LASER INTERFEROMETER GRAVITATIONAL WAVE OBSERVATORY

- LIGO -

CALIFORNIA INSTITUTE OF TECHNOLOGY

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

| **Document Type** | **LIGO-T040164-01-Z** | 2004/11/18 |
|---|---|---|

# SFT Data Format Version 2 Specification

The Continuous Waves Search Group

Generated from CVS

$Source$ : $/usr/local/cvs/lscsoft/sftlib/T040164.tex,v$,

$Revision$ : 1.3

*Distribution of this draft:*

LIGO Scientific Collaboration


**California Institute of Technology**
**LIGO Project - MS 51-33**
**Pasadena CA 91125**
Phone (626) 395-2129
Fax (626) 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/



Processed with LaTeX on 2004/11/18

# 1 Changelog

| | |
|---|---|
| November 18, 2004 | All times in SFT filenames should be in seconds; simplify SFTtype description. |
| September 22, 2004 | Added new section: SFT File Naming Convention |
| September 17, 2004 | After discussion with the pulsar search group, added detector type to SFT header. |
| September 14, 2004 | After discussions with the pulsar search group, updated spec to require that all SFT blocks share some common header info. |
| September 9, 2004 | Release 2.3 of reference library. Spec updated as per comments in SCCB_2004_09_09 |
| August 17, 2004 | First public release of version 2.2 of reference library. |

# 2 SFT Data Format Version 2 Specification

An SFT is stored in a file. See below for file name convensions. The file is composed of concurrent SFT BLOCKS. Each SFT BLOCK is organized as follows, in Table 1.

| | |
|---|---|
| HEADER | 48 bytes |
| ASCII COMMENT | 8*n bytes, where n is a non-negative integer |
| DATA | 8*N bytes, where N is a positive integer |

Table 1: Summary of SFT BLOCK structure.

The total length of the SFT BLOCK is 48+8*n+8*N bytes. The SFT BLOCK may be written in either big-endian or little-endian ordering. All floats and doubles follow the IEEE-754 floating-point conventions.

The HEADER contains 48 bytes as follows, in Table 2.

| | | |
|---|---|---|
| 8 bytes | REAL8 | version |
| 4 bytes | INT4 | gps_sec; |
| 4 bytes | INT4 | gps_nsec; |
| 8 bytes | REAL8 | tbase; |
| 4 bytes | INT4 | first_frequency_index; |
| 4 bytes | INT4 | nsamples; |
| 8 bytes | UNSIGNED INT8 | crc64; |
| 2 bytes | CHAR | detector[2]; |
| 2 bytes | CHAR | padding[2]; |
| 4 bytes | INT4 | comment_length; |

Table 2: Summary of HEADER structure.

The SFT blocks in a given SFT file are required to:

1. come from the same instrument, and have:

2. identical version numbers
3. monotonically increasing GPS times
4. identical values of tbase
5. identical values of first_frequency_index
6. identical values of nsamples

NOTE: SFT blocks in a given SFT file are in general NOT contiguous. In other words the GPS start time of a given block may or may not equal the GPS start time of the previous block plus the time baseline.

Note that the HEADER corresponds to the C structure below

```
struct SFTtag {
                    REAL8     version:
                    INT4      gps_sec;
                    INT4      gps_nsec;
                    REAL8     tbase;
                    INT4      first_frequency_index;
                    INT4      nsamples;
                    UINT8     crc64;
                    CHAR      detector[2];
                    CHAR      padding[2];
                    INT4      comment_length;
} SFTheader;
```

when the structure is packed, i.e. no zero padding between fields is allowed. Note that several of these quantities that might be taken as unsigned are in fact signed. This makes it easier and less error-prone for user applications and code to compute differences between these quantities.

The structure of the ASCII COMMENT is comment_length==8*n arbitrary ASCII bytes, where n is a non-negative integer. The following rules apply to NULL bytes appearing in ASCII COMMENT, if n is non-zero:

1. There must be at least one NULL byte in the ASCII COMMENT
2. If a NULL byte appears in the ASCII COMMENT, all the following bytes in the ASCII COMMENT must also be NULL bytes.

The reason for these two rules is so that if the ASCII comment has nonzero length then it may always be treated as a C null-terminated string, with no information 'hidden' after the null byte. If the SFT comes from interferometer data, then the full channel name used will be contained in the comment block. If a window function is used (see below), then the window name (along with parameters if the name is not sufficient) of the window function will also be contained in the comment block.

