‘Steim’ Compression

Joseph M. Steim, Quanterra, Inc.

1.0 History

The ‘Steim’ compression algorithms for seismic data grew from work originally done at Harvard in the early 1980’s to construct a digital ‘very-broad-band’ seismograph capable of recording most teleseismic signals in a single high-dynamic-range data stream . The first VBB system included a new type of A/D converter, exhibiting more than 24 binary-bits dynamic range.

A new recording format was necessary to represent this type of output without losing information present in the broad-band signal, and without using prohibitively large amounts of storage. Without compression, 32-bit binary data for a 3-component seismic station recording continuously at 20 samples/sec exceeds 20Mb. Additional data channels, such as detected higher sample rate data and continuous lower sample rate data could push the volume to near 30Mb/day. A network of 100 stations could therefore produce 3Gb/day. This remains unwieldy even today, and was not manageable in 1984.

The compression technique had to:

- represent at least 28-binary-bit dyanamic range.
- be general applicable, and therefore lossless.
- be very fast, not requiring intensive floating-point operations, of which many field data loggers, even today, are not capable.
- be easy to understand and code (a page or so), and therefore be more likely to gain acceptance.
- have few states, and be easily testable.
- reduce ‘significantly’ the average global network volume for continuously-recorded data.
- reduce data volume sufficiently to allow effective telemetry of the compressed data, at least at a rate exceeding ‘real time’.

The Level 1 method, which simply codes first differences as groups of 8, 16, or 32-bits, was first coded in 1984. A ‘bit map’ preceding each group of fifteen 32-bit words gives the decoder the information necessary to reconstruct the original samples. The detailed description of the Level 1 and 2 coding appears in the *Standard for the Exchange of Earthquake Data, Reference Manual, Version 2.3*, Federation of Digital Seismic Networks.

As an aside, it is interesting to note that had the original test data been taken, say, at the Albuquerque Seismological Lab (ASL), rather than Harvard’s Oak Ridge Observatory (HRV), a rather different technique may have emerged. As shown in Peterson’s recent

report, *Observations and Modeling of Seismic Background Noise*, **USGS OFR 93-322**, *1993*, on levels at global network sites, HRV is in the top 1/3 of average short-period noise levels (which dominate the compressibility of the broad-band data) for the global sites he examined, while ASL is by far the quietest SP site in the network. Broad-band data at ASL are a playground for compression algorithms, yeilding additional returns from many technques, while HRV data are fairly resistant to compression.

1.1 Level 1

The technique that emerged, and has gained wide acceptance, is the 'Level 1' compression algorithm that uses the property of typical seismic data that a simple first-difference linear predictor works well for lengths on the order of a few samples. The grouping of a varying number of first-differenced samples into groups of 32 bits simultaneously encodes the data with low overhead, and implicitly uses the advantageous few-sample correlation length. Because each group of 32 bits is free to represent differences between consecutive samples from as small as 8 bits up to 32 bits, arbitrary differences up to 31-bit absolute value (actually beyond, because nobody bothers to detect 32-bit arithmetic overflow) can be represented by the technique. While extremely easy to code, and very fast, the maximum compression ratio of about 3.7-3.8 relative to 32 bits can be improved substantially.

1.2 Level 2

In 1990, the method was extended with Level 2, which is very similar to Level 1, with the addition of several other possible representations of binary differences. Level 2 can represent groups of differences of 4,5,6,8,10,15, and 30 bits. The additional flexibility in representation improves compression efficiency up to 80% without a major increase in CPU time. Many field stations have been capable of this since 1991, but few have enabled the feature because of lag time in central collection facilities for handling the format. The IRIS DMC now has the capability, soo to be followed by the ASL DCC.

