A Feature-Driven Fixed-Ratio Lossy Compression Framework for Real-World Scientific Datasets

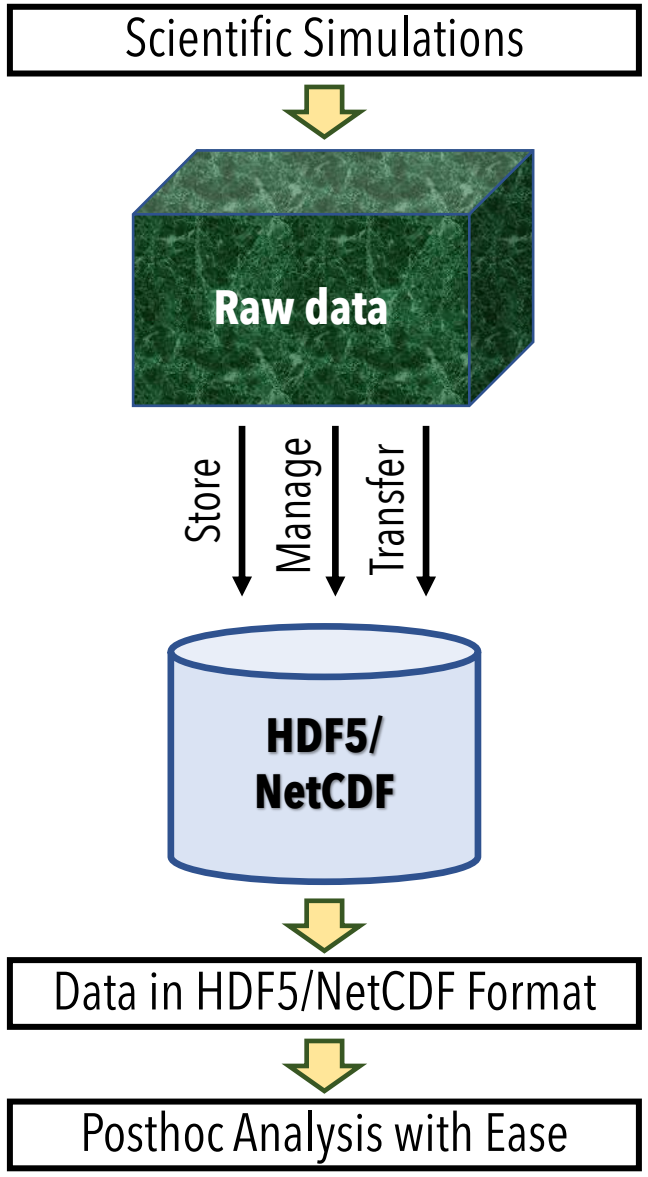
Md Hasanur Rahman★, Sheng Di†, Kai Zhao♂, Robert Underwood†, Guanpeng Li★, Franck Cappello†

★ *Computer Science Department, University of Iowa, IA, USA*

† *Mathematics and Computer Science Division, Argonne National Laboratory, IL, USA*

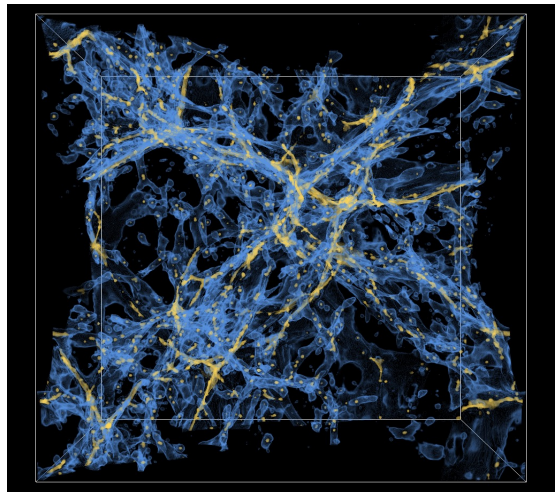
♂ *Computer Science Department, University of Alabama at Birmingham, AL, USA*