The DATA region consists of N COMPLEX8 quantities. Each COMPLEX8 is made of a 4-byte IEEE-754 float real part, followed by a 4-byte IEEE-754 float imaginary part. The packing

and normalization of this data comply with the LSC specifications for frequency-domain data. The current version of this specification may be found at
`http://www.ligo.caltech.edu/docs/T/T010095-00.pdf`.

- version: shall be 2.0:
  Note that SFTs produced before this specification will have this field set to 1.0. Note that future versions of this specification will have version=3.0, 4.0, etc. This field will always be an integer value that can be exactly represented as an IEEE754 double. If this field is not an exact integer in the range 1 to 1000000, then software reading this data should assume that it is byte-swapped and take appropriate measures to reverse the byte ordering. If byte swapping does NOT cause the version number to be an exact integer between 1 and 1000000, then the SFT does not comply with these specifications.

- gps_sec:
  Integer part of the GPS time in seconds of the first sample used to make this SFT.

- gps_nsec:
  GPS nanoseconds of the first sample used to make this SFT. This must lie in the range from 0 to $10^9 - 1$ inclusive.

- tbase:
  The time length in seconds of the data set used to make this SFT. This must be greater than zero.
  **Note**: if the sample interval is dt, and the number of time-domain samples is S, then tbase=S dt. Note that if the data is produced with heterodyning, tbase still refers to the total time length of the data set. Note that the frequency spacing in between samples (df, as defined in T010095-00) is 1.0/tbase. If df can not be exactly represented as an IEEE-754 double, then the closest IEEE-754 double to 1.0/tbase will be the closest IEEE-754 double to df.

- first_frequency_index:
  This is the subscript of the first complex FFT value that appears in DATA. It's allowed range is 0 to (Nyquist_Frequency * tbase)/2 = S/2 inclusive. Note: if S is odd, then in this document S/2 shall mean the integer part of S/2.

- nsamples:
  The number of complex samples in DATA. nsamples=N. Its allowed range is 1 to S/2+1 inclusive

- crc64:
  The 64-bit CRC checksum of the 48+8*n+8*N bytes that make up the SFT, with the 8 bytes labeled crc64 set to zero. The CRC checksum will be evaluated using the polynomial D800000000000000 (base-16) = 1101100000000000000000000000000000000000000000000000000000000000 (base-2). The primitive polynomial is $x^{64} + x^4 + x^3 + x + 1$. The CRC will be initialized to all ones ($\sim$0ULL).

- detector:
  two characters of the form 'Xn' characterizing the detector, following the naming convention for the channel-name prefix in the Frame-format, cf. LIGO-T970130-F-E. X is a single capital letter describing the site. n is the detector number. Currently allowed names are:

| | |
|---|---|
| "A1", | ALLEGRO |
| "B1", | NIOBE |
| "E1", | EXPLORER |
| "G1", | GEO_600 |
| "H1", | LHO_4k |
| "H2", | LHO_2k |
| "K1", | ACIGA |
| "L1", | LLO_4k |
| "N1", | Nautilus |
| "O1", | AURIGA |
| "P1", | CIT_40 |
| "T1", | TAMA_300 |
| "V1", | Virgo_CITF |
| "V2", | Virgo (3km) |

- padding:
  These two bytes will be set to zero. They are here so that all multi-byte quantities are byte-aligned with respect to the header. This may permit certain efficiencies and library usages on certain platforms/architectures.

- comment_length:
  The number of bytes that appear in ASCII COMMENT. comment_length=8*n with n a non-negative integer.. Note that if comment_length==0 then the SFT contains no comment.

The complex quantities contained in DATA REGION are defined by the following equations, complying with `http://www.ligo.caltech.edu/docs/T/T010095-00.pdf`

The data set (with the native fundamental sample interval dt) is denoted by $x_i$ for i=0, ..., S-1. The $x_i$ are all real. Let

$$n_k = \sum_{j=0}^{S-1} x_j \exp(-2\pi i j k/S)$$

be the values of the DFT with LSC sign conventions. The values in DATA REGION are

$$data_k = dt * n_k$$

for k=first_frequency_index to k=first_frequency_index+nsamples-1. [Note: the interesting range of k is from 0 to S/2 inclusive.]

The allowed range of first_frequency_index is from 0 to S/2 inclusive.

The allowed range of nsamples=N is 1 to S/2+1-first_frequency_index inclusive.

[Here we assume that the window function is rectangular, eg each sample is weighted by a window function whose value is 1. If the data IS windowed then the normalization conventions of T010095-00.pdf still apply.]

