The 38th IEEE International Conference on Data Engineering (ICDE 2022)

Improving Prediction-Based Lossy Compression Dramatically via Ratio-Quality Modeling

Sian Jin⋆, Sheng Di†, Jiannan Tian⋆, Suren Byna‡, Dingwen Tao⋆, Franck Cappello†
⋆School of Electrical Engineering and Computer Science, Washington State University, Pullman, WA, USA
†Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA
‡Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

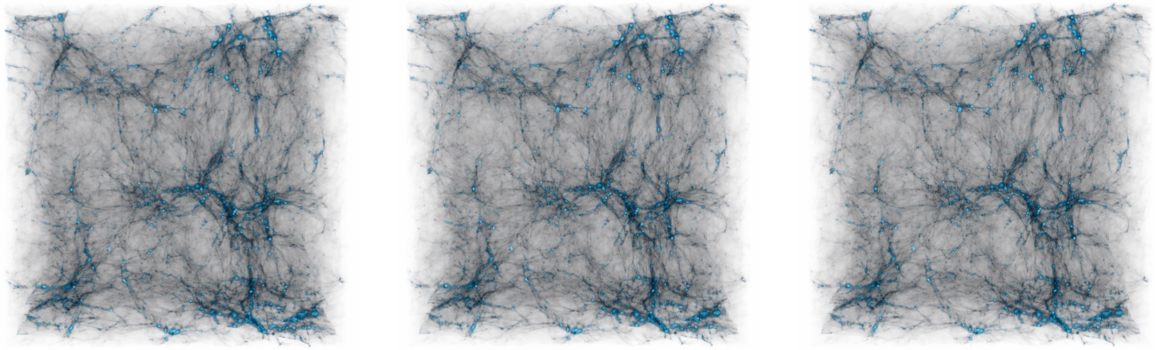
Introduction

Why Compression

- Large-scale scientific simulations generate extremely large amounts of data
- Limited storage capacity even for large-scale parallel computers
- The I/O bandwidth required to save this data to disk can create bottlenecks in the transmission

Lossy Compression

- High compression ratio
- Controllable compression error



(a) Original (b) Reconstructed with PW_REL = 0.1 (c) Reconstructed with PW_REL = 0.25

Jin, Sian, et al. "Understanding GPU-based lossy compression for extreme-scale cosmological simulations." *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2020.

Introduction

Why Compression

- Large-scale scientific simulations generate extremely large amounts of data
- Limited storage capacity even for large-scale parallel computers
- The I/O bandwidth required to save this data to disk can create bottlenecks in the transmission

Lossy Compression

- High compression ratio
- Controllable compression error

Take Advantage Of Lossy Compressors

- Identify the optimal trade-off between the compression ratio and compressed data quality
- No analytical model available
- Trial-and-error experiments
  - High computational cost
  - Identified configuration setting is dependent on specific conditions and input data

Introduction

WASHINGTON STATE UNIVERSITY

Our Ratio-Quality Modeling

- Estimate compression ratio and compressed data quality
  - General model suiting most scientific datasets and applications
  - High accuracy
  - Low computational overhead

Contributions

- We decouple prediction-based lossy compressors to build a modularized model
- We theoretically analyze how to estimate the encoder efficiency and provide essential parameters for compression ratio estimation
- We propose a theoretical analysis to estimate the qualification of lossy decompressed data on post-hoc analysis
- We evaluate our model using 10 real-world scientific datasets involving 17 fields.

Background

WASHINGTON STATE UNIVERSITY

Data Management in Scientific Applications

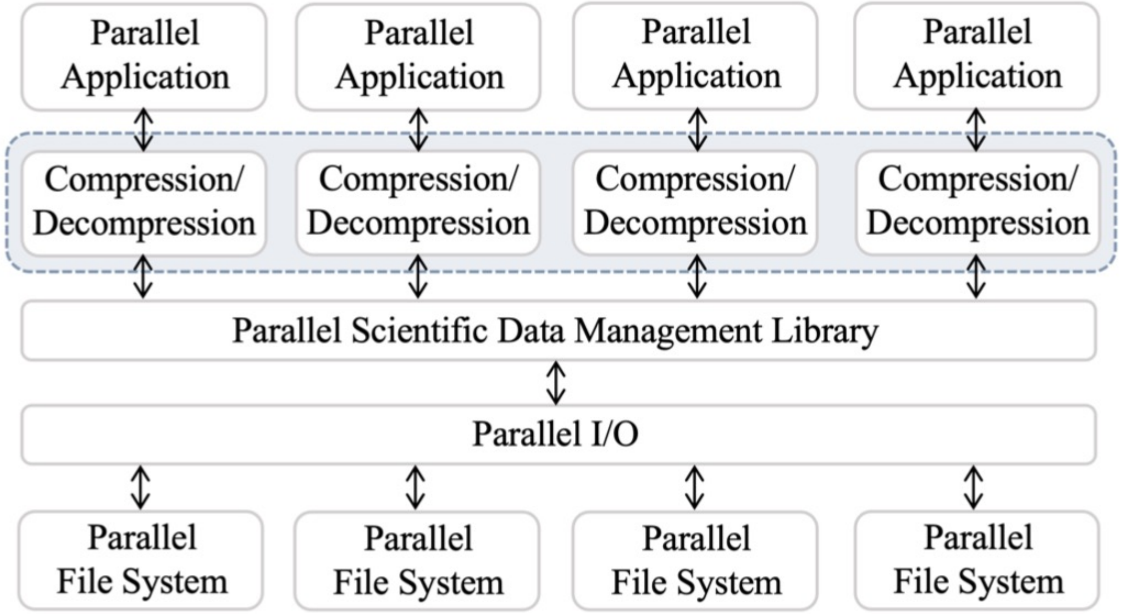
- HDF5, netCDF, and Adaptable IO System (ADIOS)
- Compression techniques are often adopted