Introduction



Challenges in scientific data management

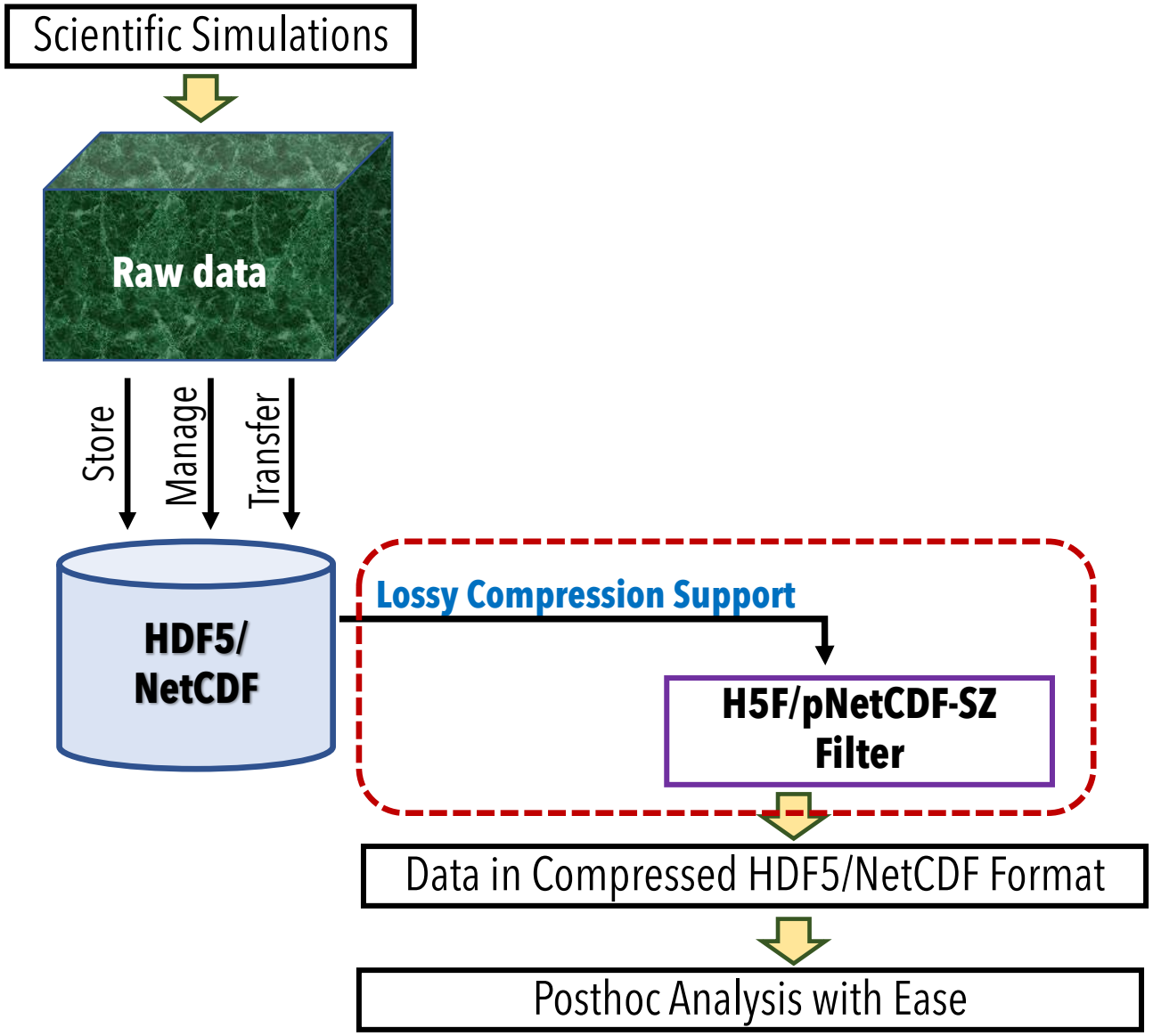
- Extremely large volume of data per simulation
- Nyx[1] can produce upto hundreds of perabytes of data per simulation
- Threat to limited available memory capacity, storage, space and I/O bandwidth