Note that if a data stream is band-limited (for instance by filtering) and then decimated or downsampled, the values stored in DATA REGION for a given set of frequency bins will be unchanged compared to those computed with the original data set. This is true even though the sample interval dt' of the new downsampled data set is larger than the original native sample time. In fact, except for the DC (k=0) and Nyquist (k=S/2+1) frequency bins, the power spectral density may be written as:

$$\mathrm{psd}_k = (2/\mathrm{tbase})|\mathrm{data}_k|^2$$

(except for k=0 or Nyquist).

# 3 Examples

## 3.1 EXAMPLE 1

Consider the case where the fundamental time-domain data set consists of 16 samples, taken at a sample rate of 16 Hz. All 16 samples are $\mathrm{x}_0 = ... = \mathrm{x}_{15} = 1$ which gives $\mathrm{n}_0 = 16, \mathrm{n}_1 = ... = \mathrm{n}_8 = 0$. Since dt=1/16, we find

$\mathrm{data}_0$ = 1 + 0i
$\mathrm{data}_1$ = 0
$\mathrm{data}_2$ = 0
$\mathrm{data}_3$ = 0
$\mathrm{data}_4$ = 0
$\mathrm{data}_5$ = 0
$\mathrm{data}_6$ = 0
$\mathrm{data}_7$ = 0
$\mathrm{data}_8$ = 0

If we store only nsamples=5 frequency bins in the SFT, then DATA REGION will contain the 40 bytes corresponding to identical values for $\mathrm{data}_k$:

$\mathrm{data}_0$ = 1 + 0i
$\mathrm{data}_1$ = 0
$\mathrm{data}_2$ = 0
$\mathrm{data}_3$ = 0
$\mathrm{data}_4$ = 0

These values could be obtained by considering a subset of the original SFT. Alternatively they could be obtained by low-pass filtering the original time series, and downsampling it, and using the previous definitions. For example if the downsampled time series had 8 samples $x_0$=...=$x_7$=1 with a sample time of dt=1/8, then $n_0$=8, and $n_1$=...=$n_4$=0. This gives the same values as above.

## 3.2 EXAMPLE 2

Consider a sinusoid function at 2 Hz, x(t)= 1 *cos (2 pi 2 t). Using again 16 samples taken at a sample rate of 16 Hz,

$x_{00}$ = 1.000000
$x_{01}$ = 0.707107
$x_{02}$ = 0.000000
$x_{03}$ = -0.707107
$x_{04}$ = -1.000000
$x_{05}$ = -0.707107
$x_{06}$ = -0.000000
$x_{07}$ = 0.707107
$x_{08}$ = 1.000000
$x_{09}$ = 0.707107
$x_{10}$ = 0.000000
$x_{11}$ = -0.707107
$x_{12}$ = -1.000000
$x_{13}$ = -0.707107
$x_{14}$ = -0.000000
$x_{15}$ = 0.707107

giving

$$n_0 = n_1 = 0 + 0\mathrm{i}$$
$$n_2 = 8 + 0\mathrm{i}$$
$$n_3 = ... = n_8 = 0 + 0\mathrm{i}$$

and

$$\mathrm{data}_0 = \mathrm{data}_1 = 0 + 0\mathrm{i}$$
$$\mathrm{data}_2 = 0.5 + 0i$$
$$\mathrm{data}_3 = ... = \mathrm{data}_8 = 0 + 0i$$

If we down-sample the original data stream by a factor of two we get:

$x_{00}$ = 1.000000
$x_{01}$ = 0.000000
$x_{02}$ = -1.000000
$x_{03}$ = -0.000000
$x_{04}$ = 1.000000
$x_{05}$ = 0.000000
$x_{06}$ = -1.000000
$x_{07}$ = -0.000000

giving

$$\mathrm{n}_0 = \mathrm{n}_1 = 0 + 0\mathrm{i}$$
$$\mathrm{n}_2 = 4 + 0\mathrm{i}$$
$$\mathrm{n}_3 = ... = \mathrm{n}_4 = 0 + 0\mathrm{i}$$

and

$$\mathrm{data}_0 = \mathrm{data}_1 = 0 + 0\mathrm{i}$$
$$\mathrm{data}_2 = 0.5 + 0\mathrm{i}$$
$$\mathrm{data}_3 = ... = \mathrm{data}_4 = 0 + 0\mathrm{i}$$

# 4 SFT File Naming Convention

1. SFT file names are to follow the conventions of LIGO technical document LIGO-T010150-00-E, "Naming Convention for Frame Files which are to be Processed by LDAS," for class 2 frames, except with an extension of .sft rather than .gwf:

   **S-D-G-T**.sft,

   where

   - **S** is the source of the data = an uppercase single letter designation of the site, e.g., G (GEO), H (Hanford), L (Livingston), T (TAMA), or V (Virgo).
   - **D** is a description = any string consisting of alphanumeric characters plus underscore (_), plus (+), and number (#). In particular, the characters dot (.), dash (-), and space are prohibited, as is any description consisting of a single uppercase letter, which is reserved for use by class 1 raw frames.
   - **G** is the GPS time at the beginning of the first SFT in the file (an integer number of seconds). This is either a 9-digit or 10-digit number. (If the beginning of the data is not aligned with an exact GPS second, then the filename should contain the exact GPS second just before the beginning of the data.)
   - **T** is the total time interval covered by the file, in seconds. If only 1 SFT is in the file, then T is tbase in the header. For multiple SFTs, if the data is aligned with exact GPS seconds, then T is simply the number of seconds between the beginning of the first SFT and the end of the last SFT. If the data is not aligned with exact GPS seconds, then T should be calculated from the exact GPS second just before the start of the first SFT to the exact GPS second just after the end of the last SFT. Data gaps (i.e., non-contiguous SFTs within a file) are permitted, though the SFTs in the file must be time ordered.

2. Note that even though SFTs do exist outside the LDAS diskcache, adopting the class 2 frame naming convention (except for the extension) ensures that all SFTs can be indexed by LDAS and other tools adopting the LDAS conventions. Note that LDAS v1.2.0 allows a list of extensions as a diskcachAPI resource variable. Thus LDAS v1.2.0 and higher can automatically index SFTs with names as specified here.

3. SFT file names will follow these additional rules for the description field D:

   (a) The description field, D, for SFTs will be an underscore "_" delimited alphanumeric string with these subfields:

   **D = numSFTs_IFO_SFTtype[_Misc]**, where

   - **numSFTs** is the number of SFTs in the file.
   - **IFO** is a two character abbreviation of the interferometer data used to generated the SFT, e.g., G1, H1, H2, L1, T1, or V1. This field must always begin with an uppercase letter.
   - **SFTtype** is the type of SFT(s) in the file, and must be a concatenation of tbase in seconds and SFT, e.g., 1800SFT for 30 minute SFTs, 60SFT for 60 second SFTs. Note that this information is redundant but required since it ensures that SFTs of a given type are indexed uniquely (many scripts and programs use the D field as an index; see also comment b below).
   - **Misc** is an optional field that contains any other pertinent information about the SFTs, e.g., the run name, the input channel name, the calibration version, the data quality version, the frequency band, etc.... This field can be anything the group or individual that generates SFTs wants to include in the name, and should be used to distinguish a set of SFTs as unique from other sets for the same site, IFO, SFTtype, and GPS times. Possible examples of the Misc subfield are:
     i. S1: The SFT is from the S1 Science run.
     ii. S2v1Cal: The SFT is from S2 data using v1 calibration.
     iii. S3ASQv3Calv5DQ: The SFTs was generated from the LSC-AS_Q channel using v3 of the calibration and v5 of the data quality segments. (Note that LIGO channel names contain prohibited characters; thus if channels names are included in the Misc subfield a nickname or abbreviation must be used.)
     iv. S4hot: The SFT is from the S4 run using calibrated h(t).

   (b) Even though some of the information required in the description field, D, is redundant, many scripts and programs (such as LSCdataFind and the LDAS diskcacheAPI) rely on this field, plus the GPS interval and source letter, to find files. The required subfields of D will ensure that SFTs files are uniquely identified by these scripts and programs.

   (c) Example SFT files names

   - A 30 minute H2 SFT:
     H-1_H2_1800SFT-735627918-1800.sft
   - A 30 minute S2 H1 SFT:
     H-1_H1_1800SFT_S2-733467931-1800.sft
   - A file with 1887 30 minute H1 SFTs for the 352 to 353 Hz frequency band, (with gaps in time):
     H-1887_H1_1800SFT_352to353Hzband-733467931-4622400.sft
   - A 30 minute S3 GEO SFTs produced from h(t):
     G-1_G1_1800SFT_S3hot-732465218-1800.sft

- A 60 second S2 L1 SFT from the L1:LSC-AS_Q channel for v2 of the calibration and version 6 of the data quality segments:
  L-1_L1_60SFT_S2ASQv2Calv6DQ-788901256-60.sft

(d) Thus note that Misc is left as an arbitrary subfield of the D field in this specification to allow flexibility. It is up to the groups using SFTs to agree on the information to include in this subfield for the generation of SFTs for each run.