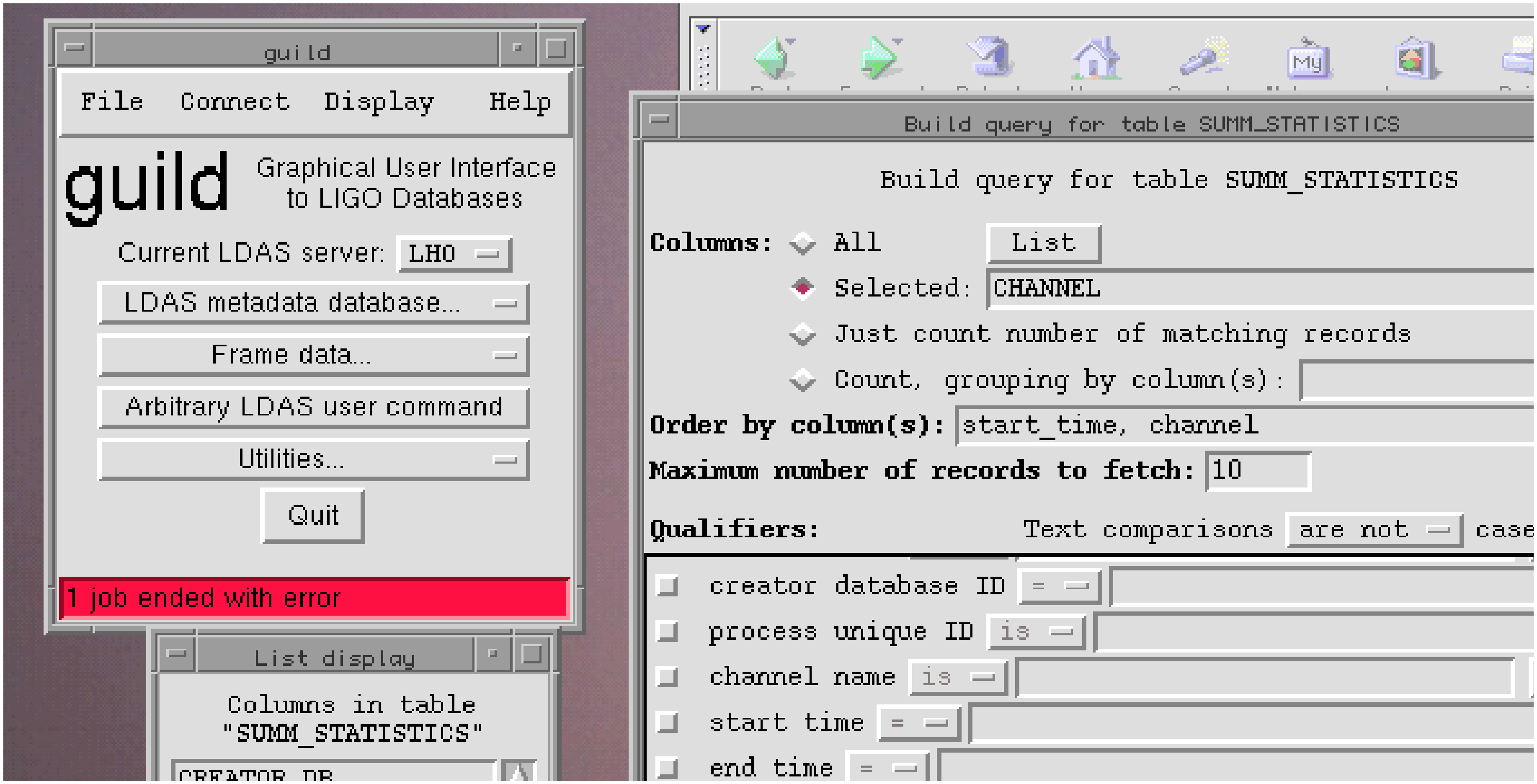
PSS 21 Nov 1999



Arrows indicate foreign key referential integrity constraints. Values near the ends of the arrows (1, N, etc.) indicate the possible multiplicities. Dashed lines indicate optional relationships. Stacked tables (grouped by thick lines) have common relationships with other tables, except for relationship arrows connecting along the right edge. Examples: 1) Each segment is related to one segment_definer; 2) Each segment_definer is (generally) related to many segments; 3) A frameset is related to one frameset_chanlist entry and to one frameset_writer; 3) A summ_value (or summ_statistics, etc.) entry may or may not be related to a segment and/or a frameset; 4)A single-interferometer event (gds_trigger, sngl_inspiral, etc.) entry may be related to up to one sngl_datasource and/or any number of sngl_transdata entries.

