Adaptive Configuration of In Situ Lossy Compression for Cosmology Simulations via Fine-Grained Rate-Quality Modeling

Sian Jin (Washington State University)
Jesus Pulido (Los Alamos National Laboratory)
Pascal Grosset (Los Alamos National Laboratory)
Jiannan Tian (Washington State University)
Dingwen Tao (Washington State University)
James Ahrens (Los Alamos National Laboratory)

2021

Outline

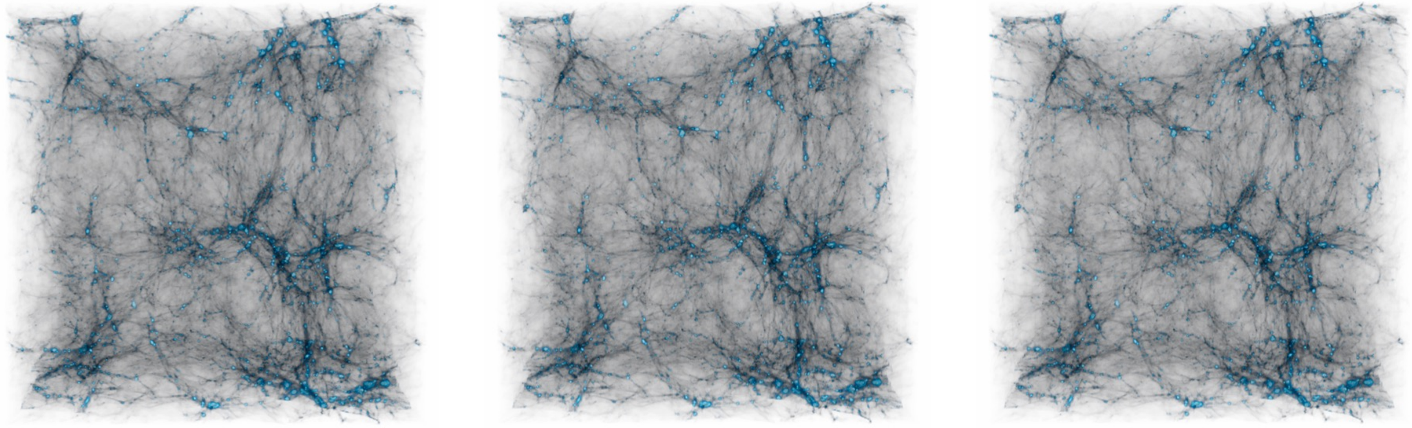
- **Introduction**
  - Fine-grained lossy compression
  - Rate-quality modeling
- **Background**
  - Cosmology post hoc analysis
  - Lossy compression for scientific data
- **Designs**
  - Modeling error impact on post hoc analysis
  - Modeling compression ratio
  - Optimization strategy
- **Experimental Evaluation**
  - Rate-quality evaluation
  - Performance evaluation

Outline

- **Introduction**
  - Fine-grained lossy compression
  - Rate-quality modeling
- **Background**
  - Cosmology post hoc analysis
  - Lossy compression for scientific data
- **Designs**
  - Modeling error impact on post hoc analysis
  - Modeling compression ratio
  - Optimization strategy
- **Experimental Evaluation**
  - Rate-quality evaluation
  - Performance evaluation

Introduction

- **Why Compress/Lossy Compression?**
  - Huge amounts of data from cosmological simulations.
    - Write speed.
    - Data storage.
  - Much higher compression ratio compared to Lossless Compression.

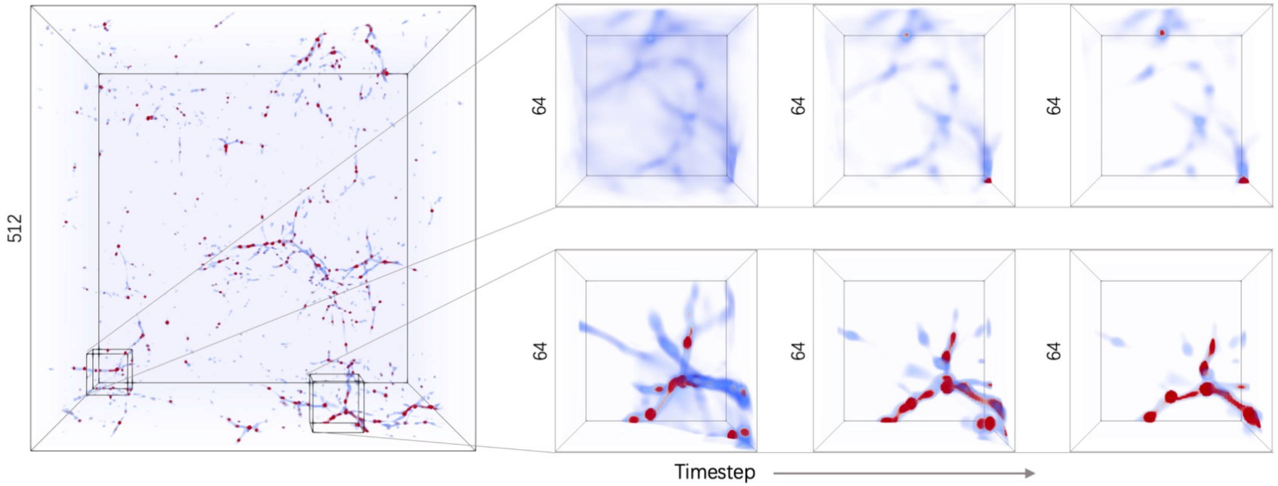


(a) Original (b) Reconstructed with PW_REL = 0.1 (c) Reconstructed with PW_REL = 0.25