Error Bounded Lossy Compressors

- Transform-based lossy compressor (ZFP)
- Prediction-based lossy compressor (SZ)
- Data distortion metrics
  - Peak signal-to-noise ratio (PSNR)
  - Structural similarity (SSIM)

Compression Mode

- Error bounded mode
  - Absolute error bound (ABS)
  - Relative error bound (REL)
  - Point-wise relative error bound (PW_REL)
- Fix rate mode



Scientific data management with compression

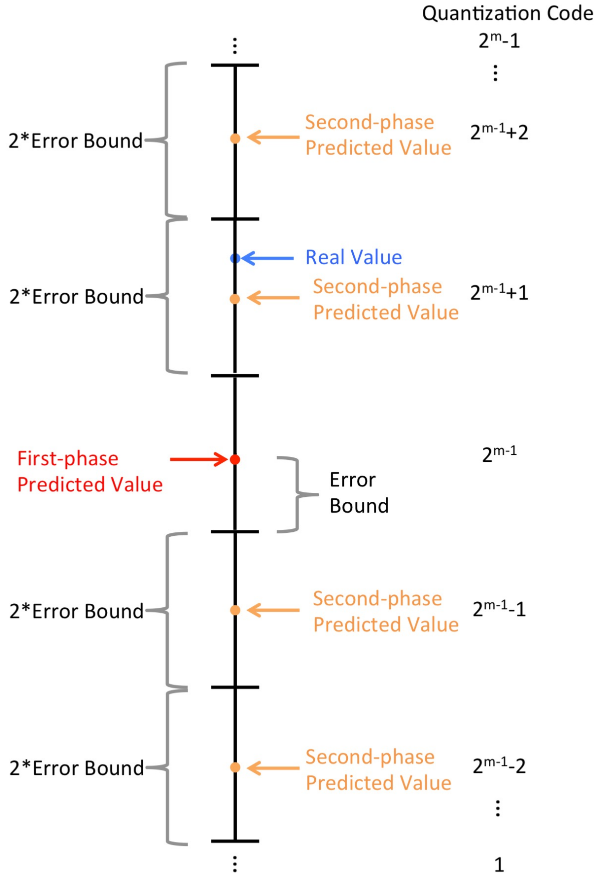
Background

Prediction-Based Lossy Compression

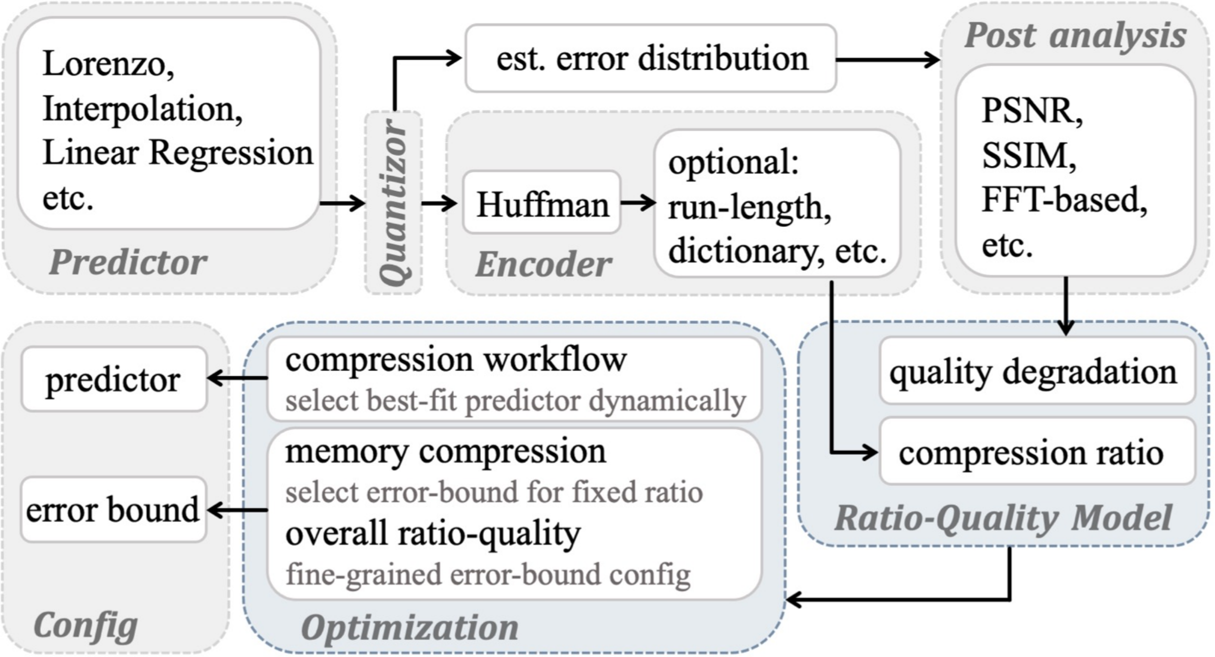
- Each data point's value is predicted based on its neighboring data points by an adaptive, best-fit prediction method
- Each floating-point weight value is converted to an integer number by a linear-scaling quantization based on the difference between the real value and predicted value and a specific error bound.
- Lossless compression is applied to reduce the data size thereafter

