Towards Efficient Error-controlled Lossy Compression for Scientific Data
Sheng Di1, Dingwen Tao2, Hanqi Guo1, Zizhong Chen2, Franck Cappello1  
1 Argonne National Laboratory, Lemont 2 University of California, River Side

Abstract
Due to vast volume of data being produced by today’s scientific simulations and experiments, lossy data compressor allowing user-controlled loss of accuracy is widely used, and scientific data compression provides a good solution for significantly reducing the data size. In this work, we first define a new error-controlled lossy compressor, namely SZ-1.4, for large-scale scientific data, and then we propose several components based on their application and improvement. We argue that the SZ-1.4 can be extended to other related works, with comparable compression/decomposition rate. We also provide some tools to visualize the data and compressions on output. Our goal is to provide a production quality lossy compressor for scientific data respecting user set error bounds.

Features of ANL SZ lossy compressor:
- Error-bounded feature: maximum compression errors can be strictly bounded based on user demand.
- High compression rate: higher compression factor based on specific error-bound than other related works, with comparable compression/decomposition rate.
- Flexible prediction-based compression model: allowing users to customize their own prediction method to optimize the compression quality.

SZ Compression Technique
1) Multi-dimensional Multi-layer prediction
2) Uniform Scalar Quantization of the prediction error from the error bound
3) Error Bound
4) Lossless compression (L77, Huffman)

Z-checker tool
We integrate in Z-checker assessment algorithms and functions that are as comprehensive as possible. (1) Z-checker can be used to characterize critical properties (such as entropy, distribution, power spectrum, PCA, autocorrelation) of any data set, such that the difficulty of data compression can be presented clearly in the granularity of data blocks. (2) Not only is Z-checker able to check the compression quality (compression ratio, bit rate), but it also provides various global distortion analysis comparing the original data with the decompressed one (PSNR, normalized MSE, rate-distortion, rate-compression error, spectral, distribution, derivatives) and statistical analysis of the compression error (maximum/minimum/average error, autocorrelation, distribution of errors). (3) Z-checker can also assess the impact on the lossy decompressed data on some common transform functions, such as discrete Fourier transform (DFT) and discrete wavelet transform (DWT). (4) Z-checker also provides two ways to visualize the data and compression results on demand. Specifically, Z-checker may help generate data figures by static scripts or by an interactive system.

We show how we used Z-checker to improve the compression performance of the SZ lossy compressor for hard-to-compress data sets.

We implemented the Z-checker software and will release it as an open-source community tool, under a BSD license. To the best of our knowledge, Z-checker is the first tool designed to comprehensively assess compression results for scientific data sets across multiple lossy compressors.

Z-checker Results
Z-checker Visualization of the Entropy (Block) with Different Accuracies on CESM Data Sets

Compresson Ratio Result and Rate-Distortion Result
Comparison of time to compress/decompress and write/read compressed data against time to write/read initial data on Blues.

Acknowledgment
This research was supported by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations - the Office of Science and the National Nuclear Security Administration, responsible for the planning and preparation of a capable exascale ecosystem, including software, applications, hardware, advanced system engineering and early testbed platforms, to support the nation’s exascale computing imperative. The submitted manuscript has been created by UChicago Argonne, LLC, Operator of Argonne National Laboratory (“Argonne”). Argonne, a U.S. Department of Energy Office of Science laboratory, is operated under Contract No. DE-AC02-06CH11357. The U.S. Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.