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

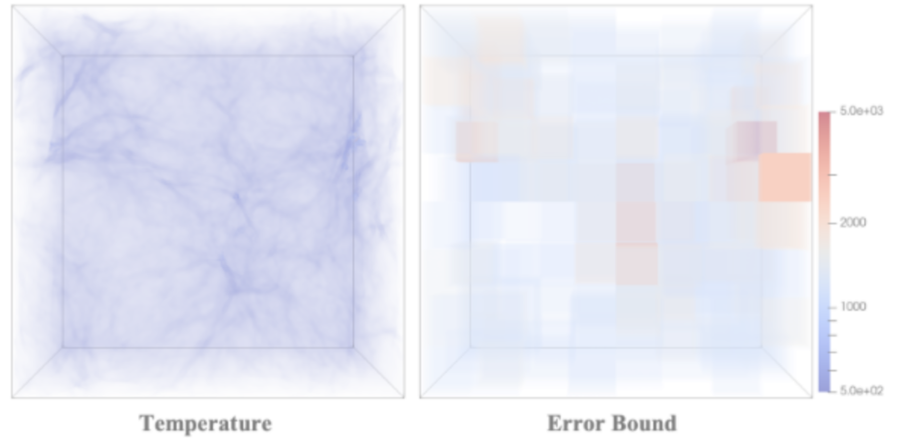
- **Structured Data**
  - Generated by mesh-based simulations in ranks
  - Different ranks (data partitions) have different **densities of information**



Visualization of Baryon Density in Nyx simulation under resolution of 512 × 512 × 512

- **Previous Solution** (Jin et al., IPDPS'20)
  - Optimize rate-quality by **trail-and-error**
  - All partitions use same compression configuration
  - Visual metrics (e.g., PSNR) are not sufficient

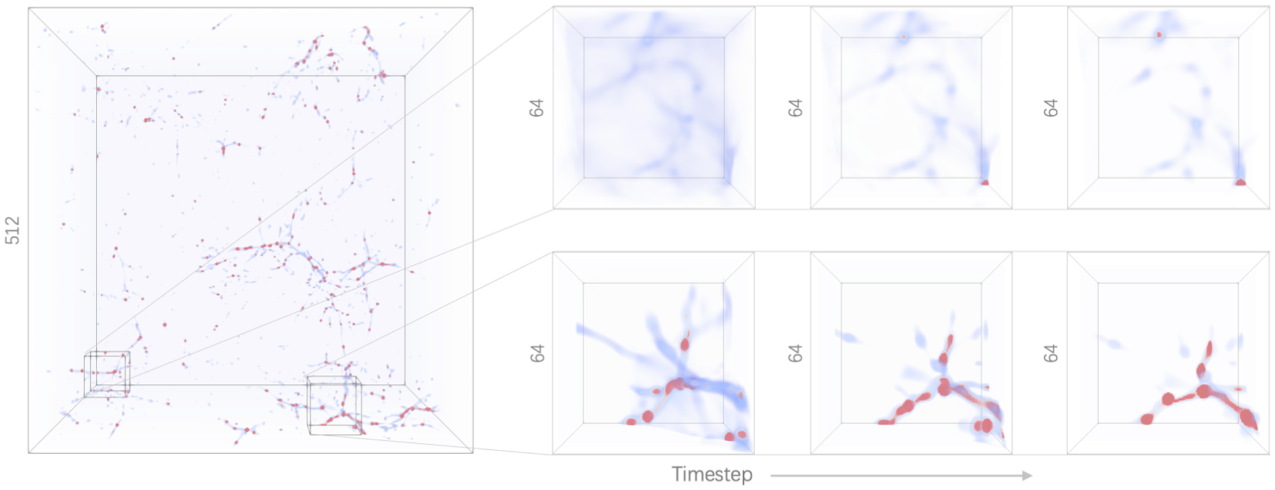
- **Our Goals**
  - Guarantee domain-specific analysis quality
  - In-situ compression
  - Towards optimal compression ratio



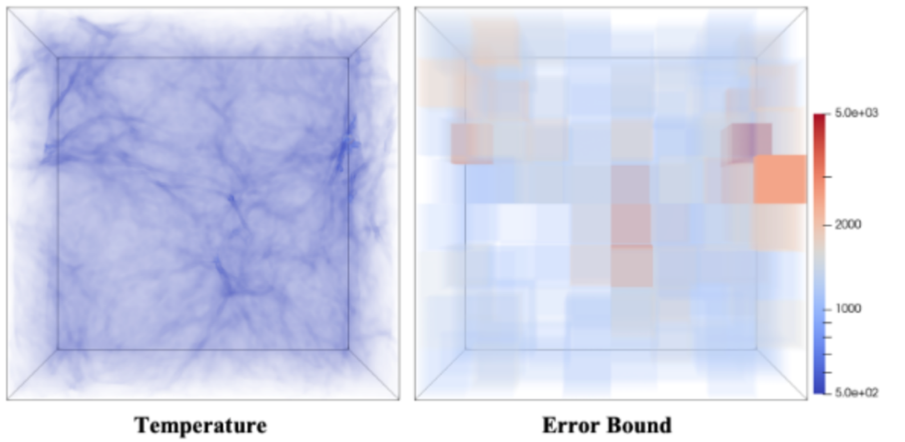
Fine-grained lossy compression control for different data partitio