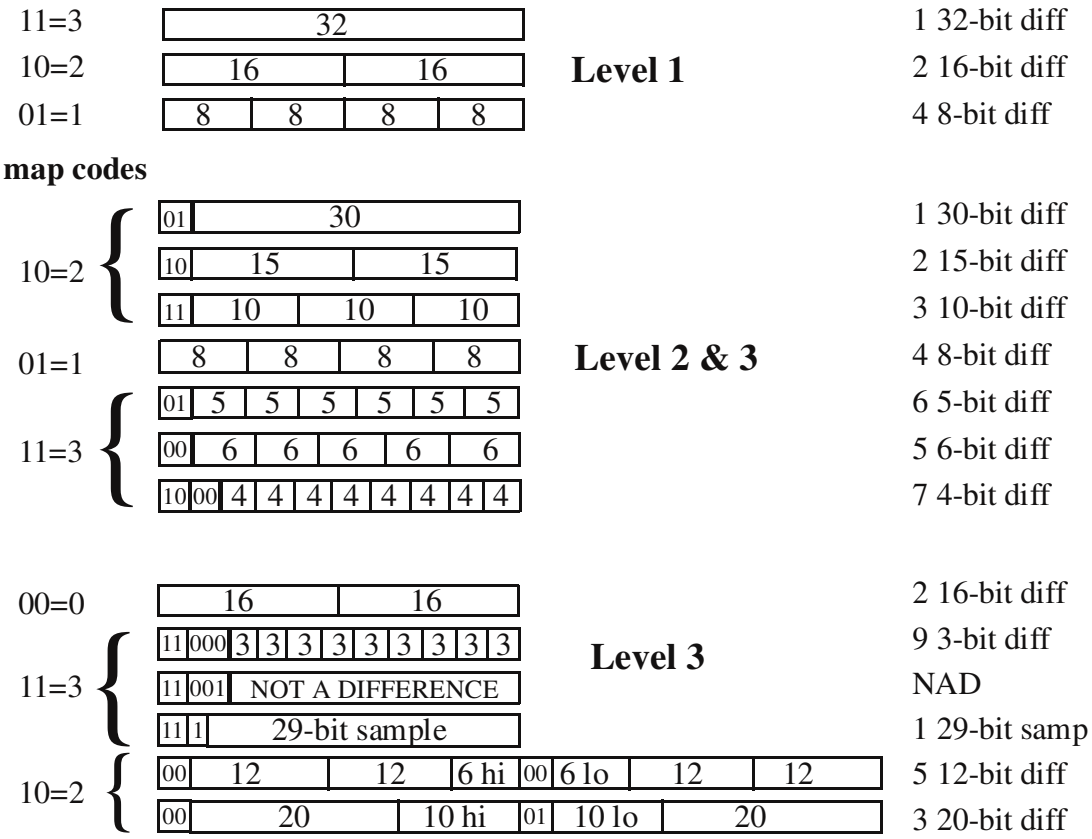
1.3 Level 3

1.3.1 Method

Also in 1990, a third, increased level of compression was proposed that essentially further expanded the Level 1 and 2 techniques by including:

- replacement of common compression bit-map patterns with one-byte codes, liberating up to about 5% additional space. This is called "flag squeezing".
- spanning two 32-bit groups to allow a representation of 5 consecutive 12-bit differences and 3 20-bit differences; representing up to 9 consecutive 3-bit differences; and using the previously-reserved bit map code 0 to represent 2 16-bit differences. The latter required definition of a Not-A-Difference 32-bit value.
- Adaptively using either 1st or 2nd differences (by frame if desired), coding the difference in the sign bit of bit-map word 0, which maps to itself, and was therefore not used.

Level 3 was not implemented in field stations of the global network at that time. The illustrative C program “steim123” embodies the technique. The coding scheme is described here, along with a comparison of results of Level 1 and 2 compression and Stearns’s LPBLC compression on some of the same data (and additional data) used by Peterson and Hutt, *Lossless Compression of Seismic Data, ASL, 1992*.



In level 3, common 32-bit bit-map words are also replaced with 8 bit codes (“flag squeezing”), freeing an additional 24 bits per 512-bit frame. The common bit-map words and their replacements are shown below. The lower nibble (4 bits) of the replacement byte represents one of 6 possible data types are represented in the 24 bits freed by using the replacement byte instead of the original 32-bit map. The 29-bit (two’s complement) samples are used to represent integration constants, and are not differences, nor decompressed.:

TABLE 1. bit-map replacement bytes

bit map 32-bit word	8-bit replacement, upper nibble
0x15555555	0x50
0x2AAAAAAA	0x60
0x3FFFFFFF	0x70

1.3.2 Data

Peterson, 1993, compiled background noise estimates for a wide cross-section of the available digital stations. The typical background levels of VBB data (which control the compression effectiveness) range from about 2 dB above the Low-Noise Model at 1Hz (ANMO) to 46dB above, with a mean and median 20-22 dB above the LNM. For minimizing network data volume, the compression test data should be weighted as the actual network. Other objectives, such as maximum compression on very well-behaved (or poorly-behaved) data, or events, can be looked at individually.

Highly simplified, the network background noise levels divide the network into, say, fourths, or thirds, from which data from representative stations within each group, appropriately weighted, can give an estimate of the typical whole-network behavior. The following summarizes the representative data, and the source. Some of the files are provided in the "EARTH1" data files used by Stearns, from whom the description is taken, representing both exceptionally quiet background at Albuquerque, and high background at RAR; some files use the data in Peterson and Hutt, 1992, to compare Level 3; and some data from PFO and MAJO are new, illustrating the average, to slightly better- than-netowrk-average noise levels. The stations (not each file) are ranked by the number of stations that are quieter, from a total of 74.

