Evaluating SZ3 Compressor Performance on High Energy Physics Data

Amy Byrnes, Serhan Mete, Jahred Adelman, Peter van Gemmeren, Michael E. Papka

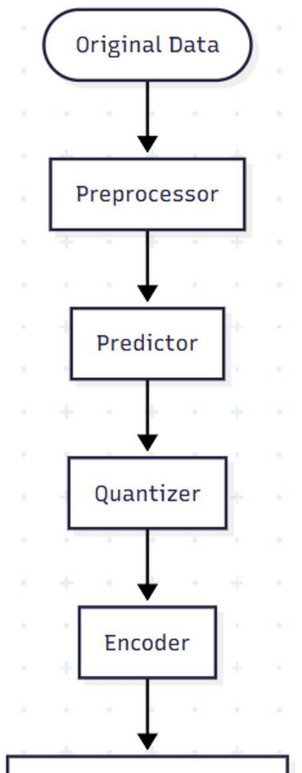
Motivation

- **The Large Hadron Collider (LHC) has created hundreds of petabytes of data. The High-Luminosity LHC (HL-LHC) will produce orders of magnitude more.**
  The HL-LHC is a major upgrade to the LHC, aiming to be operational by the end of 2030. **It will produce more data in one year than the LHC produced during its first 10 years of operation.**
- **Significant R&D efforts are needed to handle the incoming tide of data.**
  Without using new technologies, the data volume of the HL-LHC will rapidly outgrow CERN's compute and storage capacities. This means everything is on the table, even historically disfavored methods like lossy compression.

ATLAS Data

- **ATLAS (A Toroidal LHC ApparatuS) is one of two general-purpose detectors at the LHC.** ATLAS and CMS (the other general-purpose detector) are responsible for the observation of the Higgs boson in 2012.
- **The data collected by ATLAS represents physics objects resulting from particle collisions.** Hadronic jets represent partons (quarks and gluons), which cannot be measured directly.
- **Quantities collected include:** Energy (E), Transverse momentum (pt), Pseudo-rapidity (eta), and Azimuthal angle (phi)
- **Many TB of real and simulated ATLAS data are available on the CERN Open Data portal** – including the data used in this work. The data are provided under a Creative Commons C0 license.

SZ3 Compression Framework



- **SZ3 is "a modular, error-bounded lossy compression framework for scientific datasets".** SZ3's flagship features are its modularity, making it straightforward to build a custom compression pipeline, and its use of best-fit predictors. It's source code is available on GitHub under a BSD license.
- **Predictors are generally used in compression pipelines to reduce the entropy of the input data.** This ultimately leads to smaller outputs. All predictors rely on the assumption that the value currently being compressed can be predicted with reasonable accuracy from some small number of previous values.
- **SZ3's predictor attempts to fit the current data to a (linear, quadratic) curve with the immediately preceding data points.** The initial SZ compressor was developed with scientific simulation data in mind. This prediction model is well suited to data where indices correspond to contiguous physical coordinates (e.g. temperature at a point (x, y, z)).
- **The "loss" incurred during lossy compression is the result of quantization**. Scientists are understandably hesitant about losing any amount of precision from their data. Quantization is a well-defined process, so maximum error bounds can be imposed. SZ3 defaults to lossless compression with zstd if it detects that lossy compression is not possible with the given error bound.

Acknowledgements

- This work is supported by the Chicagoland Computational Traineeship in High Energy Particle Physics (C^2P^2) under US Department of Energy Office of Science grant DE-SC0023524.

Evaluation

- **ALGO_NOPRED reduces data sizes by 66-83%, while keeping the data within 0.1-2% of its original values, on average. This is better than any of SZ3's other (lossy) compression methods.**
  ALGO_NOPRED achieves the highest compression ratios outright, and the least error and fastest compression/decompression times on average.
- **Events are independent, negating the assumptions that make prediction effective.** Data are recorded per-collision-event, and the outcomes of these events are independent from one another. There is no meaningful relationship between adjacent data points. ALGO_NOPRED is the only lossy method in SZ3 that skips prediction entirely. Its superior performance supports our intuition that this data is not suited for prediction-based methods.
- **Some data is more compressible than other data, but more investigation is needed to understand why.** The data which achieves the highest compression ratio of 6 has, predictably, the lowest entropy. But other data which compress relatively well (ratios of 4.5, 5.4) have entropies equivalent to data that compress relatively poorly (2.9, 3).