LIGO-G000315-00-E

GUILD: Database Query Tool



Event data uses

- Collaboration research, individual exploratory research
- Event data available from observatories (recent) or Caltech DB archive (long term)
  - – LIGO-Lightweight data format (XML)
- Use LDAS resources to query DB.
  - » Download to locally owned,managed resources for exploratory research
    - – Internet (small data sets), ftp, pftp
  - » Analysis environments:
    - – Commercial tools
      - Matlab, Mathematica, IDL, …
    - – Replica installation of LDAS tools, APIs
      - Other LSC institutions
      - *Off-lline LDAS Development & Test systems at Caltech*
    - – Prototype tools, public domain code -- ROOT, GRASP, ...

LIGO Science Run Data Availability

- Data will generally not be placed in the public domain:
  - » Instrumental idiosyncrasies require collaborators with intimate working knowledge of the interferometers
  - » Maintaining a public domain archive service is not in LIGO Lab. Operational scope at present
- Access to LIGO I Science Data through the LIGO Science Collaboration (LSC)
  - » LSC formed by LIGO in 1996 at the recommendation of NSF's *Panel on the Future Uses of LIGO*
  - » Individuals join by establishing an MOU with LIGO defining scope of collaborative work
    - – ***26 institutions,***
    - – ***245 people participating in LIGO I science***

LIGO Science Run Data Availability

- LSC plans to participate in SNEWS
  - » Supernova Early Warning System (neutrino expts.)
  - » Early participation will be to subscribe to trigger distribution

GriPhyN: The Grid Physics Network

Relevance to LIGO Data Analysis

Drexel University Workshop on Astronomical Sources

Philadelphia,Pennsylvania
30 October 2000

Albert Lazzarini
LIGO Laboratory
California Institute of Technology
Pasadena, California 91125

LIGO-G000315-00-E

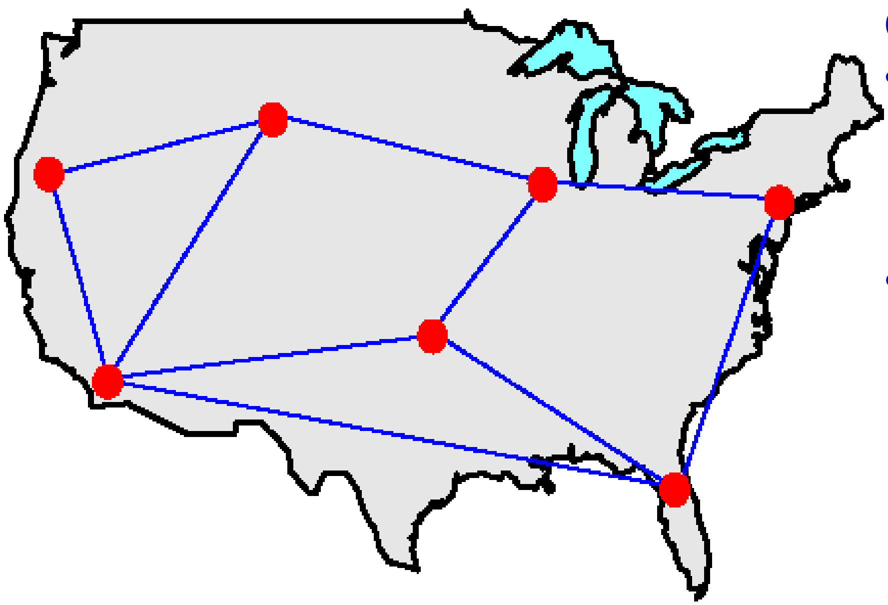
GriPhyN & LIGO Data Analysis

Overview

- GriPhyN concept, organization
- LIGO/LSC within GriPhyN
- LIGO data analysis challenges for GriPhyN
- GriPhyN hardware for LIGO/LSC
- Status

The GriPhyN Concept

- GriPhyN = Grid Physics Network
- Vision: build production-scale Computational Grids
  - » Mobilize large-scale IT resources for scientific research
  - » Emphasis on massive data movement, high-speed networks
  - » “Data Grid” rather than “Computational Grid”



Collaboration among:

•4 Multi-disciplinary experiments:

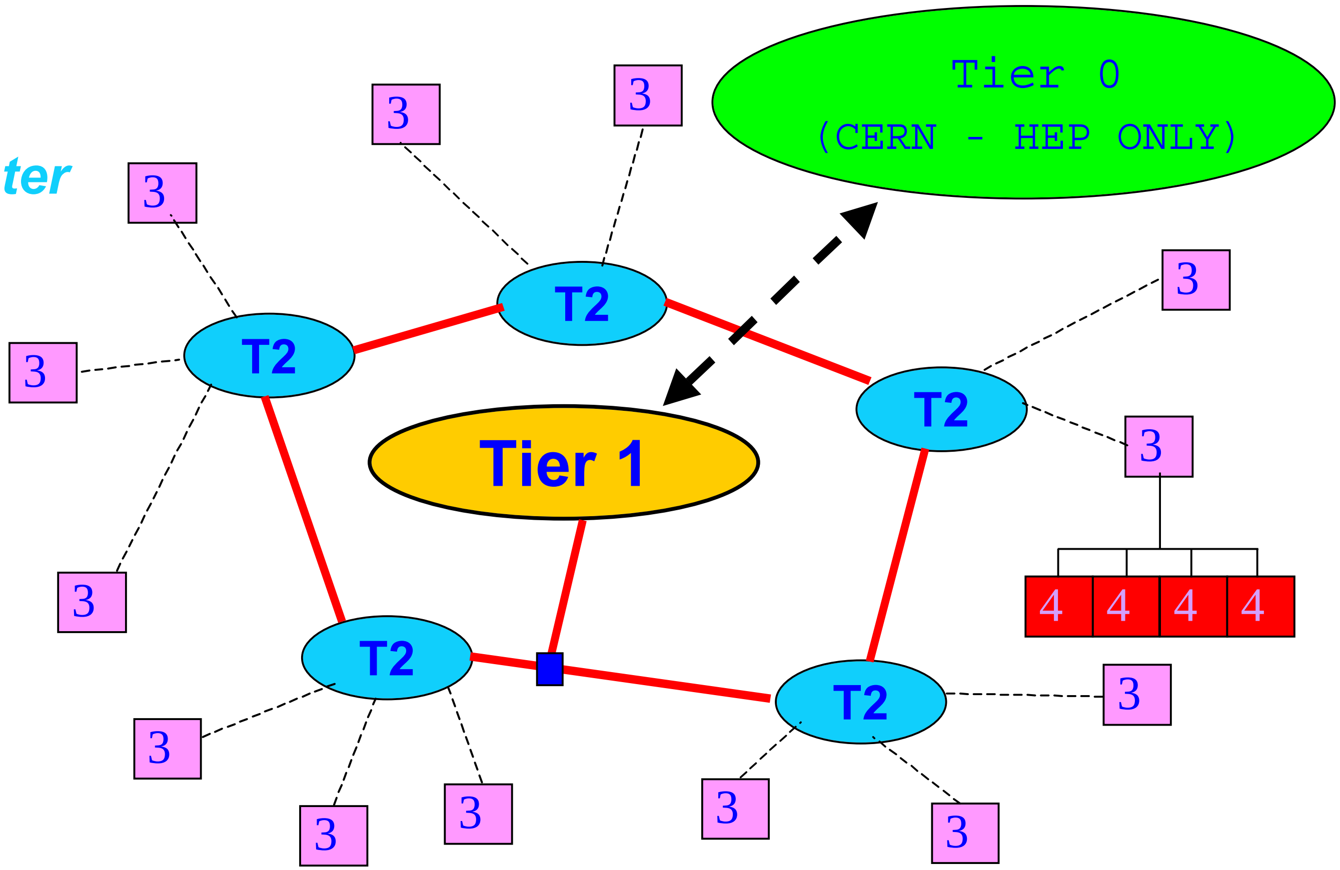
- **•HEP (CMS, ATLAS @ LHC/CERN)**
- **•Gravitational Physics (LIGO/LSC)**
- **•Astronomy (SDSS)**

•Computer Science

- **•SDSC**
- **•ANL/UC**
- **•USC**
- **•LBL**
- **•UW/Madison**
- **• ...**

Data Grid Hierarchy

- *Tier0 CERN*
- *Tier1 National Lab*
- *Tier2 Regional Center*
- *Tier3 University*
- *Tier4 Workstation*



LIGO-G000315-00-E

LIGO/LSC Organization Within GriPhyN

- LIGO Laboratory (Tier 1)
  - » Caltech/MIT principals under NSF Cooperative Agreement
  - » Sites at Hanford, WA and Livingston, LA
- LIGO Science Collaboration (Tier 2)
  - » LIGO Science Collaboration
  - » 26 institutions
  - » ~ 350 people

What are the LIGO Data?