DESCRIPTION OF EARTH 1 FILES

##	File Name	Stat	Yr	Rate	Samples	rank	Data	Source
1	ANMBHZ89.ASC	ANMO	89	20	67000	0	Loma Prieta earthquake	Stearns
2	ANMEHZ92.ASC	ANMO	92	100	14000	0	High freq. event	Stearns
3	ANMBHZ92.ASC	ANMO	92	20	72000	0	Broadband background	Stearns
4	RARBHZ92.ASC	RAR	92	20	72000	65	Broadband high bgnd.	Stearns
5	ANMLHZ92.ASC	ANMO	92	1	86000	0	Long-period low bgnd.	Stearns
6	RARLHZ92.ASC	RAR	92	1	86000	65	Long-period high bgnd.	Stearns
7	KIPEHZ13.ASC	KIP	92	100	62000	63	High freq. quiet bgnd.	Stearns
8	KIPEHZ20.ASC	KIP	92	100	61000	63	High freq. noisy bgnd.	Stearns

In addition, the following files were added to represent other groups that comprise the network:

9	PFOBHZ94.DAT	PFO	94	20	71995	26	Quiet background	Dial-Up
10	ANMBHZ87.DAT	ANMO	87	20	16384	0	Quiet Background	P&H, 1992
11	HRVLPZS2.DAT	HRV	92	1	16771	56	Reduced-Gain LP backgr.	P&H, 1992
12	HRVLPZ92.DAT	HRV	92	1	20480	56	Very large event	P&H, 1992
13	TNSHHZ92.DAT	TNS	92	80	41200	?	Teleseism, oversampled	P&H, 1992
14	HRVBHZ87.DAT	HRV	87	20	16384	56	Typical background	P&H, 1992
15	MAJBHZ94.DAT	MAJO	94	20	72640	34	Quiet background	ASL
16	MAJLHZ94.DAT	MAJO	92	1	86628	34	Quiet background	ASL

1.3.3 Comparison of Levels 1-3 with the LPBLC of Stearns.

The following table summarizes the compression ratios for each method. The ratio for the linear-predictor (LP) is based on the compressed file size vs. an equal-number of 32-bit uncompressed samples. For each compression level(L1, L2, L3), the two number are the-compression ratio/execution time(sec) for 50000 samples on a 486/66.

	LP	L1	L2	L3
ANMBHZ89.ASC	6.23	2.20/0.85	2.71/0.86	4.85/0.71
ANMBHZ92.ASC	7.08	3.74/0.68	5.69/0.50	6.89/0.68
RARBHZ92.ASC	4.50	2.34/0.98	3.10/0.76	4.03/0.76
ANMLHZ92.ASC	2.76	1.87/1.26	2.08/1.22	2.53/1.13
RARLHZ92.ASC	2.03	1.79/1.29	1.32/1.69	1.82/1.48
KIPEHZ13.ASC	6.83	3.74/0.57	6.49/0.48	7.13/0.60
KIPEHZ20.ASC	5.57	3.74/0.65	4.25/0.69	5.21/0.77
ANMEHZ92.ASC	4.80	3.27/0.82	4.54/0.56	4.81/0.82

supplementary files.

PFOBHZ94			6.44
ANMBHZ87 DAT	7.34		6.97
HRVLPZS2 DAT	3.09		2.89
HRVLPZ92 DAT	2.02		1.79
TNSHHZ92 DAT	6.47*		6.04
HRVBHZ87 DAT	5.39		4.81
MAJBHZ94 DAT	5.88		5.78
MAJLHZ94 DAT			2.36

The following table summarizes execution times for the “steim123” C compressor on a number of computers, the 68030 and 68000 being of particular interest, since they power all of the global-network processors, and many field-portable instruments as well. The execution times were all normalized to 50000 samples to allow comparison with the Peterson and Hutt results. All of these times were measured with the program’s internal profiling, which measures separately time spent purely in decompression. The first number is the actual decompression-only time/the second number is the program’s total run time, of which the majority is typically reading and writing text files, which is notoriously slow. All times are in seconds. The 1062-sec execution time for LPBLC coding is calculated, because no fortran compiler is available to port the LPBLC program to that computer. Surprisingly, the Borland Pascal version on the 486 actually logged the quickest overall run times, and compared favorably to the SPARC2, which is a faster computer.

