LIGO-G000083-00-D

David Strom[a]
Jim Brau[b]
University of Oregon

Report from the Data Set Reduction Working Group

Outline:

1. Review of White Paper Data Model
   (Allen, Finn, Prince, Riles, Weiss)
   www.ligo.caltech.edu/LIGO_web/lsc/whpap1029.pdf

2. Strawman Reduced Data Set

3. Reduction versus Compression

4. An example of a successful data analysis model in HEP
   (Requested by Albert Lazzerini)

(a) strom@bovine.uoregon.edu, (b) jimbrau@faraday.uoregon.edu

White Paper Data Model

Level 0:	Full Data (not archived)	250 Tb/yr
Level 1:	Archived Reduced Data	25 Tb/yr
Level 2:	IFO Strain & Data Quality	2.5 Tb/yr
Level 3:	Whitened GW Strain Data	0.2 Tb/yr

These figures assume 50% duty cycle.

Strawman Proposal for Channels to include in Level 2 data set

signal	number	bytes	rate (Hz)	Channel number	data rate (Bytes/s)
GW strain signal					
LHO 4K	1	4	2k	0	8k
LHO 2k	1	4	2k	10000	8k
LLO 4k	1	4	2k	30000	8k
Laser Power					
LHO 4k	1	2	2k	1	4k
LHO 2k	1	2	2k	10001	4k
LLO 4k	1	2	2k	30001	4k
Control Signals					
various	12	2	2	-	48k
Max, Min Mean,RMS					
(all 1200 channels)	1200	8	1	all	9.6k
PEM					
Power Line Monitor					
LHO	1	2	0.5k	-	1k
LLO	1	2	0.5k	-	1k
Seismometers					
LHO	15	2	0.25k	20000 - 20014	7.5k
LLO	9	2	0.25k	40000 - 40008	4.5k
Accelerometers					
LHO[a]	10	2	2k	20025...	40k
LLO[a]	5	2	2k	40025...	20k
Course Accel. FFTS:					
LHO[b]	99	8	1	20025 - 20123	0.79k
LLO[b]	48	8	1	40025 - 40072	0.38k
Microphones					
LHO[c]	5	2	2k	20124...	20k
LLO[c]	3	2	2k	40124...	12k
Magnetometers					
LHO[d]	1	2	2k	20171	4k
LLO[d]	1	2	2k	40171	4k

- Total uncompressed rate $\sim$ 200 kb/s

- Assuming 50% duty cycle and 50% compression this is $\sim$ 1.6 Tb/yr

- Level 2 data could be exported to LSC member institutes

Notes:
(see last page)

a weighted sum in quadrature for 10 optical elements.

b power spectra in four bins of frequency

c one microphone /building

d weighted quadrature sum

Compression Versus Reduction

Data Compression

- *gzip* type or even simpler compression algorithms bring a factor of 50%

- Wavelett compression could bring up to a factor of 5, but this would involve some loss of information.
  (see work of Sergey Klimenko)

- Others approaches may also help.
  (c.f. Natalia Zotov)

Data Reduction:

Detector Characterization studies will clarify how control and PEM information will actually be used in LIGO analysis. Possible examples:

- Use servo information to predict location of isolation tables
- Extract from line-removal algorithms current size of violin modes
- Use magnetometers and line monitor to estimate motion of masses due to EM effects

Combined information probably much more useful than individual channels.

Level 2 Evolution

- Packages such as JDclient and RDSWriter can select data to be saved in several formats. (Experience suggests that this only works well if you are actually at the site.)

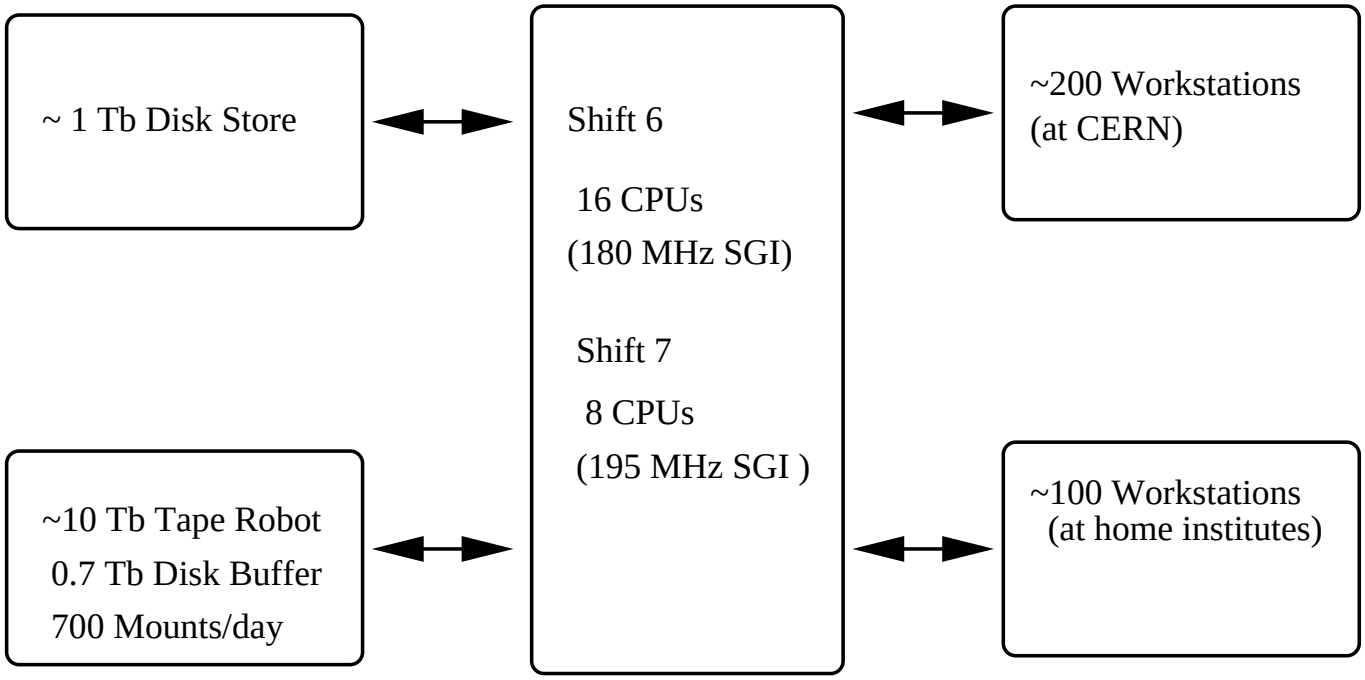
  These packages are the beginning of *Designer* datasets from which the Level 2 data set can evolve.