Main Challenges

- How to decompose prediction-based lossy compression into multiple stages and model the compression ratio for each stage?
- How to reduce the time cost of extracting data information needed by the model?
- How to model the quality degradation in terms of diverse post-analysis metrics?
- How does our model benefit real-world applications?



Rate-Quality Model



An overview of ratio-quality modeling workflow for prediction-based lossy compression and scientific data analysis

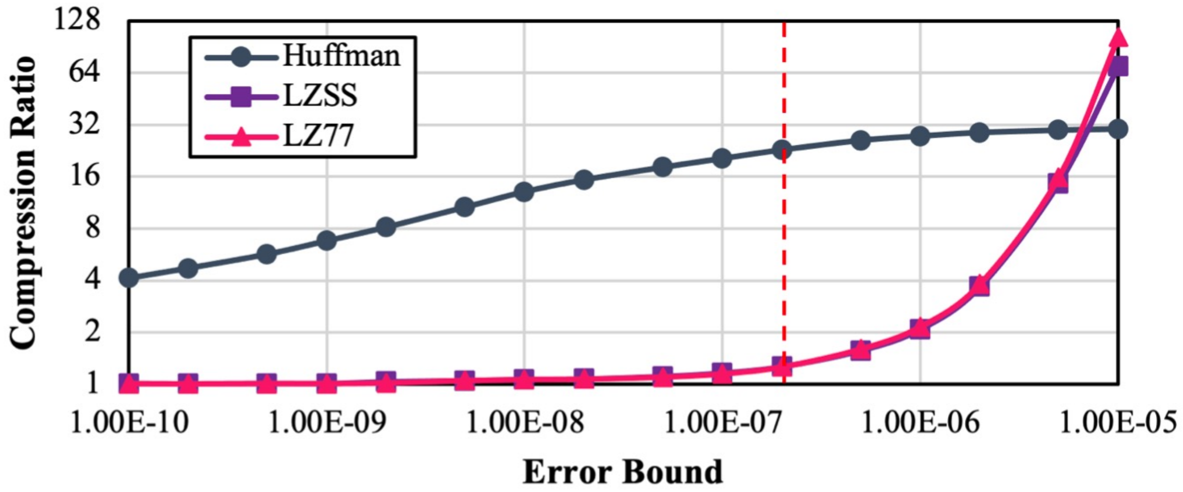
Overview

- Compression ratio
  - Predictor (prediction error histogram)
  - Quantizor (quantization code histogram)
  - Encoder (encode efficiency)
- Post-hoc analysis quality
  - Estimated error distribution

Analysis

- Model compression ratio of popular encoders
- Refine compression ratio modeling for various predictors and quantizers
- Model quality degradation for both generic and specific post-hoc analysis

Rate-Quality Model



Compression ratio from Huffman encoder and optional lossless encoder from Zstandard and Gzip on quantization code

Modeling Encoder Efficiency

- Quantization code is highly randomized
- Encoding efficiency provided by Huffman encoding is highly separated from that provided by the optional lossless encoders
- Zero would always dominate the Huffman codes after the red dashed line

Huffman Encoding

$$B = \sum_{i=0}^{n} P(s_i)L(s_i) \approx -\sum_{i=0}^{n} P(s_i)\log_2 P(s_i),$$

$$e^* = 2^{B-B^*}e,$$

Run-Length Encoding (After Huffman)

