Lossy compression for transferring, storing and analyzing huge scientific datasets

Sheng Di, Franck Cappello, University of Chicago/Argonne National Laboratory

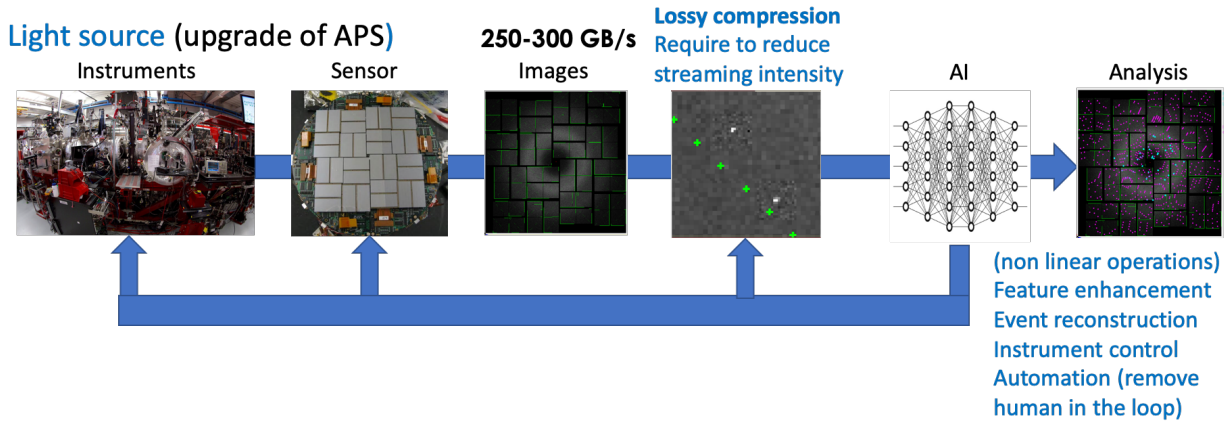
Background

Scientific and engineering simulations and experiments are producing ever-increasing amounts of data to the point where the data volume and velocity become unbearable. For some applications, it becomes impossible to transfer or store all the data either used as inputs or produced. For other applications, transferring the data through the network or storing the data on file systems become serious performance bottlenecks. This problem will worsen with Exascale systems and updated physics instruments. For example, it is estimated that phase 1 of the SKA (Square Kilometer Array) will generate ~300 PB/year in 2023 and the HL-LHC will generate 1 EB of science data in 2026.

Error-bounded lossy compression has been very effective in significantly reducing vast volume of data for scientific simulations, compared with lossless compressors such as Gzip and Zstd which have very limited compression ratios (~2:1 or less) in general. The SZ [1,2] compressor, for example, is able to reduce the data size by 10× or even 100×, with acceptable data distortion according to user requirements for different applications, such as climate, quantum chemistry and cosmology among others.

Generic or customizable lossy compressors like SZ are appealing to domain scientists because: 1) they do not need to develop and maintain their own compression tools, 2) as shown by the LHC experience, developing ad hoc data reduction algorithms reaching high performance in terms of reduction ratio, accuracy and speed requires significant development and optimization efforts that is just not bearable for many simulation and experiment users, 3) customizable compressors being community tools, users can benefit from compression performance progress made by others, 4) existing lossy compressors (SZ and ZFP) provide strict accuracy guarantees with multiple error controls, and 5) they are optimized for different platforms (CPU, GPUs, FPGAs).

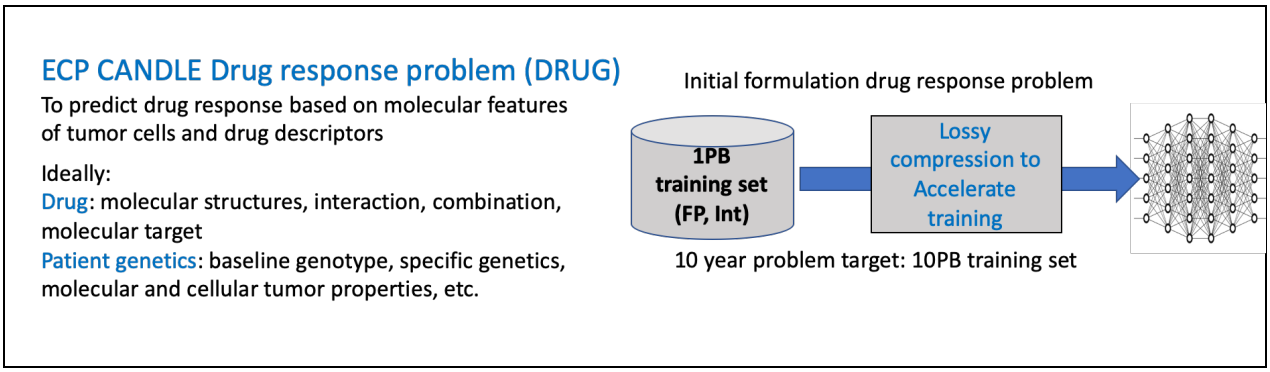
Many use cases of lossy compression for scientific data have been identified including reducing storage footprint, accelerating I/O (reducing transfer time) and reducing streaming intensity [3]. For example, collaborative research is currently under development between physics and computer science researchers in order to reduce the streaming intensity of lights source instruments.