Cosmology Simulation: Nyx[2]

[1] Almgren et al, "Nyx: A MASSIVELY PARALLEL AMR CODE FOR COMPUTATIONAL COSMOLOGY"
[2] https://deixismagazine.org/2013/12/rewinding-the-universe

Introduction



Integration of lossy compressor support

- Error bounded lossy compression (e.g., SZ[18], ZFP[12])
- High compression ratio with high data fidelity

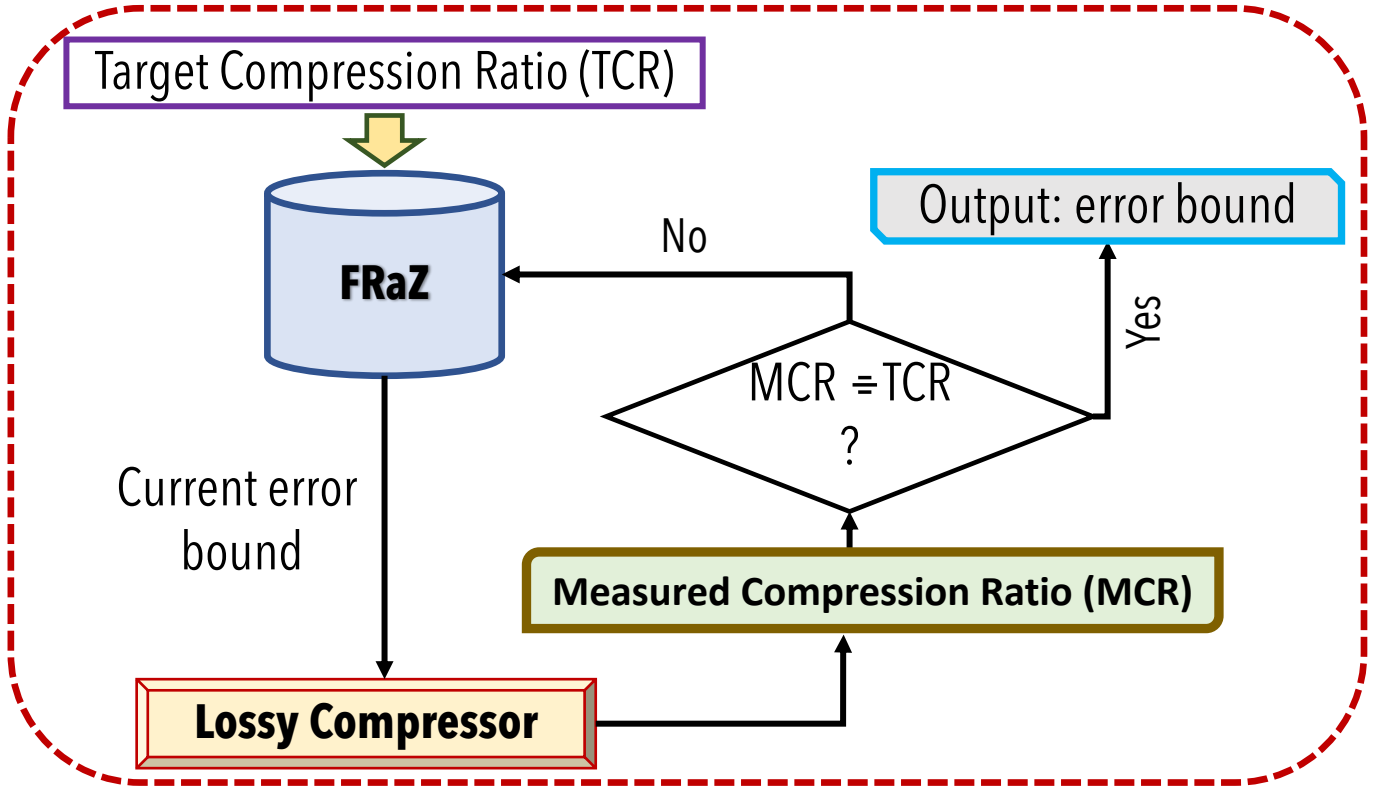
Research gap

- X Obtaining target compression ratio requires iterative executions under many error bounds
- In advance knowledge before compression
- Better utilization of limited resources
- An efficient fixed-ratio mechanism is needed
- X No lossy compressor supports fixed-ratio mode

