LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY
- LIGO -
CALIFORNIA INSTITUTE OF TECHNOLOGY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Technical Note **LIGO-T960030 - 00-E** 3/7/96

Data Analysis Scenario and Data Flow Paradigm

M. E. Zucker
J. K. Blackburn

Distribution of this draft:

Data Analysis working group

This is an internal working note
of the LIGO Project.

California Institute of Technology
LIGO Project - MS 51-33
Pasadena CA 91125
Phone (818) 395-2129
Fax (818) 304-9834
E-mail: info@ligo.caltech.edu

Massachusetts Institute of Technology
LIGO Project - MS 20B-145
Cambridge, MA 01239
Phone (617) 253-4824
Fax (617) 253-7014
E-mail: info@ligo.mit.edu

WWW: http://www.ligo.caltech.edu/

1 DATA FLOW PARADIGM

This note describes the path to be followed by detector data after it has been digitized, up to the extraction of physically meaningful signal signatures. It is intended to help determine the functional basis for hardware and software configuration, as well as to guide the division of analysis tasks.

For concreteness we assume the data analysis tasks are structured within a definite framework (Figure 1). This framework was chosen to allow unbiased definition of functions essentially without regard for their resource needs. An additional deliberate assumption, that no analysis tasks need be performed online in real time, will certainly not be true in practice, but further helps decouple functionality from computational constraints. Thus the only task of Figure 1 which needs to be considered "real time" is writing of the Master Tape by the LIGO DAQ system. We expect that after evaluation of algorithmic and resource constraints, a "real" implementation will remap some functions upstream (e.g. online pulse analysis in "real time"), further downstream to lower layers of analysis, or onto special hardware.

In this paradigm we presume all acquired[1] interferometer and PEM signals are recorded to a "master tape" in real time (one per operating interferometer). Each digitized signal is represented as a continuous time series, analogous to a multichannel strip chart recorder; at a given instant of time, a slice through this data stream would give the exact time, the instantaneous values of all the channels, and the currently valid representation of the machine state vector (Figure 2). While the data would really be divided into packets and chunks for manipulation and transport, we simply model the data stream to have arbitrary continuous length (i.e. the physical length of a "tape" is very long compared to demands of its users, though this is not necessarily possible for all users). We also ignore any effects of multiple sample and update rates. Such hardware constraints are expected to play a defining role in translating this conceptual process distribution into a real implementation.

As a further simplification, we have presumed that correlation and comparison of data from the three LIGO interferometers and other worldwide detectors will occur after significant data downselection and processing has already occurred. It should be emphasized that this is only a simplifying assumption, and is not expected to hold for all conceivable analyses. For example, periodic and stochastic background searches may combine data from distant interferometers early in the data reduction process, and diagnostic correlation of raw data from both Hanford instruments and their common PEM system will play a key role in finding and eliminating nongaussian backgrounds.

In this paradigm subsets and processed versions of the data stream are taken from the "master tape" repository and written on "user subset tapes" for specialized analytic and diagnostic functions. Conceptual outlines of some of these functions are presented below. Finally, we discuss potentially desirable features of data format and structure beyond the primitive "strip chart" analogy.

---

[1] Here we refer to the class of continuously sampled signals and continuously updated machine state information. High-bandwidth diagnostic signals, while they may be recorded, would be incorporated sporadically or after condensation to a compressed representation, if at all.