Specifically, the Linac Coherent Light Source (LCLS) [4], the Argonne Photon Source and the Spring-8 instrument in Japan are exploring the application of lossy compression to prepare for their upgrades. It is projected that these instruments will produce images at the rate of 250-300GB/s for scientists performing crystallography and ptychography analysis. The current plans are to connect these instruments to large supercomputers at Argonne, NERSC and Riken to perform analysis; which requires to transmit extremely large volumes of high velocity data on medium or long-distance networks. Current experiments with LCLS datasets show that lossy compression techniques can get high compression ratios (20+) while still meeting the required scientist post-analysis accuracy. However, the integration of lossy compression in the light sources data production/analysis pipeline is raising many

important scientific questions, in particular with the integration of AI for data analysis to resolve non-linear problems such as feature enhancement, event reconstruction and instrument control automation.

The second example is the CANDLE DRUG problem [5]. One of the goals of this project is to design a DNN capable of assessing tumor response to the drug, given information about the tumor and potential drug treatment. The training set will gather Petabytes of information from drug (molecular structure, interaction, combination, etc.) and patient genetics. This dataset will be stored in a database that would be accessible for training in different supercomputers. Communication of the information between the storage and the supercomputers is a major challenge. Early experiments have shown that lossy compression can be used in that case to reduce significantly (x30) the size of exemplar training sets.



While early results of the use of lossy compression for huge scientific datasets are encouraging, there are still many open research questions and high impact research opportunities:

(1) How to develop lossy compressors that would offer relevant solutions in terms of preservation of science opportunities for high volume and high velocity data generated by scientific simulations and instruments, stored in large databases and transferred on medium and long-distance networks? The difficulty here is that different applications and instruments need different compression algorithms and different control of the lossy compression distortion requiring the design of adaptable compression pipelines.
(2) The performance of configurable compression pipeline depends on the selection of the compression stages and the selection of the parameters for each stage. Autotuning is needed to help users find the right configuration for their problems. Autotuning of lossy compression pipeline is a high dimensional, non-convex optimization problem.
(3) How to control the compression distortion to respond to user requirements? Establishing a link between the acceptable distortion on the analysis results and the control mechanisms of lossy compressors is non-trivial. In some situations, one can establish a formal mathematical link between the two. In other cases, black-box search is needed.
(4) How to develop methodologies, tools and metrics to assess the impact of lossy compressor distortion on scientific datasets? The main problem here is that users are interested in preserving the results of the analysis they are performing on their scientific datasets. Analysis results cannot not be preserved 100% with lossy compression. So, users need to define relevant metrics and acceptable distortion due to compression.
(5) How to develop high speed implementation of lossy compression algorithms in GPUs, FPGA and ASICs responding to user needs. The difficulty here is that advance high-performance compression pipelines are quite complex and software optimizations as well as hardware implementations are non-trivial.

Reference

[1] SZ: https://collab.cels.anl.gov/display/ESR/SZ
[2] Sheng Di, Franck Cappello, "Fast Error-bounded Lossy HPC Data Compression with SZ", International Parallel and Distributed Processing Symposium (IEEE/ACM IPDPS 2016), 2016.
[3] F. Cappello, et al., "Use cases of lossy compression for floating-point data in scientific datasets", in The International Journal of High Performance Computing Applications (IJHPCA), 2019.
[4] LCLS-II: https://lcls.slac.stanford.edu/lcls-ii
[5] Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer: https://candle.cels.anl.gov/