Introduction

Existing fixed-ratio framework

- **FRaZ** [20]: a generic fixed-ratio framework
- X High computational cost
- X Need to run underlying compressor at each iteration
- X Requires trail-and-error-based iterative runs



Our Solution: FXRZ

- An efficient fixed-ratio compressor-agnostic framework
- Efficient compressor-agnostic fixed-ratio framework is non-trivial
- Exploit data characteristics such as smoothness, textures, distribution
- No need to run the underlying compressor

Research Background

- Compression ratio: ratio of original data size to the compressed one
- Data distortion: quality of reconstructed data w.r.t original data
- Error bound: Maximum error between reconstructed and original data

FXRZ Design Overview

Data features extraction

- Data smoothness, spreadness, special patten

Optimization strategies

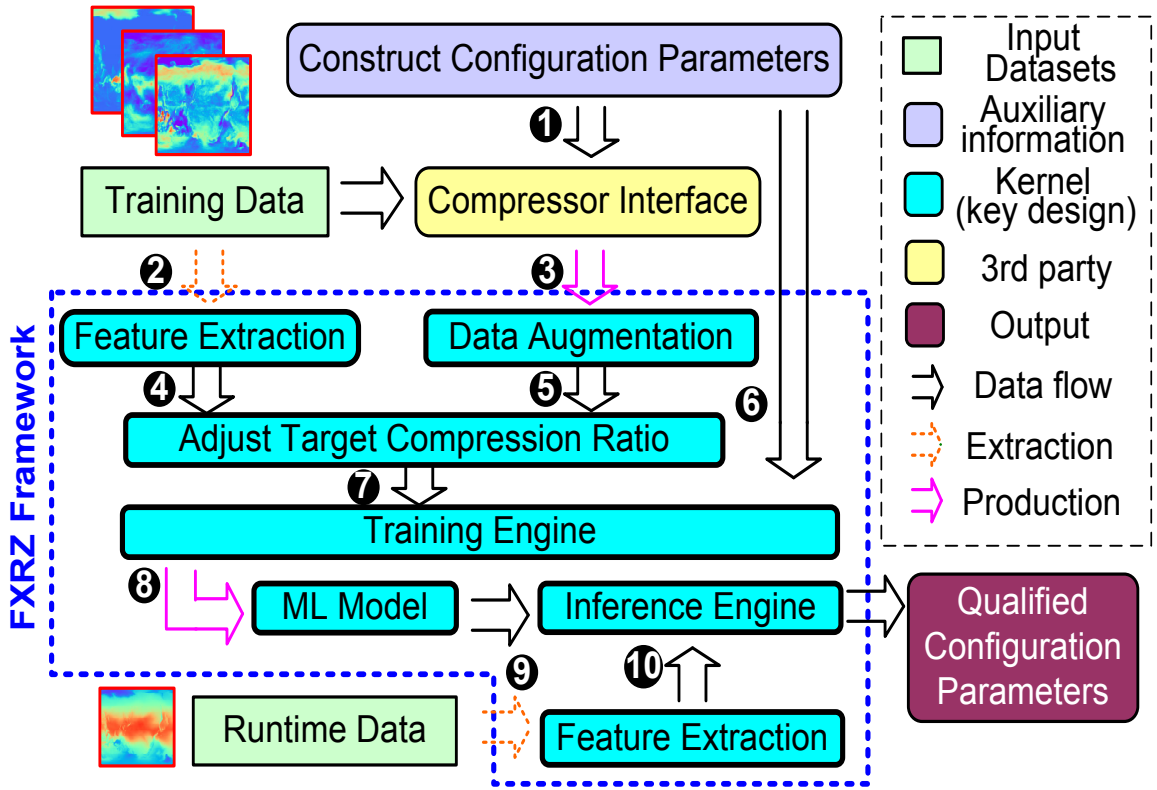
- 4-point stripe uniform sampling for efficiency
- Adjust target compression ratio for better accuracy