Figure 1: Data flow paradigm

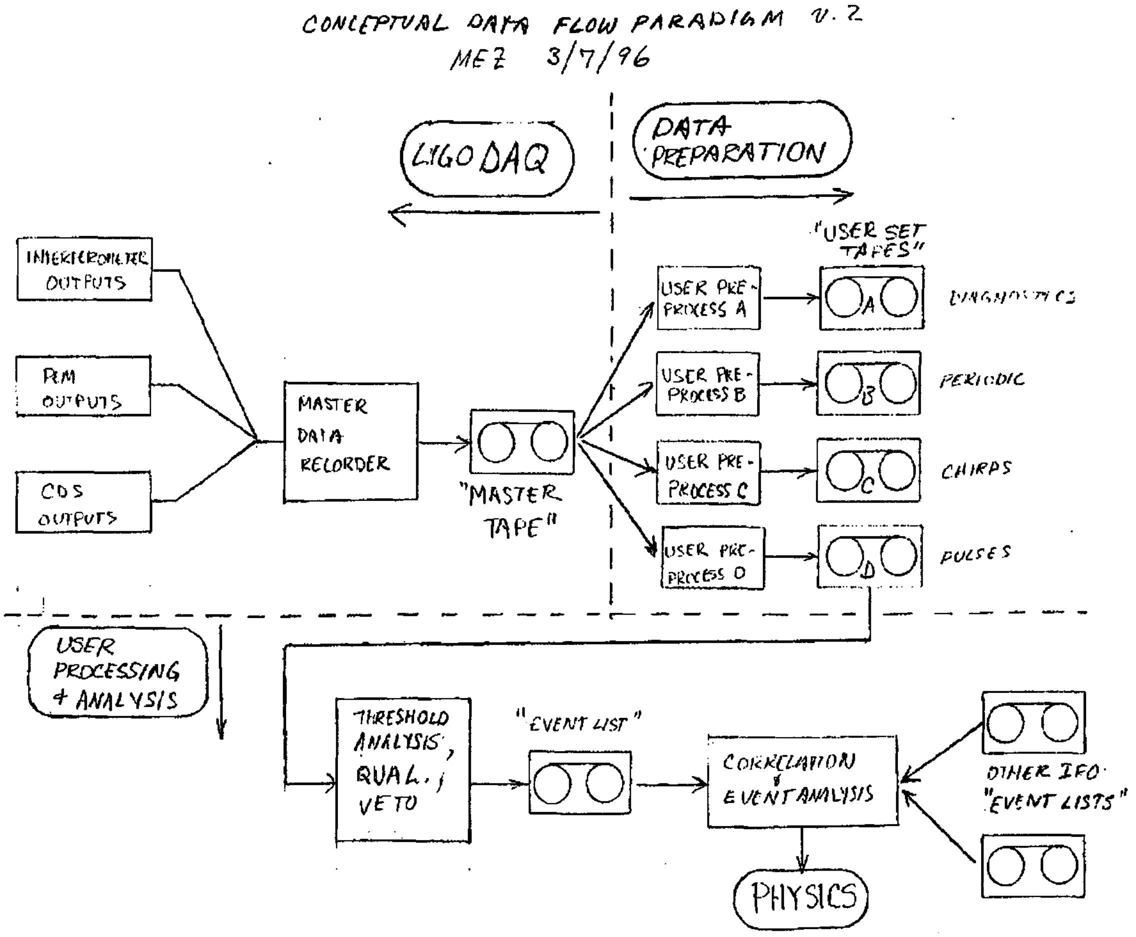
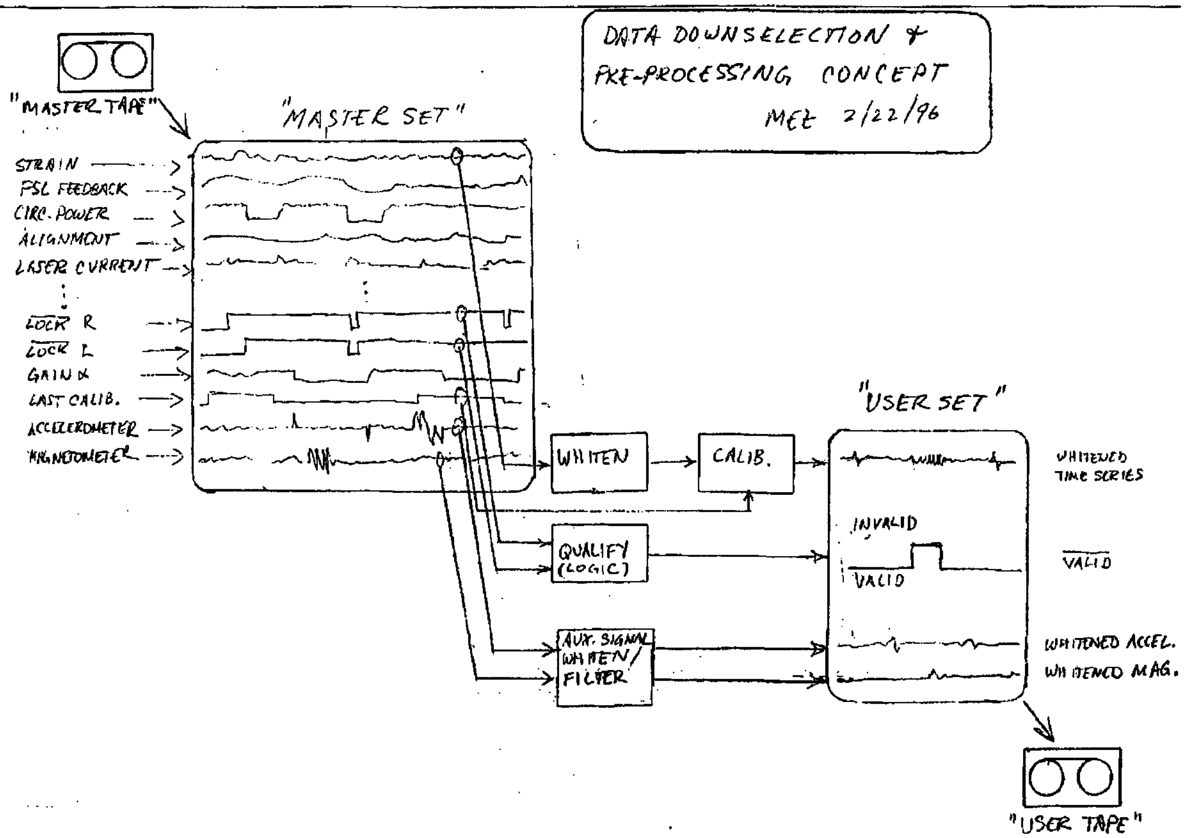


