Error-bounded Fixed-ratio Lossy Compression for Scientific Data

Jon Calhoun*, Robert Underwood*, Sheng Di' and Franck Cappello'
*Clemson University, 'Argonne National Laboratory

Large-scale scientific simulations and instruments are capable of generating large volumes of data at a high velocity. To handle large data sets, users need to reduce the data volume for a variety of use cases: lower storage requirements, improve streaming and in-situ processing of data from scientific instruments, increase network and I/O bandwidth, and reduce the memory footprint of applications. As scientific applications scale to exascale systems and scientific instruments generate data at a faster rate, this problem will worsen. For example, the SKA (Square Kilometre Array) Phase 1 will generate ~300 PB/year of science data in 2023. It is estimated that 1 EB of science data will be generated from HL-LHC (High Luminosity-Large Hadron Collider) experiments in 2026 [1].

To reduce the size of data sets, data reduction techniques such as decimation are common, which only keeps a subset of the data generated. As more data is discarded, larger errors are possible in the analysis. Data compression is another alternative to combat large data sizes. By reducing the number of bytes needed to store a data set more data sets are able to be stored. For floating-point scientific data, lossless compression fails to achieve compression ratios required by scientists; therefore, scientists use lossy compression, which trades accuracy in the data for large compression ratios [2]. Current lossy compressors for scientific data such as SZ, ZFP, and MGARD allow users to bound the error in the decompressed data with various metrics and tolerances. Currently, if these compressors are required to compress to a target ratio, the accuracy is no longer guaranteed without human-in-the-loop trial-and-error. Without the ability to simultaneously control both the accuracy and the compression ratio, it is difficult to predict storage requirements and bandwidth requirements for data sets while ensuring the data is usable for workflows. Moreover, the acceptable error bounds changes from domain to domain and data set to data set. Adapting existing lossy compressors to simultaneously address error bounding and compression to a fixed-ratio will benefit many domains. This problem requires a black-box approach and solving an multivariate optimization problem raises the following non-trivial research questions:

1. What algorithms solve the multivariate optimization problem and do not get bogged down in local minimas due to a nonmonotonic relationship between the error bound and the compression ratio?
2. How can the framework enable broad use by supporting a variety of error bounding metrics including domain specific error metrics?
3. Are there search algorithms fast and efficient enough to be used for in-situ use cases?
4. How to handle infeasible goals?

Recently we proposed a generic, efficient fixed-ratio compression framework, FRaZ, that determines the error bound accurately for various error-controlled lossy compressors, given the particular target compression ratio with a specific scientific floating-point dataset [4]. Our design involves three critical strategies. First, we develop a global optimum searching method by leveraging Davis King's global minimum finding algorithm [3] to determine the most appropriate error bound based on the given compression ratio and dataset. Second, our parallel algorithm decreases the runtime of the parameter search by splitting the search range into distinct regions, parallelizing on file, and by time-step in the offline case. Lastly, unlike previous approaches, we treat the compressors as a black-box making our approach robust to changes in the compressors.

Figure 1 shows the bit rate (the number of bits used per data point after the compression) versus the data distortion measured by the peak signal-to-noise ratio (PSNR), a common indicator to assess the data distortion in

the community, for three HPC data sets and the top HPC lossy compressions: SZ, ZFP and MGARD. In general, the higher the PSNR, the higher the quality of decompressed data. In this figure, we see that ZFP (FRaZ) provides consistently better rate distortion than does ZFP (fixed-rate) across bit rates (i.e., across compression ratios). ZFP (fixed-rate) only allows for setting the target compression ratio and does not allow simultaneously bounding the error. Moreover, SZ (FRaZ) exhibits the best rate distortion in most cases. However, ZFP (FRaZ) does give better results for small bit-rates in Figure 1(b) indicating that no single lossy compression is best for all situations. Overall, error-bounded lossy compressors that leverage FRaZ maintain high fidelity in the data during the compression, by leveraging the error-bounded lossy compression mode for different compressors.

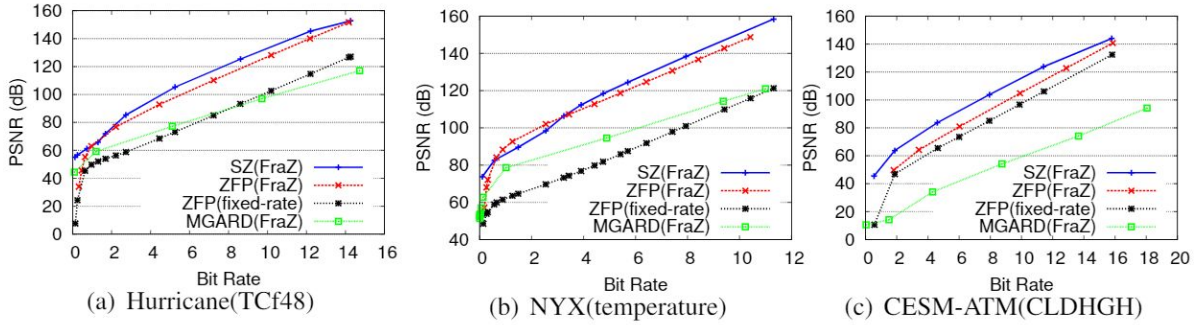

Figure 1: Rate distortion of lossy compression.

FRaZ is an initial autotuning framework that treats compressors as a black-boxs and solves a nonlinear derivative free optimization problem to select the error bound that yields a target compression ratio. Thus, FRaZ produces fixed-ratio error-controlled lossy compression for scientific floating-point HPC data sets, but a number of areas for improvement exist. Extension to respect custom user error metrics that correspond with the quality of a scientist's analysis result relative to that on noncompressed data will ensure that data is reliability reduced in size. Accelerating the performance and convergence rate of FRaZ would enable it to be used online as an in-situ fixed-ratio compressor for simulation and instrument data. Finally, selecting which compressor to use based on the accuracy and storage constraints to yield the best rate distortion.

References:

[1] M. Neubauer, IRIS-HEP Blueprint: Concepts and Process, FAST ML Workshop, Fermilab, Sept 2019. [Online] https://indico.cern.ch/event/822126/contributions/3500165/attachments/1905471/3146855/IRIS-HEP_BluePrint_FML.pdf

[2] S. W. Son, Z. Chen, W. Hendrix, A. Agrawal, W. keng Liao, and A. Choudhary, “Data compression for the exascale computing era-survey,”Supercomputing frontiers and innovations, vol. 1, no. 2, 2014.[Online]. Available: http://superfri.org/superfri/article/view/13

[3] D. King. Dlib C++ Library - Optimization. [Online]. Available: http://dlib.net/optimization.html#global_function_search

[4] R. Underwood, S. Di, J. Calhoun, and F. Cappello, “FRaZ: A Generic High-Fidelity Fixed-Ratio Lossy Compression Framework for Scientific Floating-point Data,” To appear in the IEEE International Parallel and Distributed Processing Symposium (IEEE IPDPS). IEEE, 2020.