ML model

- Random forest regressor

Data augmentation

- Why? To generate ample training samples without expensive compressor run
- How? Augment a few measured compression results based on interpolation to obtain ample results



FXRZ: Data Features Extraction

Value range

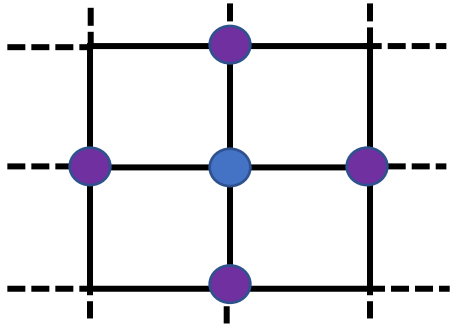
- (Max value – Min value) from sample data

Mean value

- Average of sampled data values

Mean neighbor difference (MND)

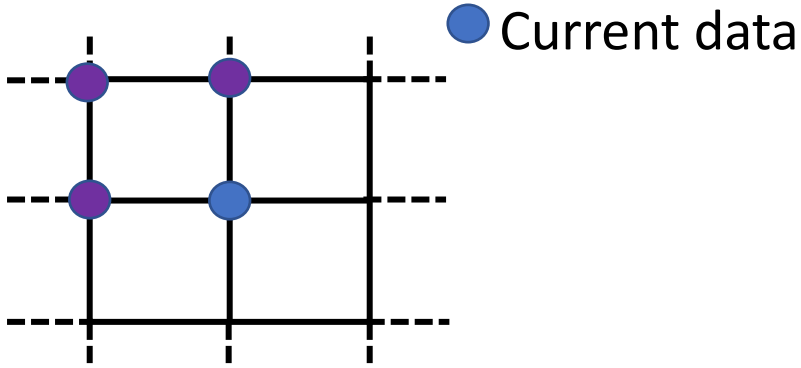
2D Case:



Mean Lorenzo difference (MLD)

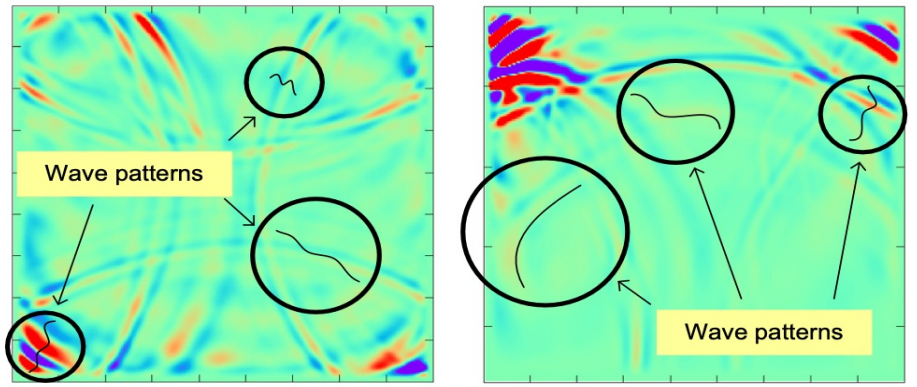
$$lorenzo_{i,j} = d_{i-1,j} + d_{i,j-1} - d_{i-1,j-1}$$

2D Case:



Mean spline difference (MSD)

[3] $$spline_i = -\frac{1}{16}d_{i-3} + \frac{9}{16}d_{i-1} + \frac{9}{16}d_{i+1} - \frac{1}{16}d_{i+3}$$



(a) Cross Section Visualization (b) Longitudinal Section Visualization

[3] Zhao et al, "Optimizing error-bounded lossy compression for scientific data by dynamic spline interpolation"