$$R_{rle} = 1/(C_1(1-p_0)P_0 + (1-P_0)).$$

$$p_0 = \sqrt{1 - R_{rle}^{-1} - ((C_1-1)/2)^2} + (C_1-1)/2$$

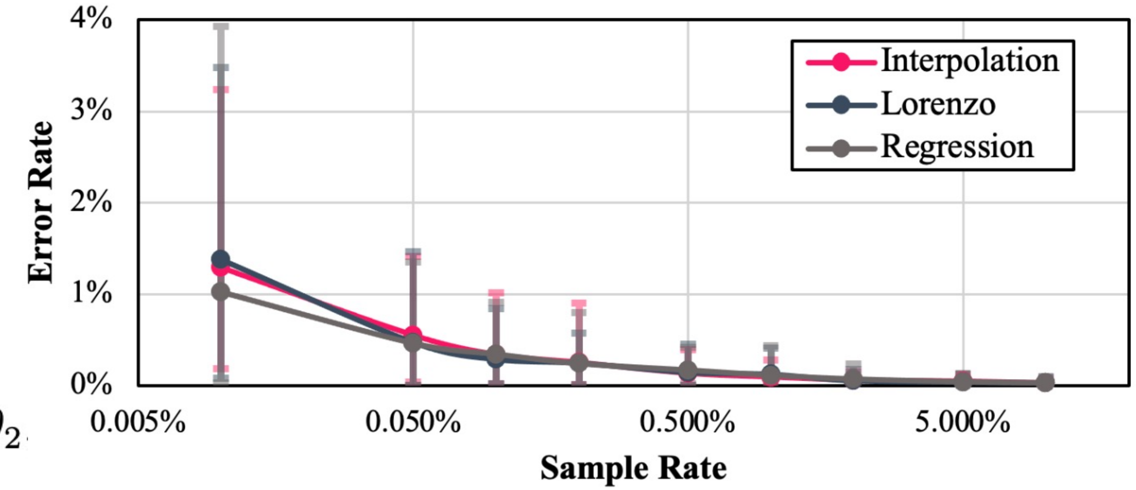
Rate-Quality Model

Modeling Quantized Prediction Error Histogram

- Prediction error histogram: different sampling solutions for different predictors
  - Lorenzo Predictor
  - Linear Interpolation Predictor
  - Linear Regression Predictor
- Quantization code histogram
  - Based on sampled prediction error
  - Large distortion under large error bounds
    - Bin transfer scheme

$$N_{tran} = P_{tran} \cdot N = C_2 \cdot (1 - p_0) \cdot N, \quad \text{when } p_0 \geq \theta_2$$

Original Value	[..., 0.0, 1.3]	
Quantization Code	[..., 0, 1]	[..., 0, 0]
	Ours	Actual



Error rate between sampled prediction error and original prediction error under different sampling rates with three predictors. The error bar indicates the max and min values

Rate-Quality Model

Post-hoc Analysis Quality Model

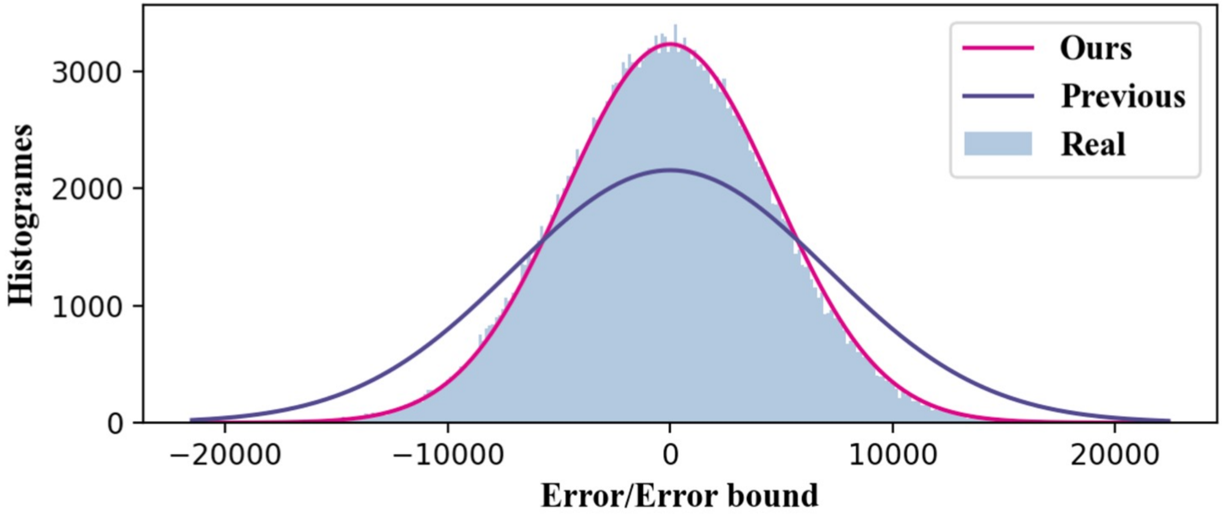
- Error distribution, described by its variance

$$\sigma(E)^2 = \sum_{i=0}^{N}(E[i]^2 - \mu^2) \approx \int_{-e}^{e} \frac{1}{2e} x^2 dx = \frac{1}{3}e^2$$

Unified distribution

$$\sigma(E)^2 = \sum_{i=0}^{(1-p_0)N}(E[i]^2 - \mu^2) + \sum_{i=0}^{p_0 N}(E[i]^2 - \mu^2)$$
$$= (1-p_0)\frac{1}{3}e^2 + p_0\sigma(B[0]),$$

Refined centralized distribution at high error bounds



FFT quality degradation estimation compared to measurement. Evaluated on Nyx temperature field at ABS 500