- Continuous Time series
  - » 16 kHz, 160 Hz, 1 Hz....
- 1% Gravitational-Wave channel, plus
- 99% other channels
  - » Environment: Seismometer, Microphone, Magnetometer, ...
  - » Engineering, Housekeeping, Health, Status, ...
- Analysis performed in both Time/Fourier domain
  - » One channel, long time or Many channel, short time
  - » How to cache, catalog, replicate, this virtual data
  - » Need CS wisdom!

LIGO data processing challenges

- Signal processing of “all data”
  - e.g.: [5-50 Mflop/byte] for inspiral search of GW channel
  - x [0.2 TB] total cleaned GW channel for LIGO I
  - System-based (LDAS pipelines)
  - Menu-based (standard toolboxes, interfaces)
  - Personal filters (individual exploratory research with data, LIGO Algorithm Library)
  - Estimating required resources
- LIGO archive (200 TB)
  - Transposed, Reduced, Filtered & other caching
  - Metadata replicas [2 TB]
  - Clients requesting data
  - Clients adding data

LIGO data processing challenges

- Search for periodic sources
  - » Very long Fourier Transforms
    - – e.g., 1 kHz for 10 days => ~ 10^9 point FFT
  - » Need to try every sky direction, frequency, d^n/dt^n [frequency]
  - » 10s - 100s Tflops required to cover parameter space

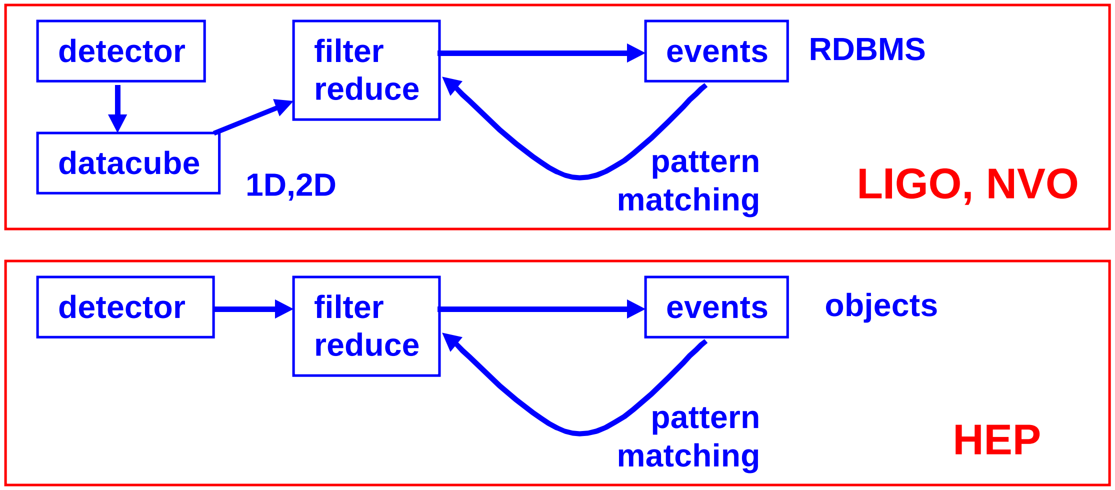
- Wide Area Networking
  - » New data from observatories
  - » Coincidence network analysis
    - – Virgo (France/Italy),
    - – TAMA (Japan),
    - – GEO (Germany/UK)

Services GriPhyN can provide

- Distributed Computing Power
  - » Code development sandbox
    - – Also menu & parameter driven processing
  - » Compute-intensive background jobs
    - – *"Pulsar@GriPhyN"* project
  - » How to make code portable within GriPhyN
- Virtual Data
  - » Data, Catalog, Reduced Data, Mirror
  - » From browsing to "all data"
  - » Data transformation