FXRZ: Analysis of Data Features

Data characteristics by data features

- Value Range and Mean Value reveal the data amplitude and spreadness
- MND and MLD reveal the spatial data smoothness
- MSD feature is particularly effective in detecting the wave textures/patterns

Feature	Nyx Baryon Density	QMCPack BigScale	RTM SmallScale	RTM BigScale	Hurricane TC
Value Range	4.90	35.36	0.16	0.05	104.81
Mean Value	0.97	16.75	0.09	0.02	45.63
MND	0.01	0.29	1.1E-4	5.5E-5	0.67
MLD	0.31	0.30	9.2E-5	4.0E-5	31.30
MSD	8.4E-3	0.33	1.3E-4	6.1E-5	0.79

Table: Feature values across datasets

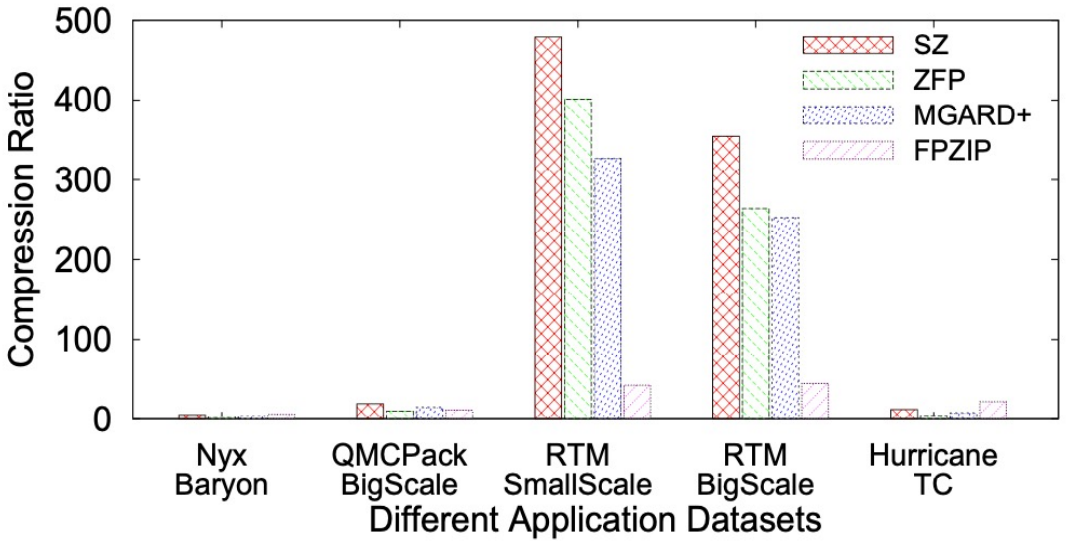


Figure: Compression ratios with datasets and compressors under same error bound

FXRZ: Optimization Strategies

Observation

- overall compression ratio is very sensitive to smooth region
- These regions lead to overestimation of true compressibility