Figure 2: Data processing from “Master” to “User” tape (example)



Reference: “Data Processing, Analysis and Storage for Interferometric Antennas.” B. F. Schutz, in **The Detection of Gravitational Waves**, D. Blair, ed. (Cambridge, 1989).

2 USER DATA PREPARATIONS

Each “user” would typically need access to only a subset of the data stream, and may also benefit from a certain degree of preprocessing. To continue the paradigm, we provide a “user set tape” from the master tape, by implementing some selection and preprocessing functions (Figure 2). Here are some ideas for such “user subsets.”

2.1. Impulsive Event Search

The assumed starting point for most pulse-searching algorithms is a white noise data stream. The searches generally would involve convolving a large number of proposed “signal” templates, derived from various astrophysical or phenomenological signal classes, with the strain signal.

The raw digitized interferometer output will not be white, but strongly colored. Even though there may be significant analog prefiltering before digitization to suit the limited dynamic range of the ADC’s, this filtering is severely limited by practical constraints, and the digitized time series is expected to consume the full dynamic reserve available from its N-bit representation. It thus makes sense to provide a generic "prewhitened" version of the strain signal for all pulse search algorithms. The pulse user also needs the exact filter function used to whiten the data and the complex calibration vector (to transform his search templates from the "physical" strain basis into the whitened basis, and to transform any candidate events back into strain for display). Each of these will be updated periodically as the background noise characteristics and machine operating conditions change.

(Note that the whitening filter impulse response will be quite long, roughly the reciprocal of the frequency bandwidth of its narrowest features; its efficient manipulation may be critical, and its length may constrain the "packetization" of data for transport and storage).

The pulse user will need data for at least two other functions; data qualification and spurious event veto. To qualify stretches of data, a hierarchical "OR" of servo unlocks, servo overranges, mode hops, settling time clocks and other state vector alarm flags will provide an essentially binary representation of which time periods contain valid data. Indeed, it may be advantageous to simply delete unqualified data; but since the total amount of unqualified data is to be minimized by detector design, this should not help much. Also, several different "severities" or hierarchical groupings of data disqualification may be useful, e.g. for continued development of the algorithms and veto/qualification procedures themselves.

