20231116-1

nanometrics

# Seismic Data Compression and Telemetry Bandwidth Considerations for EEW

Michael Laporte (michaellaporte@nanometrics.ca), Michael Perlin, Marian Jusko, Ted Sommerville, Bruce Townsend
Nanometrics, Inc., Kanata, Ontario, Canada

## BACKGROUND

Earthquake early warning systems depend on the prompt, reliable arrival of seismic data at network data centers. Network operators invest significant resources into the design, installation and operation of real-time acquisition systems to ensure maximum data completeness and minimum data latency, to allow EEW processing modules to detect events and issue warnings as quickly as possible.

These mission critical acquisition systems must perform before, during and after earthquakes, as main shocks are frequently preceded by foreshocks and followed by aftershocks, which are often just as dangerous. As such, a key consideration in the design of these systems is the impact that large earthquakes may have on their performance. Seismic data is generally encoded using Steim compression, which is a first difference algorithm. During large events the differences between samples grow, requiring more bits to record and, thus, increasing the data volume. The decrease in compression for energetic signals creates a surge of data generated by large events. Transmission and processing systems engineered for typical conditions may appear to operate well until a large earthquake strikes, and then they may become congested and fail to deliver the data just when it is most needed. System designers and network operators must be fully aware of this effect and plan for it accordingly.

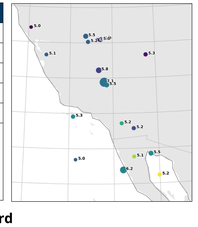
This phenomenon was identified as a major factor during the ShakeAlert processing of the 2019 Ridgecrest sequence. Many stations with marginal communications links experienced significant increases in data latency, which prevented that them from contributing to early warning solutions (Stubailo et al, 2020).

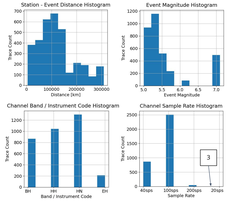
**This study assesses the impact of large events on seismic data compression and the corresponding spikes in throughput and packet rate which must be supported by real-time acquisition systems. The study examines the relationship between compression and various factors, including magnitude, epicentral distance, sample rate and instrument type.**

## METHODOLOGY

- The study examined a catalog of real events in Southern California.
- For each event, the corresponding waveform data was retrieved from the SCEDC and written to disk in miniSEED format using 512B records and Steim1 compression using the Obspy Stream write function.
  - 512B and Steim1 were chosen to be representative of the typical data encapsulation for the real-time streaming protocol SeedLink.
- The miniSEED files were then scanned using the ObsPy script RecordAnalyser to calculate the average data compression, in Bytes per sample [B/s], for each record. Maximum, or best case, compression corresponds to 1 B/s, whereas the minimum, or worst case, compression, corresponds to 4 B/s.
  - The average data compression for each record serves as the basic building block that allows modelling of any real-time streaming protocol that uses Steim1-based data encapsulation, such as SeedLink and NP.
- The average data compression for each record was then averaged over time windows of interest relative to the P phase arrival for each trace. This included a window before the onset of the arrival to establish baseline compression for the trace.

| EVENT CATALOG | |
|---|---|
| Source | Southern California Earthquake Data Center |
| Count | 23 Events, with 3423 traces |
| Time Period | 2013-01-01 to 2023-01-01 |
| Region | Latitude: 30N to 39N ; Longitude: -111E to -124E |
| Magnitude Range | M > 5 |
| Inclusion Criteria | • Must be classified an **earthquake**<br>• Must have a **manually** generated preferred magnitude that has been **reviewed**<br>• Must have a **final** origin with **P phase** arrivals that map to **manual** picks that have been **reviewed**<br>• Stations must have complete data for full event |






**Average Data Compression from miniSEED Record**

- Steim payloads comprise multiple of 64B Steim frames. The first frame has 12B of overhead, leaving 52B for compressed samples. The subsequent frames have 4B of overhead, leaving 60B for compressed samples.
- 512B miniSEED records comprise one 64B header frame and seven 64B Steim frames, yielding a total of 412B of compressed data.
- The average data bytes per sample, for each record, can be calculated based of the number of samples in the record, npts, which is reported by the RecordAnalyser script:

$$\text{Average Data Bytes per Sample} = \frac{\text{Data Bytes per Record}}{\text{Number of Samples per Record}} = \frac{52 + (60 \times 6)}{\text{npts}} = \frac{412}{\text{npts}}$$

## DATA COMPRESSION

**Seismic Data Encapsulation**

The plot below depicts the encapsulation of seismic data into packets during an earthquake. The black vertical lines identify packet start and end times. The blue trace corresponds to the average number of data bytes used to record each sample, calculated per packet. The effect of the P phase arrival is evident as the packet durations decrease due to the increased data volume as data compression becomes less effective.