Introduction

- **Structured Data**
  - Generated by mesh-based simulations in ranks
  - Different ranks (data partitions) have different **densities of information**
- **Previous Solution** (Jin et al., IPDPS'20)
  - Optimize rate-quality by **trail-and-error**
  - All partitions use same compression configuration
  - Visual metrics (e.g., PSNR) are not sufficient
- **Our Goals**
  - Guarantee domain-specific analysis quality
  - In-situ compression
  - Towards optimal compression ratio



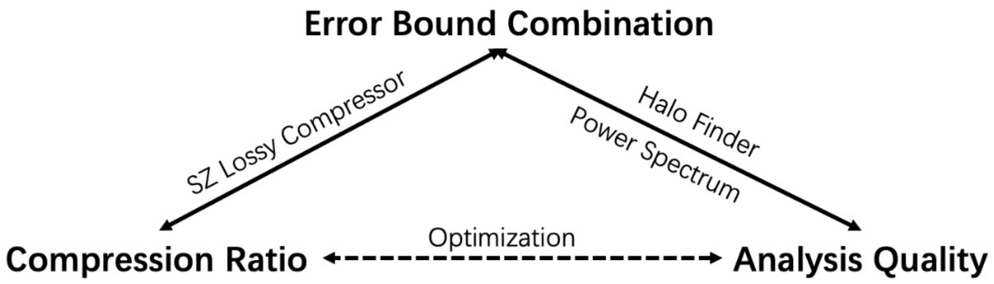
Visualization of Baryon Density in Nyx simulation under resolution of 512 × 512 × 512



Fine-grained lossy compression control for different data partitic

Our Approach

- **Fine-grained Compression**
  - Different configurations (error bound) for different partitions
  - Different configurations (eb combination) for different time-steps
- **Estimation on Analysis Quality (Loss)**
  - Predict post-hoc analysis error based on error-bound combination
    - Power spectrum (FFT based)
    - Halo finder
- **Estimation on Compression Ratio**
  - Predict compression ratio based on error-bound combination
    - SZ lossy compression



Our Approach

- **Contributions**
  - Propose an **adaptive approach** to select **feasible error-bound combinations** for **different partitions**
  - Build **theoretical models** to efficiently estimate:
    - **Loss of post hoc post-analysis** caused by lossy compression error
    - **Compression ratio** based on each partition's data feature
  - Develop an efficient **optimization strategy** to determine the best-fit configuration
  - Improve the compression ratio by up to **73%** (with only **1% performance overhead**)

Outline

- **Introduction**
  - Fine-grained lossy compression
  - Rate-quality modeling
- **Background**
  - Cosmology post hoc analysis
  - Lossy compression for scientific data
- **Designs**
  - Modeling error impact on post hoc analysis
  - Modeling compression ratio
  - Optimization strategy
- **Experimental Evaluation**
  - Rate-quality evaluation
  - Performance evaluation

Background

➢ **Nyx Cosmological Simulation**

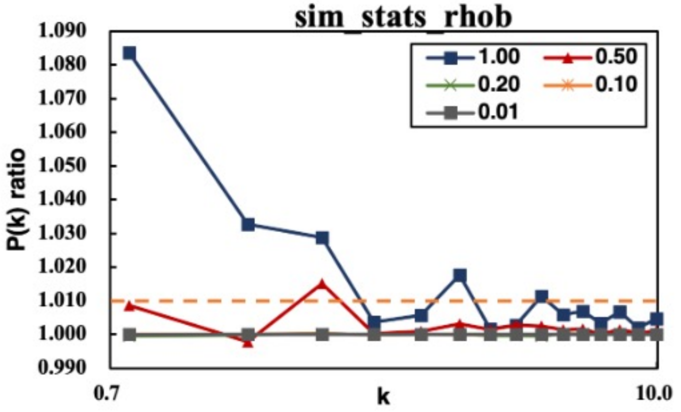
- AMR-based hydrodynamics code designed to model astrophysical reacting flows on HPC systems
- 6 3-D data fields (mainly)

➢ **Power Spectrum**

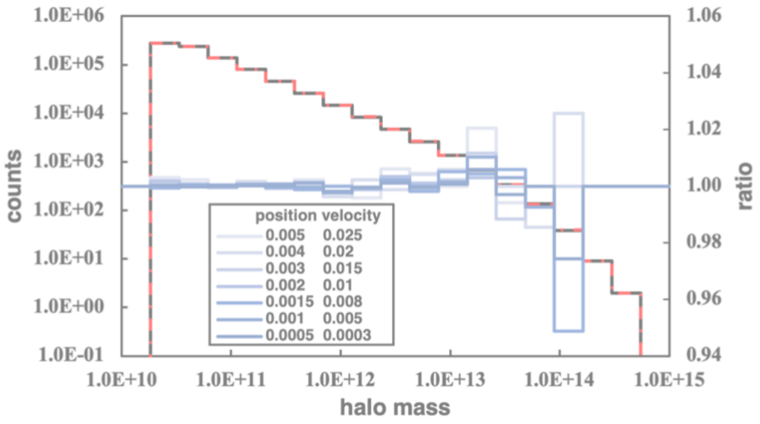
- FFT-based analysis for Universe's matter distribution
- **Target**: Ratio of P(k) on reconstructed data and original data remains within 1 ± 0.01

➢ **Halo Finder**

- Find over-densities in the Mass distribution
- **Target**: Minimize the halo mass change of each halo



Power spectrum analysis on baryon density.