After data qualification and matched filtering, the pulse user develops a list of candidate event times and properties. He/she must then compare these events with the possible local influences that might have caused spurious signals. These include "high speed" veto signals, including PEM microphone, magnetometer, powerline monitor, accelerometer, cosmic ray scintillator, and vacuum ion gauge data, plus various fast IFO servo and photodetector signals (e.g. for laser glitches or bursts of servo oscillation). The process of examining these fast data for "exceptional" events in many ways parallels the processing of the strain data; it is thus likely that prewhitening would be advantageous for some of these channels as well.

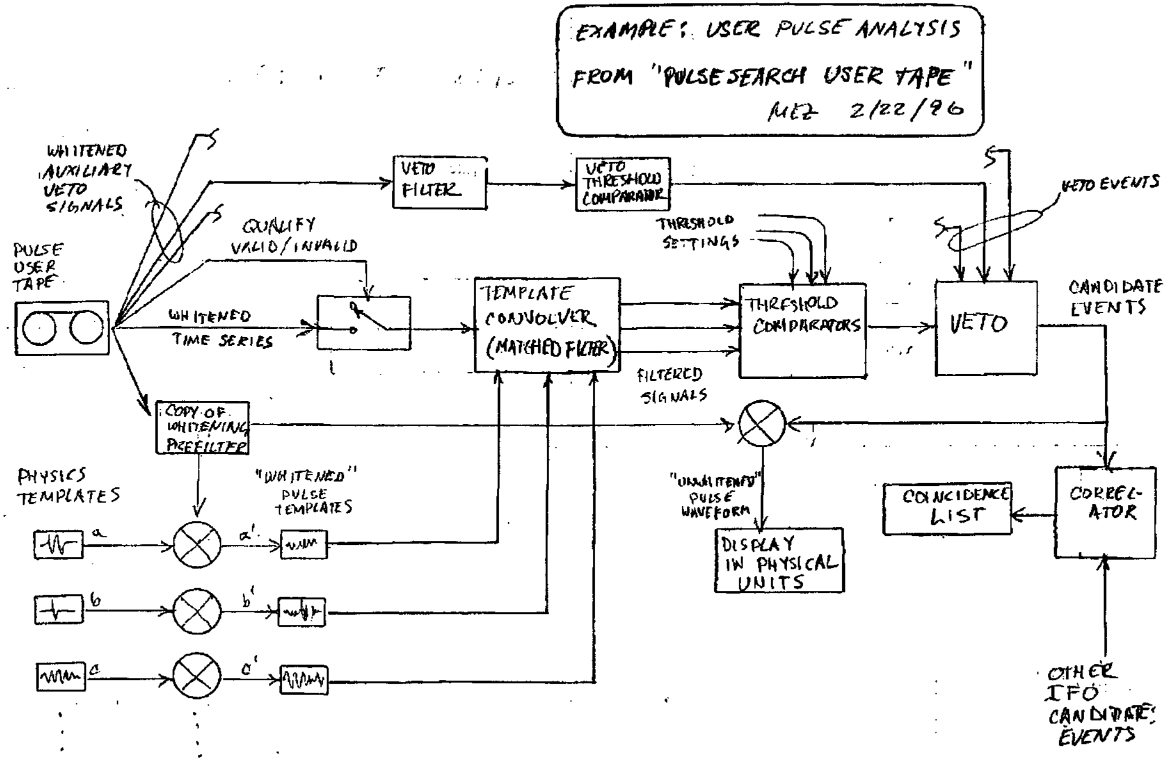
By careful construction of the "data qualification" logic hierarchy and veto processing, the spurious event "singles" rate should eventually be reduced to a level where coincidence processing between interferometers is fully effective at removing any remaining local events. Initially, we will not know the proper veto and qualification criteria, but armed with early event lists we can go back to the full data set and uncover unpredicted correlations (or record other auxiliary veto channels, or fix the machines). We should then be able to iterate and converge on a streamlined pulse detection data preparation that is convenient to use.