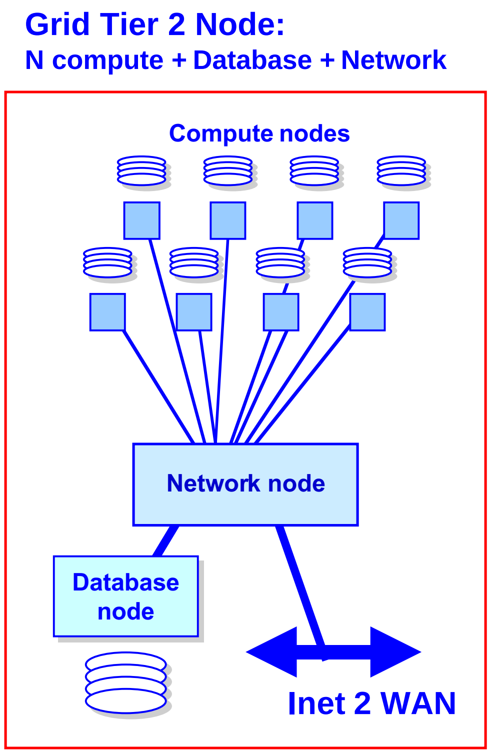
LIGO vs NVO vs HEP

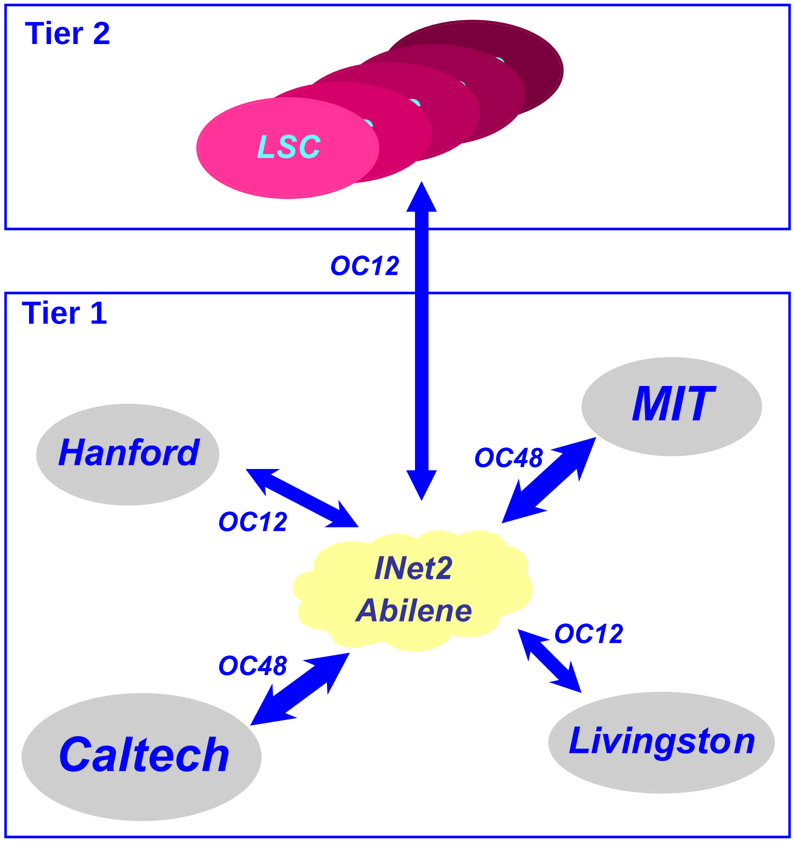
- Processing



- Network
  - » NVO is **federation**, LIGO is **coincidence**
  - » In both cases **registration** is important

GriPhyN Tier 2 Hardware





LIGO-G000315-00-E

Example Tier 2 Configuration

Typical Tier 2 Grid Node -- "*Datawulf*" machine

- Compute Nodes
  - » 48 beowulf nodes (1GHz CPUs)
  - » Each with 150 GB disk (7.2 TB total)
  - » Gigabit switch for cross-node communication
- Database server
  - » DB2 database
  - » 300 GB disk cache
  - » Multiple redundant path to grid
- Network node
  - » OC12 access to internet

GriPhyN Phase I

- 5-year, $11.9M NSF/ITR project to research, develop:
  - » Application-specific instances of Grid use (4 physics experiments, each with different need)
  - » Virtual data tookits for services required of grid
  - » Computer Science research
    - – Execution management, performance analysis, request scheduling & planning, virtual data management, ...
  - » Outreach -- promote greater involvement in the development of the US grid

GriPhyN Phase I

- 14 institutions
  - » HEP: Caltech, ANL, Harvard, UPa
  - » Gravitational physics: Caltech , UT/Brownsville, UW/Milwaukee
  - » SDSS: FNAL, JHU, NWU
  - » CS: ANL, Chicago, LBL, UI/Chicago, USC/ISI, UW/Madison,

GriPhyN Phase II

- Target NSF/ITR program for 2001, 2002
  - » Follow-on proposal to be developed
- ~ $50M needed to implement Tier 2 infrastructure & services
  - » Hardware component of GriPhyN; high speed network infrastructure connecting sites
- Must be nearly concurrent with GriPhyN I in order to support US/HEP computing needs for LHC data
  - » CMS/ATLAS start in 2005
- LSC participation to be expanded beyond Caltech, UWM, UTB