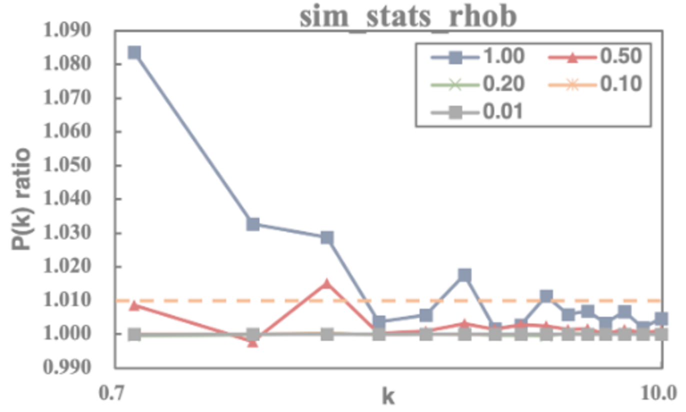


Halo Finder analysis on baryon density.

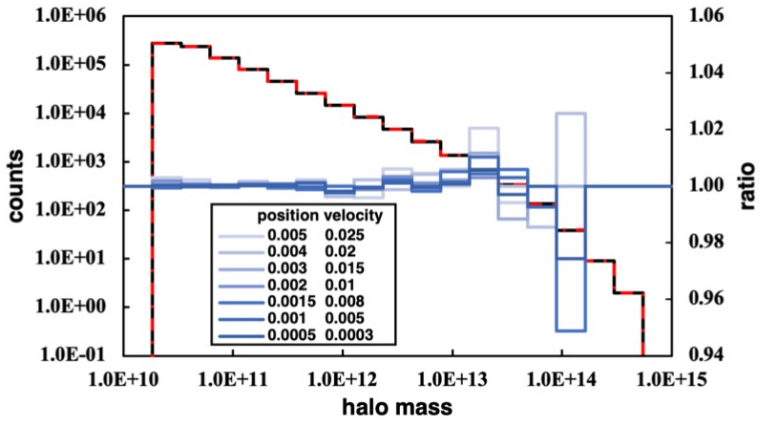
Jin, Sian, et al. "Understanding GPU-based lossy compression for extreme-scale cosmological simulations." *2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. IEEE, 2020.

Background

- **Nyx Cosmological Simulation**
  - AMR-based hydrodynamics code designed to model astrophysical reacting flows on HPC systems
  - 6 3-D data fields (mainly)
- **Power Spectrum**
  - FFT-based analysis for Universe's matter distribution
  - **Target**: Ratio of P(k) on reconstructed data and original data remains within 1 ± 0.01
- **Halo Finder**
  - Find over-densities in the Mass distribution
  - **Target**: Minimize the halo mass change of each halo



Power spectrum analysis on baryon density.



Halo Finder analysis on baryon density.

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.

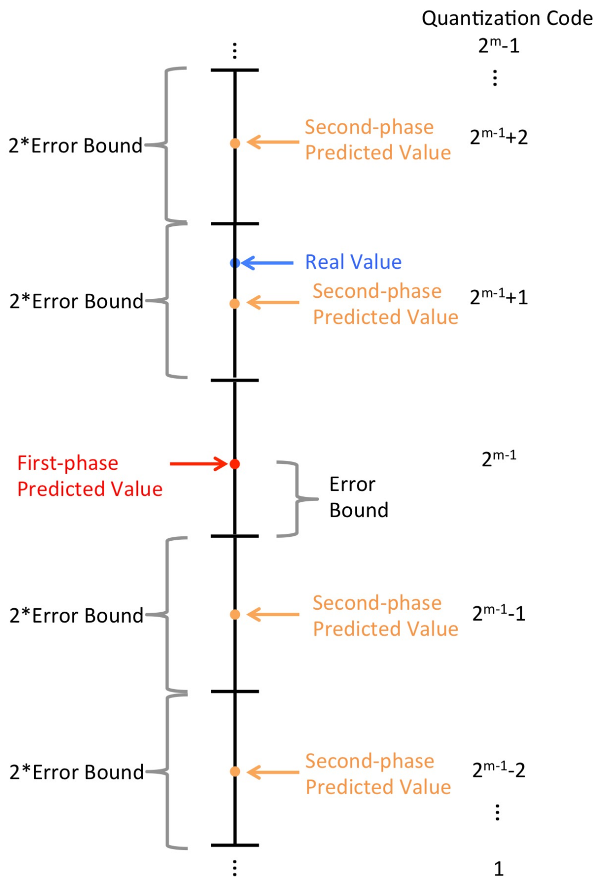
Background

➢ Lossy Compression

- Error-controllable lossy compression
  - Absolute error bound (ABS)
  - Power relative error bound (PW_ABS)
- High compression ratio
- High throughput (with GPU-based version)

➢ How SZ Works

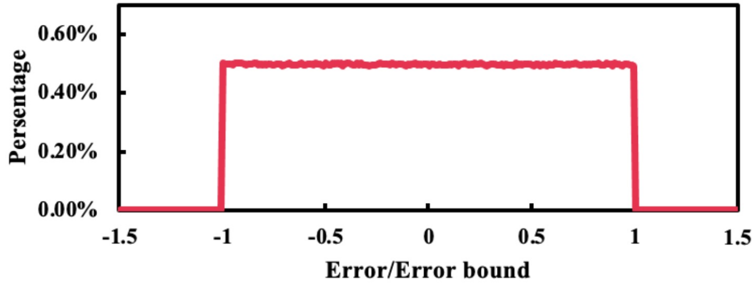
- Each data point's value is **predicted based on its neighboring data** points by an adaptive, best-fit prediction method.
- Each floating-point weight value is converted to an integer number by a linear-scaling **quantization** based on the difference between the real value and predicted value and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



