# Statistical Prediction of Lossy Compression Ratios for 3D Scientific Data

**David Krasowska[1], Robert Underwood[2] (advisor), Julie Bessac[2] (advisor), Jon Calhoun[1] (advisor), Sheng Di[2] (advisor), and Franck Cappello[2] (advisor)**

[1]Holcombe Department of Electrical and Computer Engineering - Clemson University

[2]Mathematics and Computer Science Division – Argonne National Laboratory

This material is based upon work supported by the National Science Foundation under Grant No. SHF-1910197 and SHF-1943114.

## Introduction

### Why Study Lossy Compressibility?

- Error bounded lossy compressors are used within scientific research due to larger compression ratios (CRs) in relation to lossless compressors
- Entropy[1] is the mathematical limit on lossless compression; however, there is no known bound of lossy compression
- Use of data correlation structures, heterogeneity and error bounds in lossy compression techniques
- Establish entropy-like metric for lossy compression algorithms which can guide lossy compression community to an optimal development and usage
- Anticipate compression performances and adapt compressors to correlation structures to get the best CR performance possible

**Goals:**

**(1) Explore possible models of CR and quality metrics for 3D data**

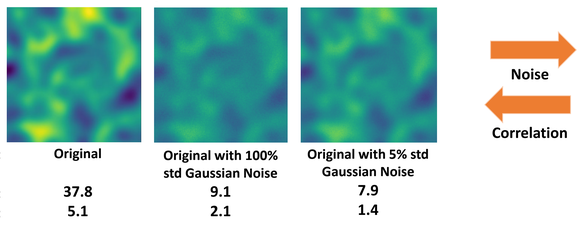
**(2) Next step towards the theoretical limit for lossy compressibility**

## Our Previous Work

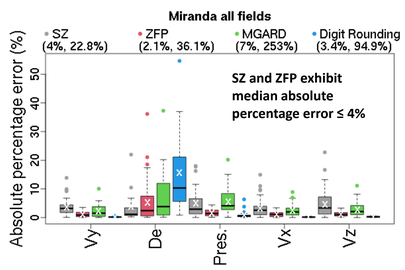
### Statistical Methods & Compression Statistics for 2D

Statistical predictors trained with datasets using notions of correlation, entropy and lossy-ness. The model, Eq. (1), relies on:

- Quantized entropy
- Singular Value Decomposition (SVD)




**How Well Does This Work?**

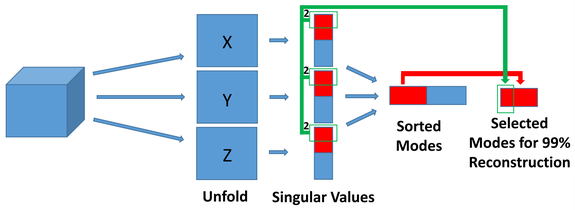



**Does this extend to 3D?**

## Methodology

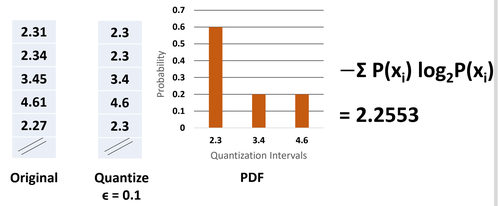
### Statistical Methods & Compression Statistics: Predictors for 3D Datasets

#### What Is Higher Order SVD (HOSVD)?




1. Unfold a tensor along each dimension
2. Perform SVD on each unfolding to obtain singular values
3. Combine singular values using at least two from each unfolded dimension to reconstruct 99% of the data
4. Perform SVD truncation (percentage of singular values needed to reconstruct 99.9% of the data)

#### What Is Quantized Entropy (Qentropy)?




$$-\Sigma\, P(x_i)\, \log_2 P(x_i) = 2.2553$$

1. Discretize original data into intervals based off the user defined error bound (ϵ)
2. Create a probability distribution function (PDF) of the different symbols used
3. Calculate the entropy using the PDF

#### Compressors and Software

Lossy Compressors:

- SZ[2] @2.1.12.2
- ZFP[3] @0.5.5
- MGARD[4] @1.0.0
- Bit Grooming[5] @2.9.0
- TTHRESH[6] @0.0.5

Software:

- Libpressio[7] @0.83.1
- LLVM @12.0.1
- Julia @1.7.2
- CUDA @11.7.0
- CUSOLVER @11.3.5

#### Regression Model for 2D and 3D

- Regression models, Eq. (1), are trained on observed CR of the studied datasets and the statistical predictors
- Least-square techniques estimate parameters from observed training datasets and then used to make CR predictions from new predictor values
- K-fold cross-validation assesses without bias or over-fitting the predictive capabilities of the regression models