- Peak signal-to-noise ratio (PSNR)
- Structural similarity index (SSIM)
- Data-specific post-hoc analysis

$$PSNR(D', D) = 20\log_{10}(minmax) - 10\log_{10}(\sigma(E)^2)$$

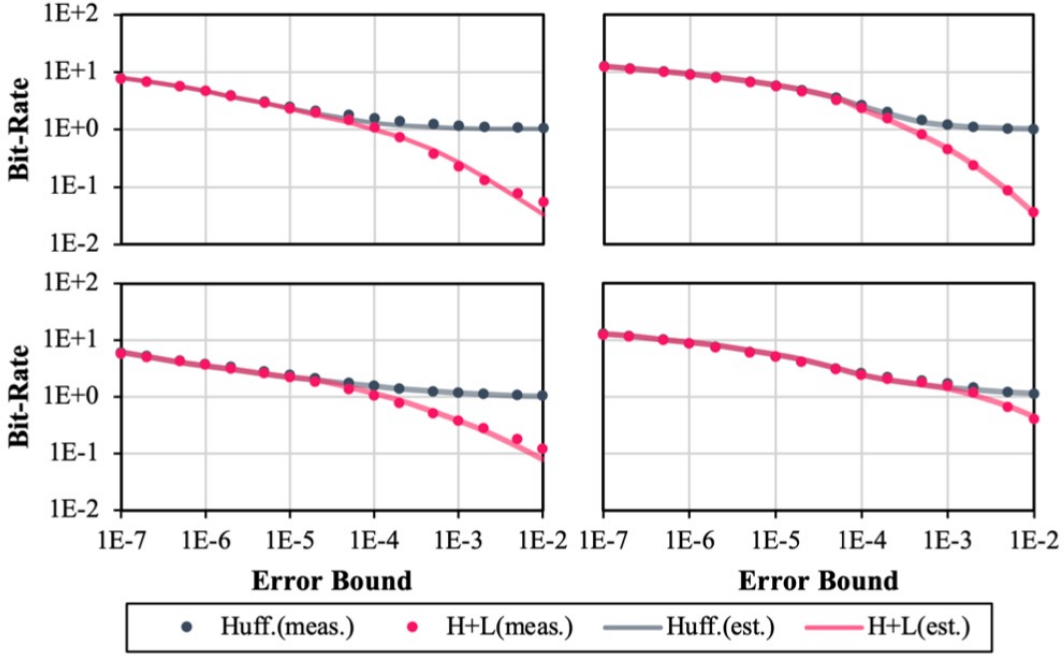
$$SSIM(D', D) = \frac{2\sigma_D^2 + C_3}{2\sigma_D^2 + C_3 + \sigma(E)^2}$$

Evaluation

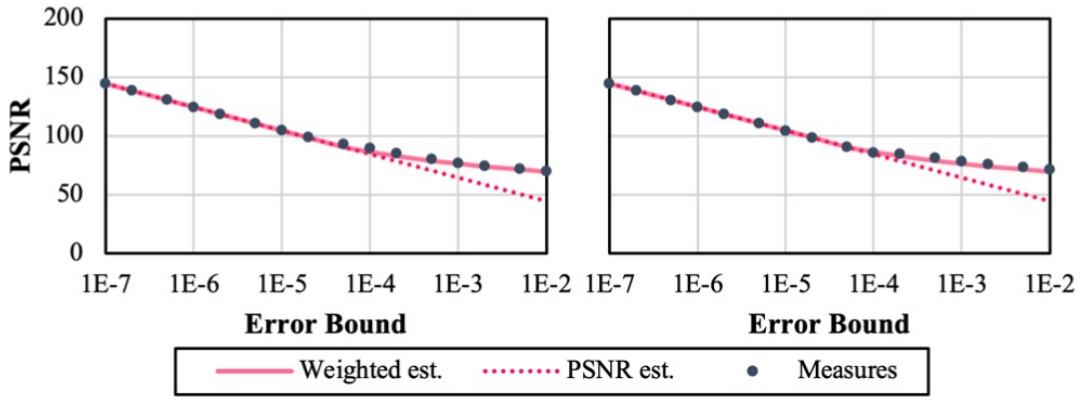
WASHINGTON STATE UNIVERSITY

Ratio-Quality Model Accuracy

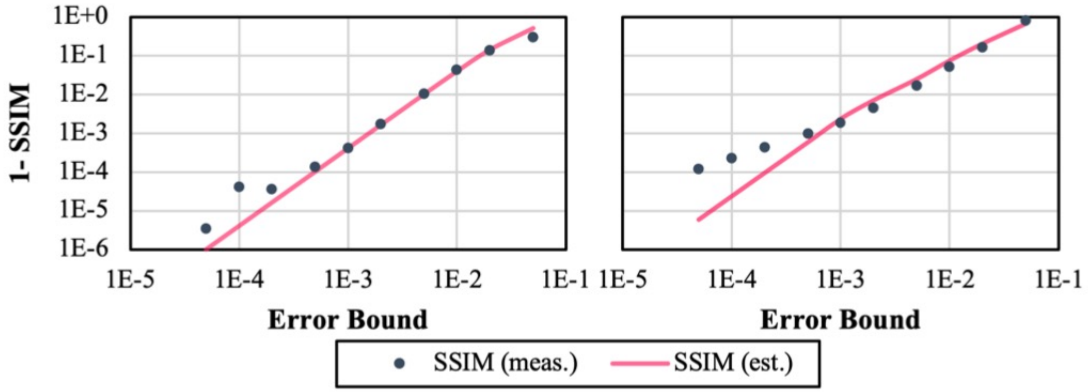
- Accuracy of Compression Ratio Model
- Accuracy of Post-Hoc Analysis Quality Model



Compression ratio (bit-rate) estimation accuracy compared to measurement by the encoders