Outline

- ➢ **Introduction**
  - Fine-grained lossy compression
  - Rate-quality modeling
- ➢ **Background**
  - Cosmology post hoc analysis
  - Lossy compression for scientific data
- ➢ **Designs**
  - Modeling error impact on post hoc analysis
  - Modeling compression ratio
  - Optimization strategy
- ➢ **Experimental Evaluation**
  - Rate-quality evaluation
  - Performance evaluation

Modeling Error Impact



Error distribution of temperature data in one Nyx dataset compressed by SZ lossy compression.

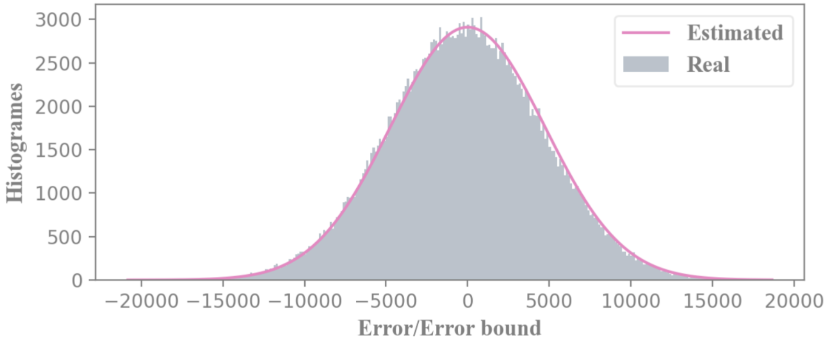
M: number of ranks/partitions
N: grid size, σ: deviation

➢ Introduced Error Distribution

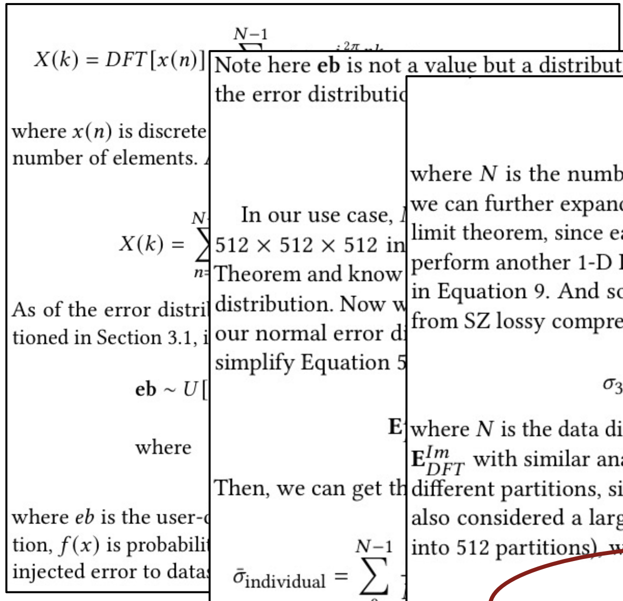
- SZ: **Uniform** distribution under ABS mode
- Can be adapted to **other compressor** with error distribution of deterministic **average** and **deviation**

➢ FFT-based Analysis

- Normal distribution based on **Central Limit Theorem**
- Works on all 1D, 2D and 3D DFTs



Comparison of real and estimated FFT error distribution based on our model using Nyx's temperature field.



Note here **eb** is not a value but a distribution function. We can get

$$\sigma = \sqrt{\frac{N}{6}}eb, \quad \mu = 0, \qquad (8)$$

where N is the number of elements in given 1-D data. Similarly, we can further expand our equation to 2-D DFT results by central limit theorem, since each row in the new dimension would further perform another 1-D DFT on values with error distribution shown in Equation 9. And so on, we can get 3-D DFT error distribution from SZ lossy compression is:

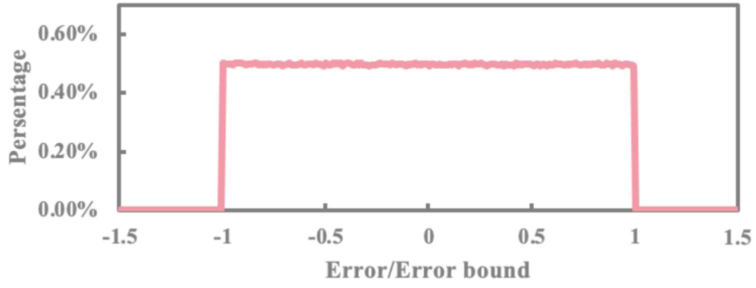
$$\sigma_{3D} = \sqrt{\frac{N^3}{6}}eb, \quad \mu_{3D} = 0, \qquad (9)$$

where N is the data dimension. The same error distribution goes to $\mathbf{E}^{Im}_{DFT}$ with similar analysis. As of applying various error bound to different partitions, since the element number in each partition is also considered a large number (e.g., 64^3 when cutting 512^3 data into 512 partitions), we can further derive our equation to:

$$\sigma_{3D} = \sum_{0}^{M-1}\sqrt{\frac{N^3}{6}}\frac{eb_m}{M}, \quad \mu_{3D} = 0, \qquad (10)$$

Modeling Error Impact

- **Introduced Error Distribution**
  - SZ: **Uniform** distribution under ABS mode
  - Can be adapted to **other compressor** with error distribution of deterministic **average** and **deviation**
- **FFT-based Analysis**
  - Normal distribution based on **Central Limit Theorem**
  - Works on all 1D, 2D and 3D DFTs



Error distribution of temperature data in one Nyx dataset compressed by SZ lossy compression.