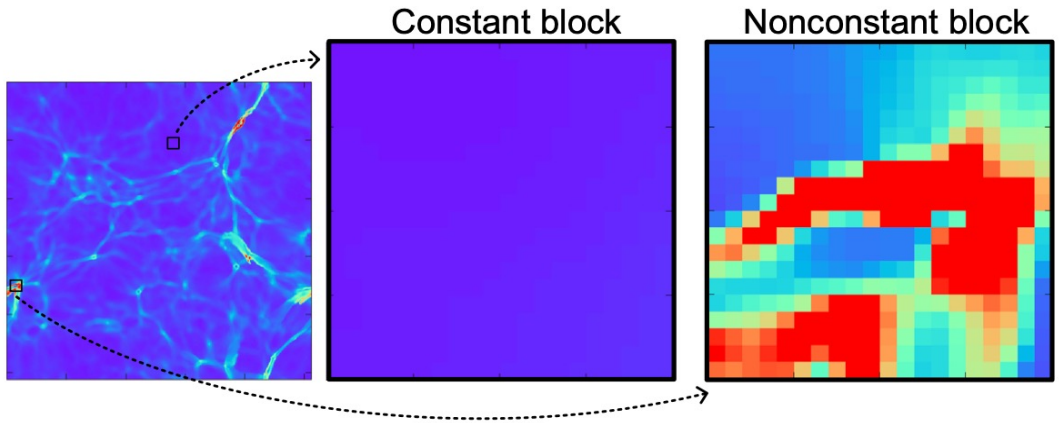


Fig.: Illustration of constant and non-constant blocks

Solution: Adjusting TCR for better accuracy

- Compressibility Adjustment (CA): Need to adjust TCR based on density of data
- Split the dataset into small blocks, e.g., 4x4 for 2D
- Constant and non-constant block: based on value range in block
- 15% of mean value is used as value range threshold
- Formula: ACR = TCR * R

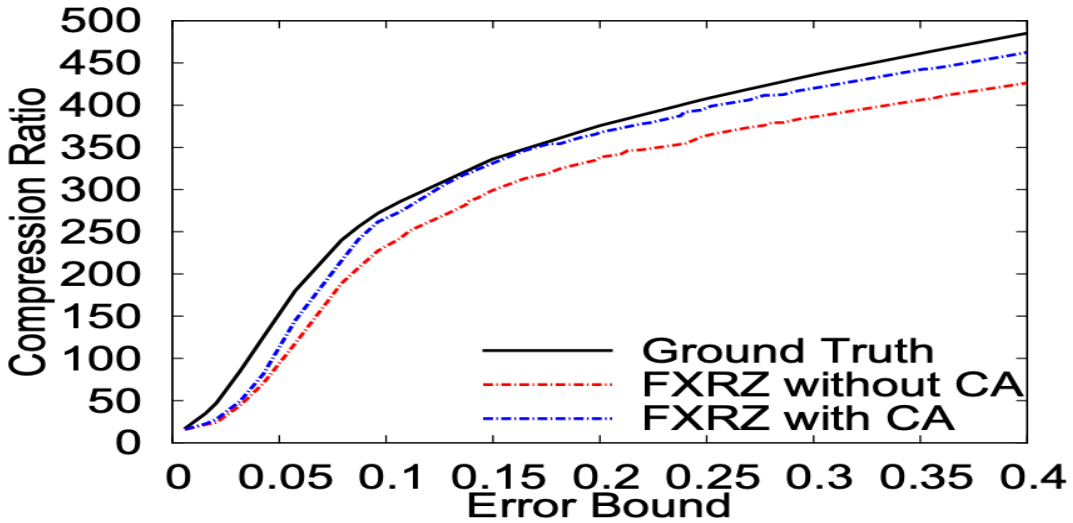


Fig.: CA optimization of Nyx Baryon Density with SZ

Experiment Setup

Datasets

- Frequently used in lossy compression studies [45-48]

App.	# Fields	TSteps	Dim	Size	Domain
Nyx-1	4	6	512×512×512	12.00GB	Cosmology
Nyx-2	4	1	512×512×512	2.00GB	Cosmology
QMCPack-1	1	1	288×115×69×69	0.59GB	Quantum Structure
QMCPack-2	2	1	480×115×69×69	1.96GB	Quantum Structure
QMCPack-3	2	1	816×115×69×69	3.33GB	Quantum Structure
RTM-Small	1	7	449×449×235	1.24GB	Seismic Wave
RTM-Big	1	2	849×849×235	1.26GB	Seismic Wave
Hurricane	2	7	100×500×500	1.30GB	Weather

Table: Datasets

Assessment levels of FXRZ capability

- Capability level 1: across different time steps of same application model
- Capability level 2: across different simulation configurations within application model