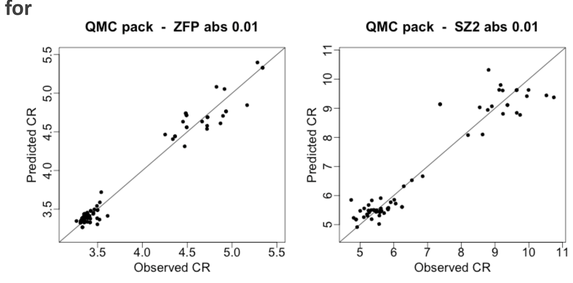
$$\log(\mathrm{CR}) = a + b \times \log(\mathrm{Qentropy}) + c \times \log\left(\frac{\text{SVD-trunc}}{\sigma}\right) + d \times \log(\mathrm{Qentropy}) \times \log\left(\frac{\text{SVD-trunc}}{\sigma}\right) + \epsilon \qquad (1)$$

## Results

### How Accurate Is The 3D Approach?

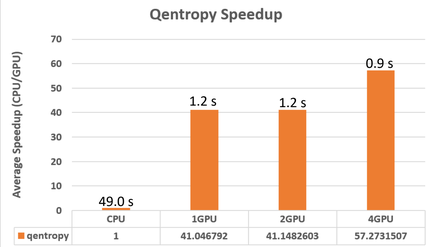
- Data: 288 3D orbitals from QMCPack from SDRBENCH [8] containing structures of atoms, molecules, and solids.
- Median absolute percentage errors (MAPE) is the difference between the predicted and observed CRs on the validation set
- The predicted CR exhibits low MAPEs (< 7.5%) for SZ2, ZFP, MARD, and Bit Grooming
- However, TTHRESH produces a higher error

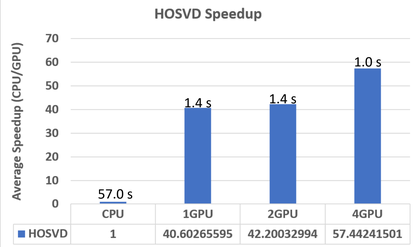
| Compressor | MAPE (median percentage error) | 10% Quantile | 90% Quantile |
|---|---|---|---|
| SZ2 | 4.5% | 3.2% | 5.7% |
| ZFP | 1.7% | 1.3% | 3.5% |
| MGARD | 0.6% | 0.4% | 1.3% |
| Bit Grooming | 7.4% | 5% | 9.3% |
| TTHRESH | 24.8% | 15.7% | 27.7% |




## Performance

### How Fast Is The 3D Approach?




The HOSVD is a slow algorithm even in parallel on the CPU; therefore, an accelerated version is needed. We implemented a multi-GPU parallel version to be used with CUDA for Nvidia cards [9].

The HOSVD and Qentropy were measured on average performance over 6 runs on the baryon_density buffer (512x512x512) from the NYX dataset [8].

NVIDIA A100 GPU scaling with DGX node on the Palmetto Cluster [10].

**The maximum speedup (CPU / GPU) is 57x**

## Conclusions

- Ability to accurately predict CRs in 3D is comparable to 2D
- Flexible across compressors, error bounds, and datasets
- Statistical predictor reuse allows for comparison of different compressors to find largest CR
- Next step towards theoretical quantification of lossy compressibility

## Future Work

- Further reduction of computation costs
  - ➢ Generate training samples from blocks of the 3D tensor data
  - ➢ Estimate CR using the samples and our predictors

CONTACT ME | Github

[1]C. E. Shannon, "A mathematical theory of communication," in The Bell System Technical Journal, vol. 27, no. 3, pp. 379-423, July 1948, doi: 10.1002/j.1538-7305.1948.tb01338.x.

[2]https://github.com/disheng222/SZ

[3] https://github.com/LLNL/zfp

[4] https://github.com/CODARcode/MGARD

[5] Zender, C. S.: Bit Grooming: statistically accurate precision-preserving quantization with compression, evaluated in the netCDF Operators (NCO, v4.4.8+), Geosci. Model Dev., 9, 3199–3211, https://doi.org/10.5194/gmd-9-3199-2016, 2016.

[6] https://github.com/rballester/tthresh

[7]R. Underwood, S. Di, J. C. Calhoun, and F. Cappello, "FRaZ: A Generic High-Fidelity Fixed-Ratio Lossy Compression Framework for Scientific Floating-point Data," presented at the 34th IEEE International Parallel and Distributed Processing Symposium, New Orleans, May 2020. https://github.com/robertu94/libpressio

[8]F. Cappello et al., "Scientific Data Reduction Benchmarks," Scientific Data Reduction Benchmarks, Jun. 18, 2018. https://sdrbench.github.io/ (accessed Jun. 02, 2020).

[9] https://developer.nvidia.com/cuda-toolkit

[10] https://www.palmetto.clemson.edu/