M: number of ranks/partitions
N: grid size, σ: deviation

$X(k) = DFT[x(n)]$

where $x(n)$ is discrete

number of elements.

$X(k) =$

As of the error distri

tioned in Section 3.1, i

$\mathbf{eb} \sim U[$

where

where eb is the user-

tion, $f(x)$ is probabili

injected error to data

Note here **eb** is not a value but a distribution function. We can get the error distributio

In our use case,

$512 \times 512 \times 512$ in

Theorem and know

distribution. Now w

our normal error d

simplify Equation 5

Then, we can get th

$\bar{\sigma}_{\text{individual}} = \sum_{0}^{N-1}$

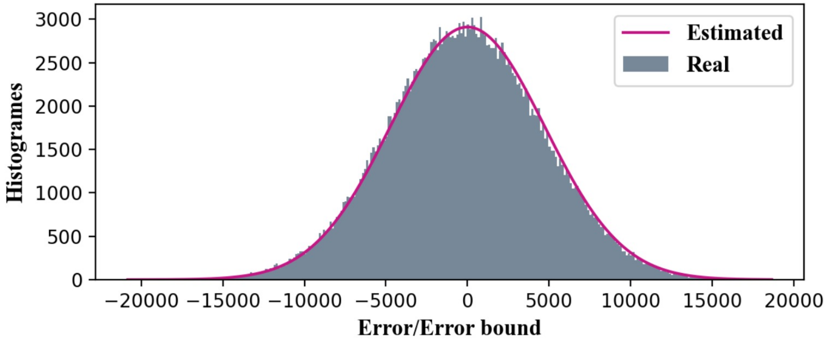
$$\sigma = \sqrt{\frac{N}{6}}eb, \quad \mu = 0, \tag{8}$$

where N is the number of elements in given 1-D data. Similarly, we can further expand our equation to 2-D DFT results by central limit theorem, since each row in the new dimension would further perform another 1-D DFT on values with error distribution shown in Equation 9. And so on, we can get 3-D DFT error distribution from SZ lossy compression is:

$$\sigma_{3D} = \sqrt{\frac{N^3}{6}}eb, \quad \mu_{3D} = 0, \tag{9}$$

where N is the data dimension. The same error distribution goes to $\mathbf{E}^{Im}_{DFT}$ with similar analysis. As of applying various error bound to different partitions, since the element number in each partition is also considered a large number (e.g., 64^3 when cutting 512^3 data into 512 partitions), we can further derive our equation to:

$$\sigma_{3D} = \sum_{0}^{M-1}\sqrt{\frac{N^3}{6}}\frac{eb_m}{M}, \quad \mu_{3D} = 0, \tag{10}$$

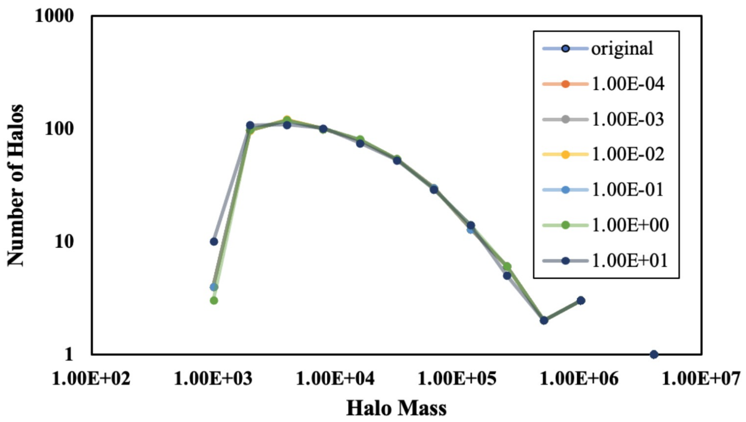


Comparison of real and estimated FFT error distribution based on our model using Nyx's temperature field.