PSNR estimation accuracy compared to measurement



SSIM estimation accuracy compared to measurement

Evaluation

Name	Field	Dim	Sample Err.	Huff Err.	Lossless Err.	Huff+LL. Err.	PSNR Err.	SSIM Err.
RTM	1000	235x449x449	0.03%	5.67%	**9.82%**	8.72%	0.77%	**9.34%**
	2000	235x449x449	0.02%	3.32%	**9.01%**	7.76%	1.56%	**6.56%**
	3000	235x449x449	0.06%	1.88%	**9.15%**	7.57%	2.84%	**4.12%**
CESM	TS	1800x3600	0.06%	6.88%	**11.26%**	8.85%	**3.97%**	2.54%
	TROP_Z	1800x3600	0.20%	7.56%	**10.52%**	9.66%	2.97%	**4.44%**
Hurricane	U	100x500x500	0.10%	**4.62%**	3.46%	5.75%	1.56%	**5.43%**
	TC	100x500x500	0.12%	**5.44%**	2.96%	5.95%	2.42%	**3.80%**
Nyx	Dark Matter	512x512x512	0.14%	**7.53%**	4.36%	7.67%	1.78%	**6.55%**
	Temperature	512x512x512	0.13%	3.92%	**5.13%**	3.99%	1.89%	**4.34%**
	Velosity Z	512x512x512	0.07%	6.85%	**8.65%**	8.08%	2.64%	**3.90%**
HACC	xx	280953867	0.26%	**2.29%**	1.34%	3.22%	1.98%	-
	vx	280953867	0.27%	**3.71%**	1.49%	3.83%	3.67%	-
Brown	Pressure	8388609	0.11%	**5.99%**	5.68%	6.46%	4.42%	-
Miranda	vx	256x384x384	0.13%	**7.90%**	6.95%	8.71%	2.55%	**8.92%**
QMCPACK	einspine	69x69x115	0.13%	6.84%	**8.83%**	6.20%	5.67%	**7.43%**
SCALE	PRES	98x1200x1200	0.16%	1.65%	**2.79%**	2.36%	1.72%	**5.35%**
EXAFEL	raw	10x32x185x388	0.12%	**5.64%**	4.25%	6.23%	3.80%	-
Average	-	-	0.12%	5.16%	**6.21%**	6.53%	2.72%	**5.59%**

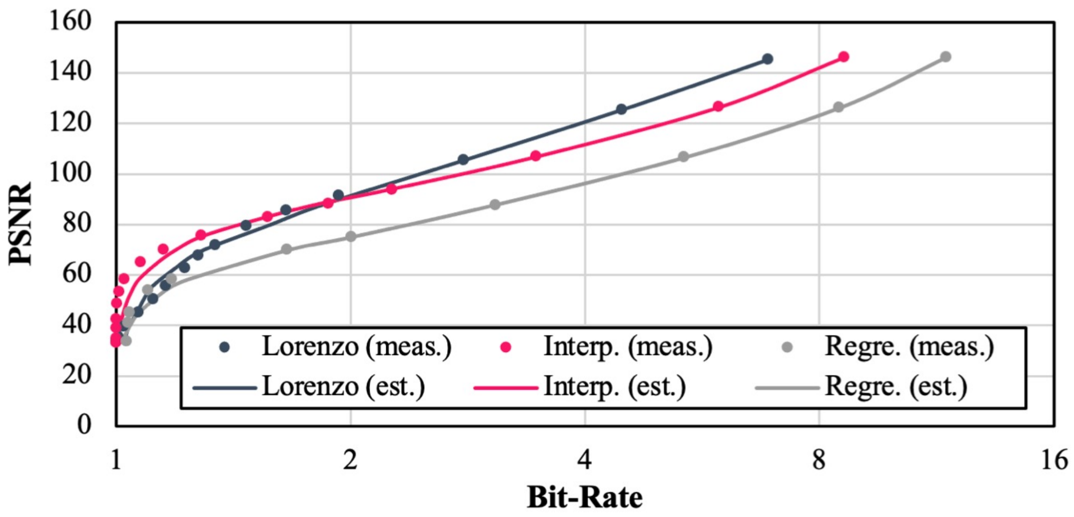
* Bold items highlight the larger prediction error between the two encoders and between the two post analyses

Details of Evaluation Results on Tested Data and Fields

Evaluation

Use Cases

- Predictor Selection
  - Select the most efficient predictor for a given dataset & error bound
- Memory Limitation Control
  - Efficiently utilize available memory
- In-Situ Compression Optimization
  - optimize the compression performance individually for each partition with overall compression ratio and overall analysis quality as objectives