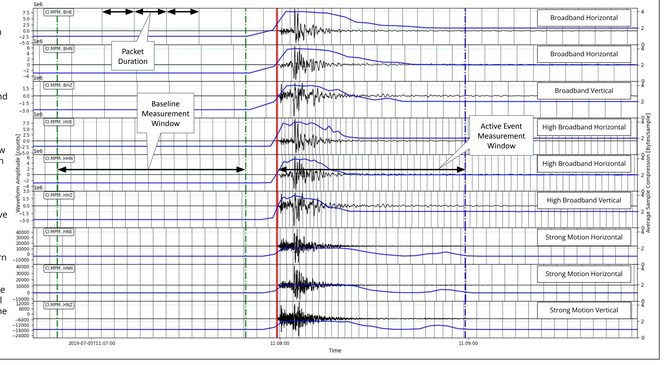
Data from station MPM for a M5.37 event in 2019 located at an epicentral distance of 34 km.

The red vertical line corresponds to the P phase arrival.

The green vertical lines correspond to the time window used to calculate baseline data compression during normal conditions. The P phase arrival to the blue vertical line is the window used to calculate the compression corresponding to an active event.

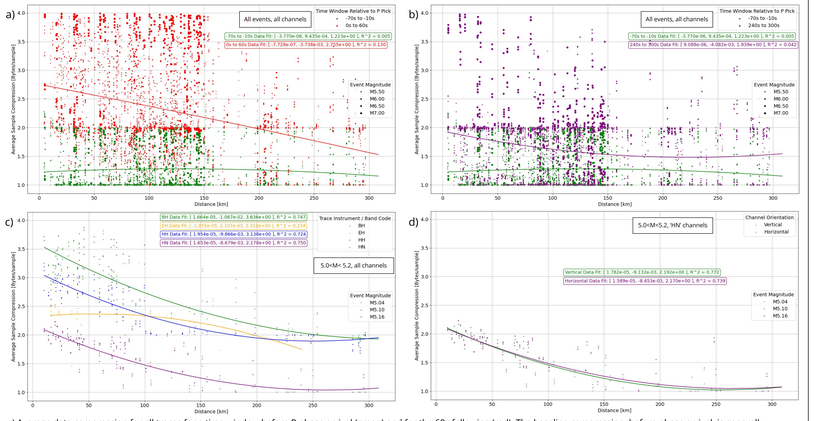
The average bytes per sample (blue trace) increases during the shaking associated with the P wave arrival. The packet durations decrease accordingly.

The strong motion channels return to the baseline compression level after approximately one minute. The broadband channels decrease from the peak during the P arrival but remain well above the baseline level for the remainder of the trace.




**Average Data Compression Influences**

The data points in the plots below present the average data compression (bytes per sample) for different waveform time windows relative to the arrival of the P phase. The x axis is epicentral distance from the event to the station. The corresponding event magnitude is reflected by the marker size. The marker color is used to highlight groupings of channel characteristics of interest. Quadratic regression was applied to model compression for each group versus epicentral distance. In order to obtain statistically relevant results, it was necessary to organize the data into smaller bins based on event magnitude and instrumentation characteristics.



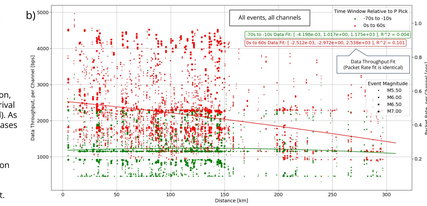

a) Average data compression for all traces for a time window before P phase arrival (green) and for the 60s following (red). The baseline compression, before phase arrival, is generally between 1 and 2 bytes per sample. As expected, compression decreases significantly, by up to a factor of 4, following the P arrival, particularly for stations closer to the event. A large number of channels exhibit reduced compression, approaching the maximum of 4 bytes per sample, even for smaller events and some distant events.
b) Average data compression for all traces for a time window before P phase arrival (green) and for a 1 minute period beginning 4 minutes following (purple). The average compression has returned to 1 or 2 bytes per sample for most traces, but many continue to exhibit reduced compression.
c) Average data compression for all traces grouped by Band and Instrument code for the 60 second period following P phase arrival for events with magnitude in the range 5.0 to 5.2. As expected, the strong motion channels ('HN') have the highest compression throughput, rarely exceeding 2 bytes / sample, due to their reduced sensitivity. The High Broadband channels (HH) exhibit better compression than the Broadband channels (BH), likely due to lower sample to sample deltas resulting from the reduced time interval between samples for HH channels, which use a 100 Hz sampling rate, in comparison to 40 or 20 Hz for BH channels. The EH channels exhibit unexpected results which are under investigation.
d) Average data compression for all strong motion traces ('HN') grouped by channel orientation for the 60 second period following P phase arrival for events with magnitude in the range 5.0 to 5.2. Compression is similar for both vertical (green) and horizontal (purple) channels.