- Detector Characterization studies should soon tell us what information is needed in the Level 1 and Level 2 data sets.

⇒ Your input is essential for determining how PEM and servo information can be *reduced* and what channels are important.
See **zebu.uoregon.edu/∼strom/reduce/table.html** for details.

An example of a successful data analysis model in HEP

The OPAL Collaboration (one of the 4 LEP experiments) uses a data analysis facility for the first step in almost every physics analysis:



SHIFT = Scalable Heterogeneous xxx Facility Testbed
(http://wwwinfo.cern.ch/pdp/serv/shift.html)

Many thanks to Ann Williamson (University of Indiana) for the statistics shown here.

Steps in a typical physics analysis

1. Debug event selection and private (ntuple) data format.

2. Run over all of the data, apply latest calibrations and corrections (usually analysis specific) Requires 1 to 2 weeks, private data samples are usually less than 10Gb.

3. Develop data analysis on private workstation(s). This step is often CPU intensive.
   Often problems in step 2 are found and step 2 is repeated.

4. Examine more detailed information for individual events (requires short access to shift data).

5. Make final distributions for paper/conference report.

Other interesting facts:

- Approximately half (175) of OPAL members use SHIFT in any given month.

- In any given year OPAL produces 50-100 analysis using the above steps.

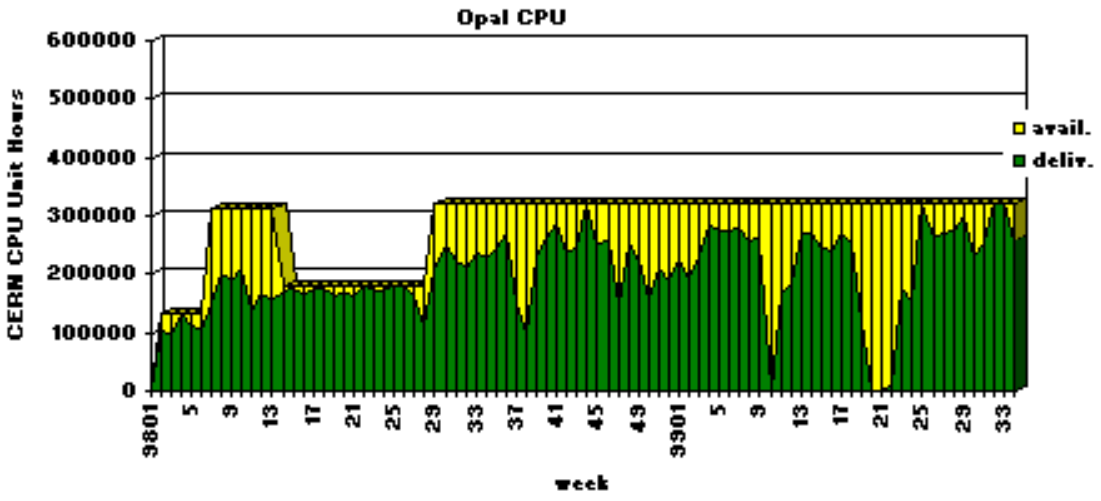
- OPAL-shift was started in 1992 (LIGO has a ten year technology advantage)

Why is OPAL-SHIFT a success?

1. CPU time is more or less divided equally among all users, except in crisis situations. (CPU intensive tasks naturally migrate off central machine).

2. OPAL management has been able to get extra processors in crisis situations, e.g. just before conferences.

3. To my knowledge, no OPAL result has ever missed a conference solely because there was not enough CPU time available on shift.



Implications for the LSC:

- Many problems faced by the LSC will be similar to those encountered in HEP.

- Warning: LIGO problems are inherently more CPU intensive than those HEP.

  $\Rightarrow$ Data analysis proposals should allow these conflicts to be resolved.