MAJBHZ94.DAT (compression-only times normalized to 50K samples)

level and	ratio	L1=3.74	L2=4.93	L3=5.78	LPBLC=5.88
SPARC2	C	0.24	0.28	0.34/3.3	31.6/51
486/66	QuickC	0.68/3.9		0.70/3.6	
486/66	Pascal	0.38/3.0		0.59/3.0	
68030/20	C	2.6/26		3.18	
386SX/16	C			5.60/34.0	
68000/12	C			11.43	(1062)

Level 3 produces compression ratios comparable to the LPBLC coding, at about 1-3% of the LPBLC execution time. In its present form, the linear-predictive filtration and coding is much too time intensive for real-time acquisition in simple equipment.

1.4 Level 2/3-like coding combined with Stearns's linear predictor

C.R. Hutt has suggested a combination of the efficient coding of the Level 1/2/3 techniques to represent rather than either fixed first or second-differences, the prediction error residuals, whose absolute value will be small if the predictor works well. This is essentially the general case of the level-3 adaptive differencing. Combined with an all-integer prediction filter, and occasional adaptation to reduce CPU time, this method may produce further gains that are usable on limited embedded hardware.

1.5 Other factors in bitwise prediction error encoding

In addition to the characteristics of the noise spectrum, the gain of the seismometer/digitizer combination critically affects compression efficiency for algorithms like Level 1,2... that depend on the predictability of the data, and try to represent prediction residuals in the smallest space possible. High gain systems (not necessarily high-resolution or high dynamic range), popular in detection applications, produce many bits of noise that do not compress very well. The VBB system was designed to have only sufficient gain to represent teleseisms adequately in a single 20 sample/sec data stream, therefore maximizing dynamic range, and minimizing the number of bits required to represent background noise. At the expense of slight resolution loss (not much, because many global sites already have high SP seismic noise) of SP signals, compression effectiveness at many global sites (and for other applications as well) could be increased significantly by numerically scaling down the output signal by a factor sufficient that the resolution referred to the system input is still capable of recording ambient background noise. At many global sites, where typical noise is 20dB above the LNM at 1 Hz (Peterson), digital scale factor could be reduced by a factor of 2-4 without loss of signal resolution. Adequate long-period resolution would remain in the broad-band signals because of the considerable over-sampling of low frequencies (below 1 Hz). This may be done automatically, in response to daily, or seasonal changes in background.

2.0 Software for Level 1,2,3 compression

The "steim123" program will read ascii files in the very simple EARTH1 format, and generate level 1, 2, or 3 compressed-binary output files, with or without a header, in either intel or sparc byte order, and will print execution time information. It is recommended to read the supplied source for further details. The C functions "compress_frame" and "decompress_frame" are written to stand alone, and to be used freely. Please use the software, but respect the public license terms.

The disk includes all of the data files above, in compressed (level 3) format, which may be decompressed using either the compiled C or Pascal (identical) version of "steim123".

To decompress a data file called “ANMBHZ89.S3C”, type

```
steim123 anmbhz89,s3c -d
```

This will generate the original ascii text file, ANMBHZ89.TXT.

An example compression run is shown below:

```
C:>steim123 majbhz94.txt
steim123 compressor version 940218-C, Copyright (C) 1994 by Joseph M. Steim
steim123 comes with ABSOLUTELY NO WARRANTY
This is free software; you are welcome to redistribute and modify it under
conditions of the GNU General Public License, Free Software Foundation, Inc.

compression level (1,2,3)? 3
iterations (to improve timing accuracy, 3-10 is a good number)? 1
iteration: 1
first 5 lines of file header info:
  0.5000000E-01  -3158.000      -1815.000      -12345.00      -12345.00
         0.       3633.050             0.      -12345.00      -12345.00
  -12345.00      -12345.00      -12345.00      -12345.00      -12345.00
  -12345.00      -12345.00      -12345.00      -12345.00      -12345.00
  -12345.00      -12345.00      -12345.00      -12345.00      -12345.00
file: majbhz94.dat    last frame:
fffffbbf aaf3998e 04cfbf86 41901d81 41cd0e81 a803ef4d 06800580 000000c8
000000c8 000000c8 000000c8 000000c8 000000c8 000000c8 000000c8 000000c8
level: 3
samples read=72640   bytes read=290560
samples processed=72640    samples_verified=72640
samples padded in last frame=9  samples not processed=0
frames=786   total compressed bytes=50304
squeezed flags=759  histogram of storage units (bits) used:
fit  3=      81  fit  4=   11212  fit  5=   46050  fit  6=   15097
fit  8=     196  fit 15=       4

compression ratio vs. 32-bit: 5.78

execution times normalized per iteration:

  3.30 sec  file i/o time
  0.88 sec  compression time
  0.67 sec  decompression and verify time
  5.44 sec  total run time

compression time (secs) for 50000 samples: 0.61
```