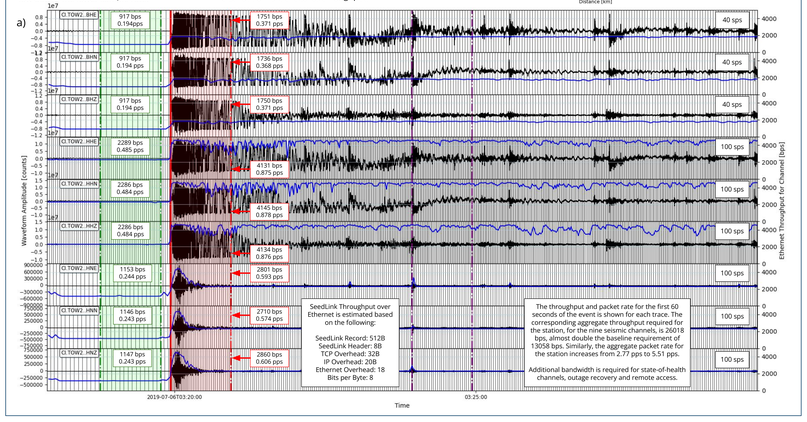
## TELEMETRY BANDWIDTH & PACKET RATE

The plots right and below present the throughput required for each channel, and corresponding packet rate, based on the measured data compression. Telemetry bandwidth and packet rate processing requirements can be determined by summing the result for each channel and adding appropriate margin for outage recovery and remote access.

a) Data from station TOW2 for the 2019 Ridgecrest M7.1 event. This station was close to the event, with an epicentral distance of 16 km, and experienced heavy shaking which caused the broadband channels to clip. The throughput (blue trace) increases during the heavy shaking associated with the P wave arrival. Peak demand is high on High Broadband (HH) and strong motion channels (HN), but drops off quickly for strong motion. The demand for Broadband channels (BH) increases but remained relatively low due to the lower sample rate.

b) Throughput demand for all traces, based on average sample compression, for a time window before P phase arrival (green) and for the 60s following (red). As expected, throughput demand increases significantly, following the P arrival, particularly for stations closer to the event. The horizontal throughput concentrations correspond to common sample rate and compression combinations. The packet rate data distribution is identical to throughput.




## CONCLUSIONS & NEXT STEPS

- The shaking associated with phase arrivals significantly limits the compression of data encapsulated in Steim1 frames, leading to a sudden spike in data volume and, correspondingly, the throughput and packet rate which must be supported by real-time telemetry, networking and data processing systems. This effect becomes more severe for larger events and as station epicentral distance decreases. Characterizing this effect requires limiting data sets to small magnitude ranges and instrumentation groupings.
- The effect is similar for vertical and horizontal channels. Both broadband and strong motion channels experience the initial spike. Strong motion channels return to baseline immediately whereas broadband channels can take a very long time to do so. Channels with lower sample rates or more susceptible due to the longer period between samples allowing greater deltas, but generally require lower throughput overall due to less samples being streamed.
- **Planning for this effect, considering the distinct loading of sudden spikes in both overall throughput and packet rate, is vital to the reliable operation of mission-critical real-time acquisition systems, such as those used for EEW. This involves both system design and ongoing preventative maintenance. Regular, proactive acquisition system testing is required to maintain confidence that the system will perform as expected when needed. Deploying systems with centralized tools that support this testing in a manner that is not disruptive to the ongoing operation of the network will help ensure it happens. A recommended technique, employed in some EEW networks, involves instructing stations to temporarily transmit data uncompressed (4B/s), but still in Steim1 format, to exercise the maximum data volume case. Since the data is real, ongoing earthquake monitoring is not disrupted.**
- More detailed analysis, and expansion of the data set to provide more uniform distribution of key characteristics (ex. epicentral distance > 150 km), is required to develop broadly applicable empirical models of the impact of heavy shaking on data compression and throughput.
- Future studies will examine the role of system sensitivity, the potential impact on outbound traffic for TCP-based protocols (SeedLink), model alternative data encoding techniques (ex. Steim2) and model alternative real-time streaming protocols (NP).

## REFERENCES


- Stubailo, I., M. Alvarez, G. Biasi, R. Bhadha, and E. Hauksson (2020). Latency of Waveform Data Delivery from the Southern California Seismic Network during the 2019 Ridgecrest Earthquake Sequence and Its Effect on ShakeAlert, Seismol. Res. Lett. XX, 1–17, doi: 10.1785/0220200211
- Southern California Earthquake Data Center, SCEDC doi:10.7909/C3WD3xH1
- California Institute of Technology and United States Geological Survey Pasadena. (1926). Southern California Seismic Network [Data set]. International Federation of Digital Seismograph Networks. https://doi.org/10.7914/SN/CI
- Beyreuther, M., Barsch, R., Krischer, L., Megies, T., Behr, Y., and Wassermann, J. (May/June 2010), ObsPy: A Python Toolbox for Seismology, Seismological Research Letters, 81 (3), 530-533.
- SEED Reference Manual, SEED Format Version 2.4, August, 2012