It may be useful to further subdivide other subsets according to special needs; for example, searches for lower frequency bursts, higher frequency bursts, bursts with "memory", etc. may benefit from different whitening operations and veto signal subsets.

Finally, we note that pulse analysis is a high priority candidate for "online" processing, for at least two reasons; first, it appears feasible to do it in real time at a physically interesting level, and second, such signals (e.g. from supernovae) would benefit from prompt correlation with optical, radio, X-ray and neutrino astronomical events.

Reference: D. Dewey Ph.D. thesis, 1986

Figure 3: Analysis algorithm example: pulse search



EXAMPLE: USER PULSE ANALYSIS
FROM "PULSESEARCH USER TAPE"
MEZ 2/22/96

2.2. Chirp Signal Search

Binary coalescence chirp searches are a very high LIGO priority and also computationally intensive, so it makes sense to condense efficient reduced data sets specifically for them. These searches will be characterized by the need to explore large, multidimensional parameter spaces with very high template accuracies (required to achieve detectable SNR's because of the phase-coherent nature of the chirp signals). These properties lend themselves to highly parallel search processes.

Like pulse algorithms, chirp algorithms will probably be best off starting with prefiltered ("whitened") data streams. However, it may be advantageous to apply slightly different weighting, e.g. to emphasize the low frequencies where these signals carry the most energy, to reduce the average filtering burden on subsequent template search algorithms.

The detailed waveform match demanded for these complex signals should reduce the reliance on environmental vetoes, but not completely; it will still probably be necessary to catch unusual events in PEM and IFO signals, but the temporal and amplitude discrimination applied to those vetoes will probably be different. For example, pseudoperiodic disturbances with many cycles, like servo oscillation bursts and powerline current transients, could be especially toxic.

The required data qualification logic and calibration information will be similar to that provided for pulse searches, although there may be adjustments necessary due to the extended duration and dominant low-frequency character of chirp signals. For example, because the seismic noise is a more important limitation to the SNR for these signals, and seismicity is intrinsically nonstationary, it may be desirable to adjust the whitening filter more often to milk better SNR during seismic lulls.

Reference: S. Smith Ph.D. thesis, 1987

2.3. Periodic Signal Search

Coherent periodic sources, e.g. nonaxisymmetric rotating neutron stars, are assumed to be relatively weak GW generators (indeed, their traditional classification is slightly tautological since, to remain coherent, they cannot dissipate much power; otherwise they would be "chirps"). Whether it represents an important class of sources or not, the problem of detecting a weak monochromatic signal has been worked extensively.

The obvious strategy, to Fourier analyze the longest data record possible and look for statistically significant peaks, is confounded by two serious limitations: motion of the detector and intermittency in the valid data record. Earth orbital motion produces Doppler frequency modulation of a signal having a fixed frequency in inertial space. Earth rotation produces Doppler frequency modulation as well, and also produces amplitude modulation as the antenna pattern rotates with respect to the source direction.

Unless the periodic search user confines his data records to periods less than tens of minutes (and thus constrains his maximum SNR), coherent accumulation of signal power is impossible without removal of the AM and FM signatures. This can be done by resampling and demodulating the strain data, but the procedure can only work for one sky direction at a time; doing an all-sky search with long data sets (several months) may be computationally prohibitive.

A less effective alternative is incoherent averaging of successive short-record spectra. Schutz et al are developing methods by which neighboring sky direction corrections can be introduced after Fourier transforming a time series which has been corrected for a single direction. If this latter approach fulfills its promise, Doppler- and AM-corrected time series could in principle be prepared for a limited number of skymap "neighborhoods" and deep searches performed in those regions with only post-processing of the reduced data set. Another conceivable aid would be to do shallow searches for weak candidate events, try to fit their Doppler and AM signatures to a specific direction (wherein many would prove unphysical), and then de-Doppler the longer data record for surviving candidates.