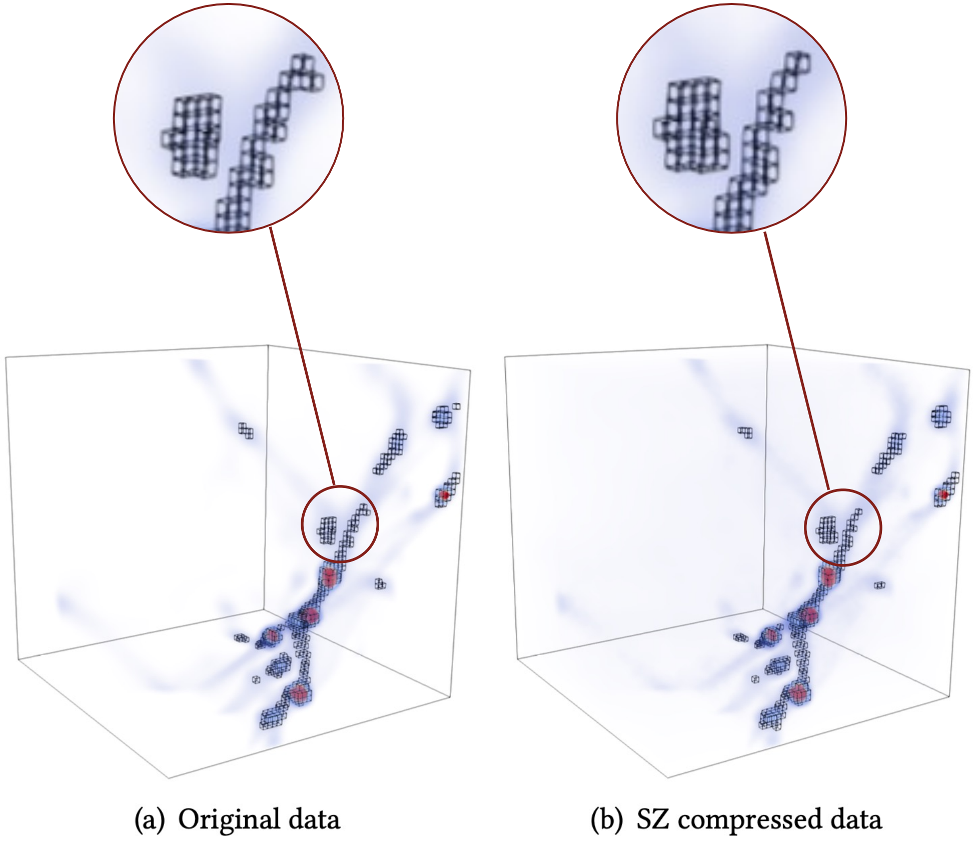
Modeling Error Impact

➢ Halo Finder

- Three key metrics:
  - Number of halos
  - Each halo's location
  - **Each halo's mass**
- Halo's mass depends on **error bound** and **cell count** of mass value within a certain range



Comparison of halo mass distribution with different error bounds using Nyx's baryon density field.



Candidate cells for halo finder

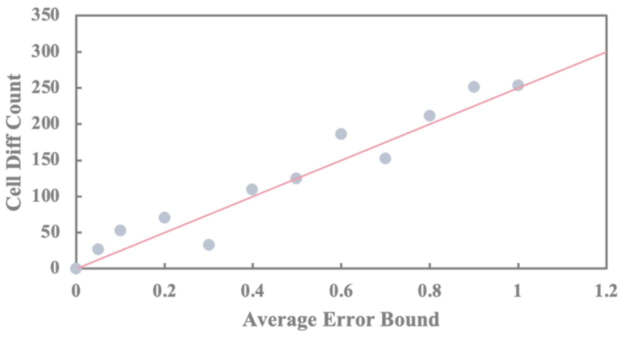
Modeling Error Impact

➢ Halo Finder

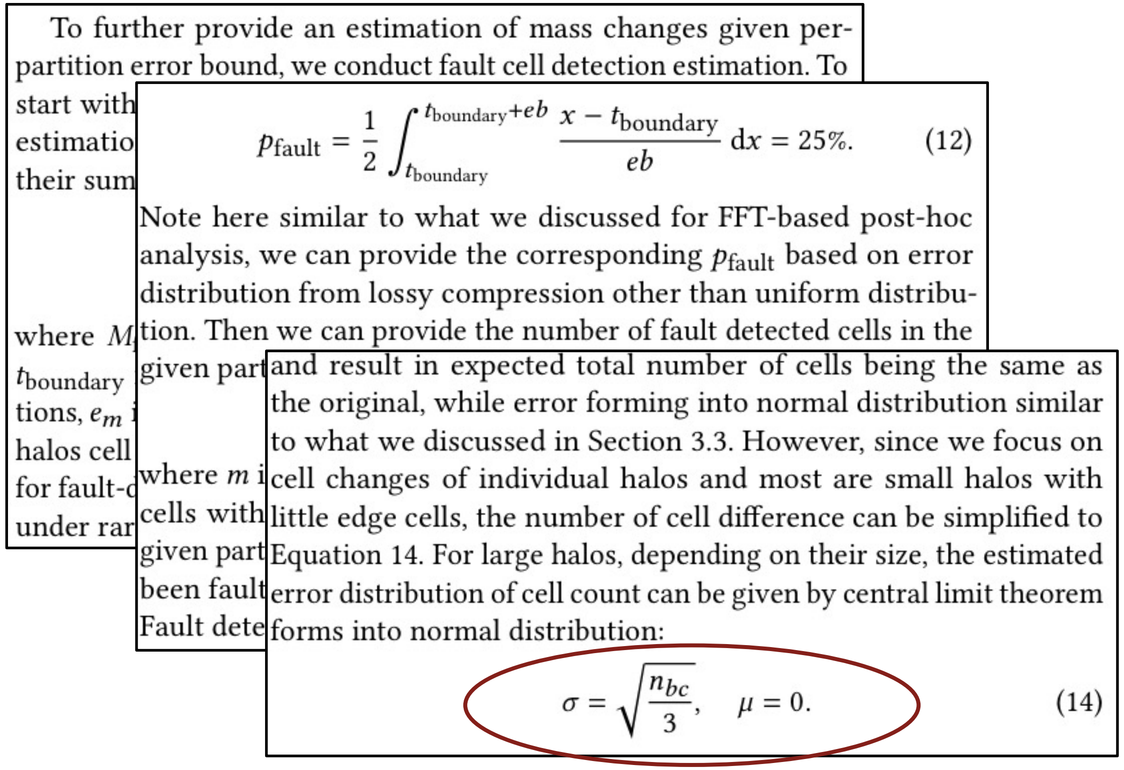
- Halo’s mass depends on **error bound** and **cell count** of mass value within a certain range
- Typically normal distribution

Error Bound	Cells	Mass	Mass Diff	Diff per cell
original	6023	3.13E+6	-	-
1E-2	6023	3.13E+6	0	-
1E-1	6011	3.13E+6	-9.8E+2	81.7
1E+0	6038	3.13E+6	1.21E+3	80.7
1E+1	6041	3.13E+6	1.66E+3	92.2

Mass difference per changed cell on a large halo.