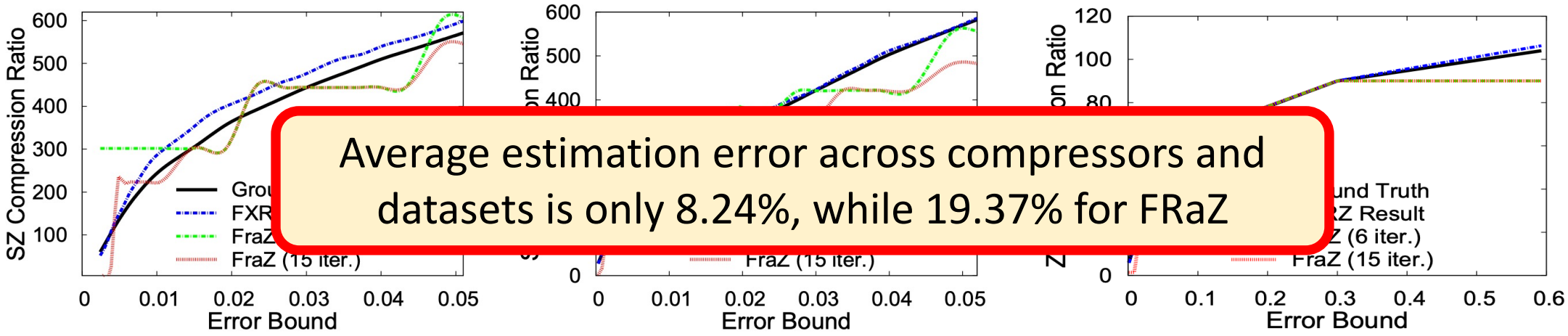
Baseline: FRaZ [20]

- Evaluate FRaZ under two max-iterations: 6 and 15

Estimation error

- |TCR - MCR| / TCR

Evaluation: Accuracy (Capability 1 and 2)



Average estimation error across compressors and datasets is only 8.24%, while 19.37% for FRaZ

(a) Nyx Velocity-X (SZ) (b) QMCPack-3 spin0 (SZ) (e) Nyx Velocity-X (ZFP)

Figure: Estimation error comparison with different testing fields and compressors

Evaluation: Efficiency W.R.T Compression Time

Comparison between FXRZ and FRaZ with 15 iterations

- Average analysis time based on 25~30 uniformly selected TCRs

App.	Test Fields/ Snapshots	SZ		ZFP		MGARD+		FPZIP	
		FXRZ	FRaZ	**FXRZ**	FRaZ	**FXRZ**	FRaZ	**FXRZ**	FRaZ
Nyx	B							0.08x	17.31x
	D							0.06x	10.75x
	T							0.11x	25.36x
	V							0.11x	20.84x
QMCPack	BigScale Spin0	0.08x	7.72x	0.07x	7.69x	0.068x	11.42x	0.08x	22.01x
	BigScale Spin1	0.08x	7.78x	0.07x	5.42x	0.07x	11.51x	0.08x	17.12x
RTM	BigScale Snapshot-750	0.09x	8.10x	0.68x	32.72x	0.12x	21.29x	0.13x	16.85x
	BigScale Snapshot-800	0.09x	7.23x	0.71x	29.41x	0.13x	19.44x	0.13x	15.79x
Hurricane	QCLOUD Snapshot-48	0.20x	9.14x	0.47x	15.57x	0.22x	14.09x	0.22x	14.78x
	TC Snapshot-48	0.17x	5.93x	0.19x	14.84x	0.18x	21.20x	0.18x	25.72x
Average	Across All Domains	0.10x	6.85x	0.24x	13.19x	0.11x	17.03x	0.12x	18.65x

FRaZ can be 108x slower than FXRZ for error bound estimation, on average

Table: Average analysis time cost with respect to compression time

Conclusion & Acknowledgement

- We propose FXRZ – a feature-driven compressor-agnostic framework
- Goal: efficient estimation of appropriate error bound setting based on given TCR
- Evaluate FXRZ using 4 lossy compressors with 10-real world scientific datasets
- The average estimation error is only 8.24%
- FXRZ's online analysis time is only 0.14x compared to compression time
- FXRZ is 108x faster than FRaZ

Md Hasanur Rahman
University of Iowa
Email: mdhasanur-rahman@uiowa.edu
Webpage: https://hasanur-rahman.github.io