The orbital modulation problem has received more attention than another artifact which may pose equally serious limitations. The "live time" of the interferometer is likely to be broken up into a large number of valid data stretches, between which the cavities have lost lock or hopped modes, the servos have exceeded their ranges, etc. The qualified data, and any periodic signals in it, are thus multiplied by a binary semirandom "telegraph wave" window function. This function will modulate any periodic signals in the data and produce spectral leakage, inducing spurious sideband/harmonic signals out of "line" features in the background noise and diluting the spectral power of real signals

In the face of these obstacles, it is hard to predict what kind of data reduction would be most appropriate for periodic-source users. Certainly it would be of some help to provide whitened signals and the corresponding filters used to make them, although the frequency resolution of the filters would have to be constrained by the projected search bandwidths. Precise timing information should also be included.

Veto signals will be extremely important, since periodic disturbances are common (e.g. acoustics from transformers and equipment fans, clock subharmonics from digital equipment, etc.); these may also be distinguished by their lack of Doppler/AM signature, but are likely to be numerous and powerful enough that such discrimination is not adequate.

Catalogs of known interference frequencies, included in the user set distribution, will be very helpful in reducing the amount of frequency space algorithms need to address. For example, exact power line phase/frequency should be tracked over time and reported periodically; typical analyses would reject candidate signals at powerline harmonics. Mechanical resonances of the apparatus, many of which have high intrinsic Q, can also be reported as “a priori vetoes”. Tracking and updating these frequencies vs. temperature or aging may also be useful in cutting down false alarms. Certainly any known transient affecting their excitation amplitudes need to be reported (e.g. a lock acquisition force transient “pinging” all the wires on the test mass suspension; they will take many minutes to settle down to their equilibrium thermal amplitudes).

References: J. Livas Ph.D. thesis, 1987; M. Zucker Ph.D. thesis, 1988

2.4. Stochastic Signal Search

Stochastic background searches will be used to place upper limits on the energy density of stochastic gravitational waves. This background gravitational signal is analogous to the 3K microwave background in the electro-magnetic spectrum. Sources of the stochastic background include the random superposition of gravitational waves from binary star systems, decaying cosmic strings, first-order phase transitions in the early Universe, and parametric amplification of quantum mechanical zero-point fluctuations in the metric tensor during inflation. The predicted strengths of all of these sources is highly uncertain. Placing upper limits on the energy density of the stochastic background or more importantly, the detection of the stochastic background, could introduce a wealth of information to better our understanding of the Universe. LIGO with its detectors in Washington and Louisiana very nearly aligned provides for the optimal orientation of two detectors for stochastic background searches. The proposed VIRGO - GEO orientation minimizes that detector pair’s sensitivity to stochastic background searches amplifying the role that LIGO will play in this type of signal search.

The stochastic background signal will be buried in the noise sources of a single detector and can not be detected unless the signal strength is significantly larger than currently predicted. However, for two detectors with no correlated sources of noise, the only correlated fluctuations in the detector pair will be those due to stochastic gravitational waves. Therefore the detection scenario for the stochastic background involves a properly weighted correlation of the strain outputs from two detectors having uncorrelated noise sources. The proper weighting is given by the optimal filter and is a function of the separation between the two detectors. This is due to the fact that the arms will on average move strongly in phase only for gravitational waves having wavelengths longer than twice the distance between the detector pair. For the Washington - Louisiana separation of 3000 km, this corresponds to stochastic signals of frequency less than approximately 50 Hz.Even though the optimal filter for stochastic signal detection peaks strongly below this frequency, the full spectral range of the interferometers is expected to be used in the search and non-trivial weighting carried out to at least 1kHz.

In order to properly carry out the two detector signal correlation using the optimal filter, the pre-whitened signals from the two detectors will need to be un-whitened to bring the two signals back into the proper phase relationships for the gravitational wave, unless the pre-whitening filters from the two detectors are identical in there phase response over the frequencies of interest. This will require that the full amplitude and phase characteristics of the pre-whitening filters be recorded and available to the stochastic data analysis. Additionally, since the optimal filter is peaked for frequencies below 50 Hz, it is important that the pre-whitening filters do not zero out the strain signal at these lower frequencies. In fact, the dependency on the frequency in the signal to noise ratio goes as one over the frequency to the third power. The signal to noise ratio also increases with time. Searches will most likely utilize months if not years of data. However, this type of search does not require the vast computer power of the periodic search since the strain data does not require doppler shifts and data can be handled in smaller stretches.