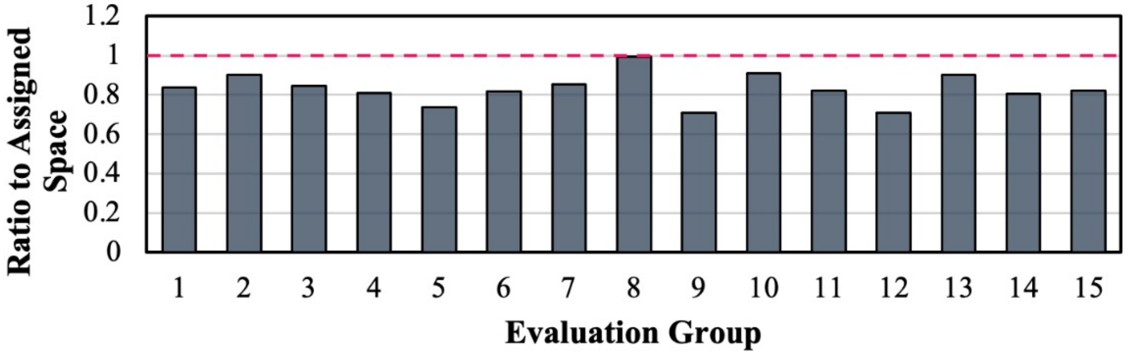
Rate-distortion curve of multiple predictors with different error bound. Evaluated with RTM dataset

Evaluation

WASHINGTON STATE UNIVERSITY

Use Cases

- Predictor Selection
  - Select the most efficient predictor for a given dataset & error bound
- Memory Limitation Control
  - Efficiently utilize available memory
- In-Situ Compression Optimization
  - optimize the compression performance individually for each partition with overall compression ratio and overall analysis quality as objectives

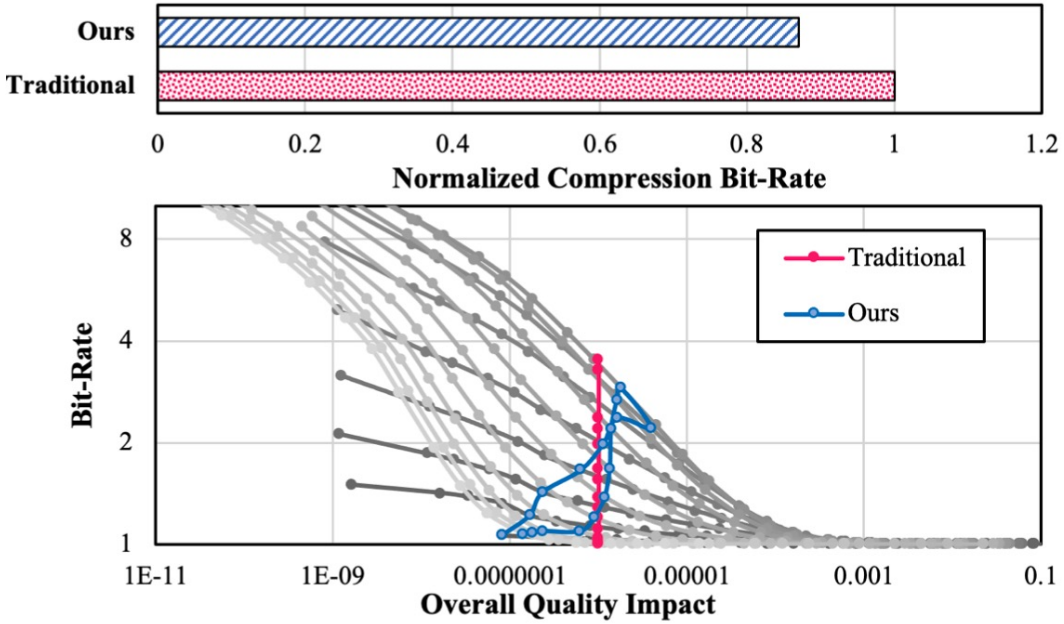


Ratio of measured space consumption to assigned space. Evaluated with RTM dataset, randomly choose time steps and error bound for 15 groups

Evaluation

Use Cases

- Predictor Selection
  - Select the most efficient predictor for a given dataset & error bound
- Memory Limitation Control
  - Efficiently utilize available memory
- In-Situ Compression Optimization
  - optimize the compression performance individually for each partition with overall compression ratio and overall analysis quality as objectives

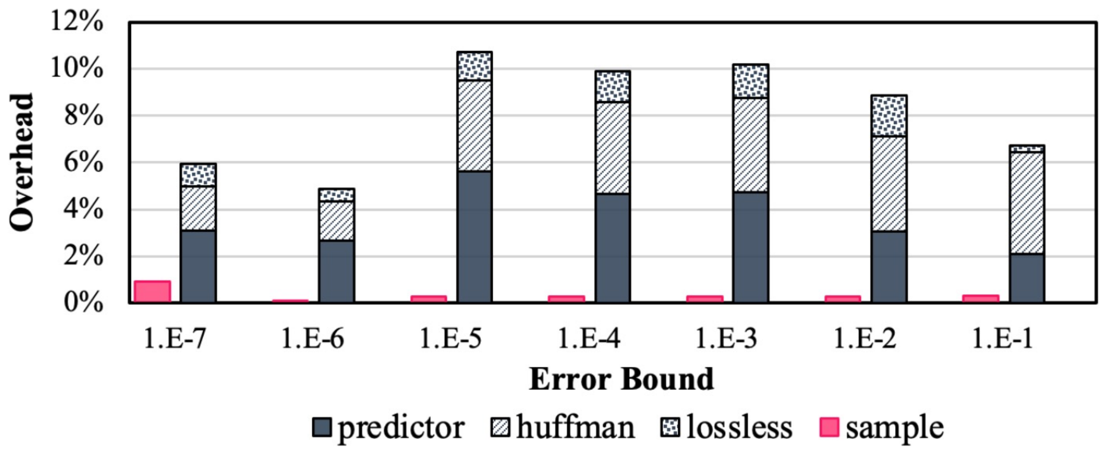


Error bound optimization for RTM dataset with multiple time steps in consideration for post-hoc analysis