Intrasite searches for stochastic signals are not expected to be possible do to the strong correlations in the common noise sources such as power line, gas column density fluctuations in the beam tube, and seismic noise. Even if these correlated noises were recorded, there would be no hope for separating out the contributions from these noises and those from the stochastic background.

References: Allen, private communications
Flanagan, Phys. Rev. D48. (1993) 2389
Christensen, Phys. Rev. D46. (1992) 5250
Christensen Ph.D. thesis, 1989

2.5. Machine Diagnostics & Development

To perform routine diagnostics we would like to have a convenient, physically meaningful set of machine parameters and signals expressed in flexible, readily manipulated formats. The need to uncover unknown correlations and relationships might mean this set also includes all or nearly all the "raw" signals from the master tape as well; as a result, this set may be physically _bigger_ than the original master. However, we also probably want to further subdivide the diagnostic set into subsets directed at specific subsystems, e.g. Suspension, Alignment, or Laser, or specific phenomena, e.g. Length-Alignment Interaction, Seismic Noise Propagation, Acoustics, etc.

The following "processed" signals, in addition to the raw data, come to mind;

- prefiltered strain signal A (good high-frequency resolution, LF suppressed)
- " B (good low-frequency resolution, HF and lines suppressed)
- whitened strain signal (filtered by inverse mean power spectrum, see Pulses above)
- averaged FFT (PSD estimate) for most recent N seconds of data (possibly several resolution

bandwidths & frequency spans)
- logic “or” of “cavity locked” status bits from state vector (and of subsets)
- logic “or” of other veto flags (controls at range limits, etc.)
- current strain calibration vector (updated as necessary)
- theta, phi error signals for each suspended optic, in EU (rads)
- slow cavity feedback signals in EU (meters)
- filtered time derivatives of slow feedback signals
- “user defined” signals

3 ‘REAL’ DATA FORMAT CONSIDERATIONS

Actual data storage formats were deliberately ignored in the above discussion, but some of the procedures and algorithms seem to recommend consideration of file format capabilities like those described by the Cardiff group (e.g. “Work in progress on Data Acquisition and Analysis,” Cardiff internal report by Barnett, Schutz, Shuttleworth and Nicholson, March 1992; section 4.)

Briefly, a standard of this type permits relatively free mixture of different data types by providing “chunks” or “frames” with timestamps, contents dimensions, and format identifiers. In principle these allow readers to scan for the desired type and only invest in reading relevant data types while ignoring unwanted information. It also conforms with the likely sequence of analysis steps; the user will typically apply completely different initial procedures to each type of data, and combinations of data from different sources (e.g. calibrations with time series, vetoes with strain signals) will only occur later after several processing steps have been performed on each. (cf. Figure 3).

Using this model, one could propose to use special-purpose frames designed for such things as:

- “Fast” IFO data (e.g. strain, laser power)
- “Fast” PEM data (e.g. microphone, accelerometer)
- “Slow” IFO data (e.g. suspension & stack signals)
- “Slow” PEM data (e.g. temperature, RGA pressure)
- Qualified data timespan list (e.g. “ignore frame U, record V through frame X, record Y”
- Measured frequency response transfer functions (e.g. strain calibrations)
- Machine state descriptors (e.g. EPICS database records)
- Remote diagnostics & video snapshots
- Operator log entries

Making frames come out to a convenient size might mean, for example, that some number of fast channels and a correspondingly greater number of slow channels would each be grouped to make one second’s worth.

In addition, one would be able to use efficient frame types for intermediate storage, as for the “user subset” tapes discussed above. By virtue of the reduced data rate these can in principle be defined entirely for the convenience of the subsequent analysis, without the real time performance concerns governing the “master” tape format.