Evaluation

Performance

- Significantly lower overhead compared to previous solution
- One sampling, prediction on all error bound setting
- Outperforms the trial-and-error solution by 18.7× on average when considering 7 candidate error bounds to estimate with the Lorenzo and interpolation predictors as candidates

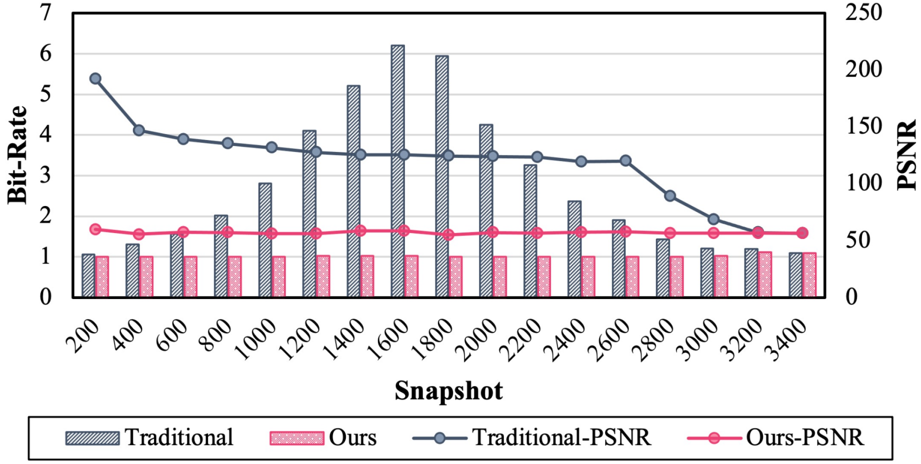


Performance comparison between proposed modeling solution and previous trial-and-error approach

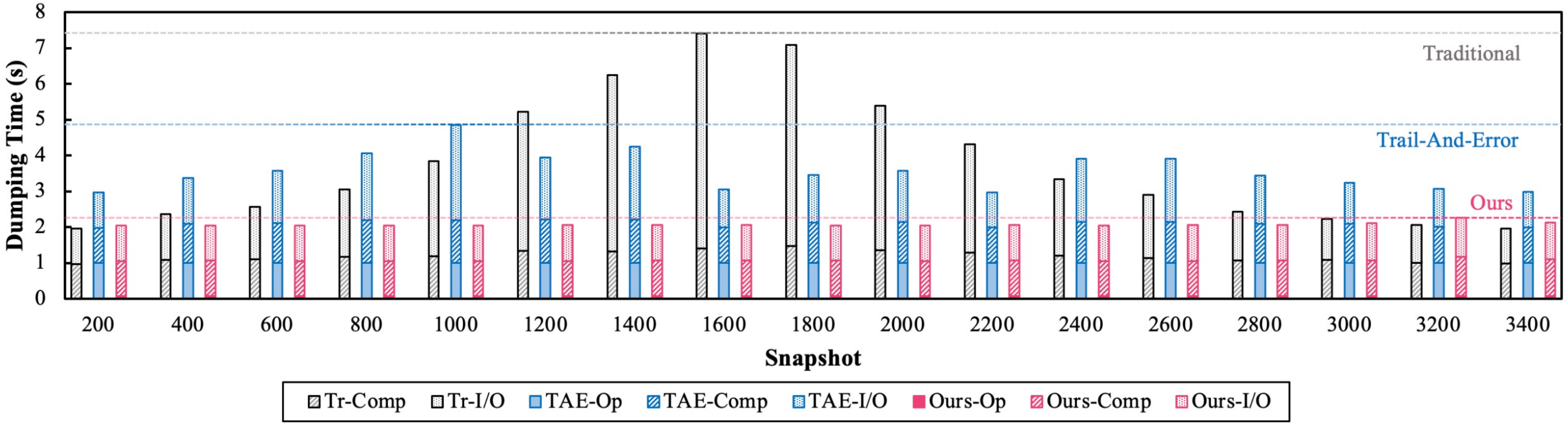
Evaluation

Overall Performance of Data Management

- Optimize the most efficient compression configuration for each snapshot
- Provide consistent and the fastest data dumping time



Comparison between our modeling-based method with offline optimization method in terms of both bit-rate and corresponding PSNR across different snapshots when target PSNR is 56 dB.



Overall data dumping performance with parallel HDF5. Comparison between traditional method, trial-and-error and our modeling-based method. Dashed lines highlight the maximum dumping time occurred in the simulation. “Tr” refers to the traditional approach, “TAE” refers to the in-situ trial-and-error approach. ‘Comp’, ‘I/O’, and ‘Op’ refer to times of compression, I/O, and optimization, respectively

Thank you!

Any questions are welcome!

Contact Dingwen Tao: dingwen.tao@wsu.edu
Sian Jin: sian.jin@wsu.edu