Number of candidate cells changed with different error bounds



To further provide an estimation of mass changes given per-partition error bound, we conduct fault cell detection estimation. To start with estimatio their sum

$$p_{\text{fault}} = \frac{1}{2}\int_{t_{\text{boundary}}}^{t_{\text{boundary}}+eb} \frac{x - t_{\text{boundary}}}{eb}\,dx = 25\%. \quad (12)$$

Note here similar to what we discussed for FFT-based post-hoc analysis, we can provide the corresponding p_{fault} based on error distribution from lossy compression other than uniform distribution. Then we can provide the number of fault detected cells in the given part

where M t_{boundary} tions, e_m halos cell for fault-d under rar

where m i cells with given part been fault Fault dete

and result in expected total number of cells being the same as the original, while error forming into normal distribution similar to what we discussed in Section 3.3. However, since we focus on cell changes of individual halos and most are small halos with little edge cells, the number of cell difference can be simplified to Equation 14. For large halos, depending on their size, the estimated error distribution of cell count can be given by central limit theorem forms into normal distribution:

$$\sigma = \sqrt{\frac{n_{bc}}{3}}, \quad \mu = 0. \quad (14)$$

n_{bc}: number of cells near boundary (~81.66)
σ: deviation

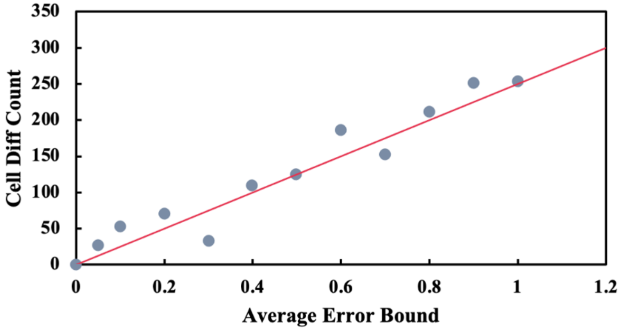
Modeling Error Impact

➢ Halo Finder

- Halo's mass depends on **error bound** and **cell count** of mass value within a certain range
- Typically normal distribution

Error Bound	Cells	Mass	Mass Diff	Diff per cell
original	6023	3.13E+6	-	-
1E-2	6023	3.13E+6	0	-
1E-1	6011	3.13E+6	-9.8E+2	81.7
1E+0	6038	3.13E+6	1.21E+3	80.7
1E+1	6041	3.13E+6	1.66E+3	92.2

Mass difference per changed cell on a large halo.



Number of candidate cells changed with different error bounds

To further provide an estimation of mass changes given per-partition error bound, we conduct fault cell detection estimation. To start with
estimatio
their sum

where M
$t_{boundary}$
tions, e_m i
halos cell
for fault-d
under rar

$$p_{\text{fault}} = \frac{1}{2}\int_{t_{\text{boundary}}}^{t_{\text{boundary}}+eb} \frac{x - t_{\text{boundary}}}{eb}\,dx = 25\%. \qquad (12)$$

Note here similar to what we discussed for FFT-based post-hoc analysis, we can provide the corresponding p_{fault} based on error distribution from lossy compression other than uniform distribution. Then we can provide the number of fault detected cells in the given part

where m i
cells with
given part
been fault
Fault dete

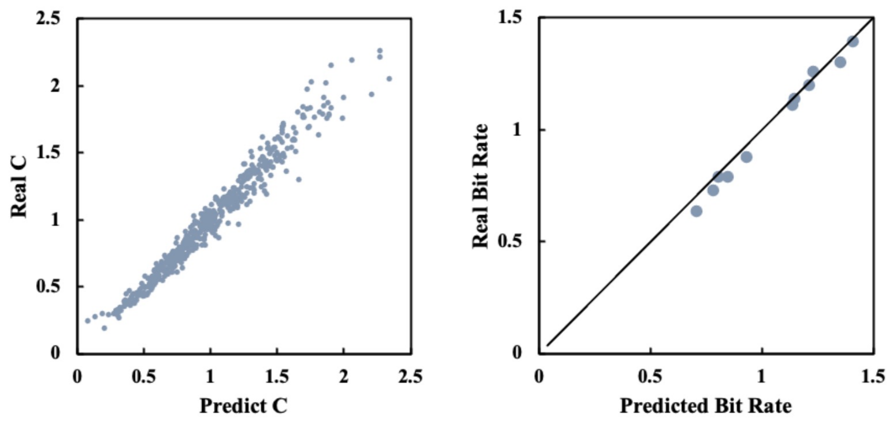
and result in expected total number of cells being the same as the original, while error forming into normal distribution similar to what we discussed in Section 3.3. However, since we focus on cell changes of individual halos and most are small halos with little edge cells, the number of cell difference can be simplified to Equation 14. For large halos, depending on their size, the estimated error distribution of cell count can be given by central limit theorem forms into normal distribution:

$$\sigma = \sqrt{\frac{n_{bc}}{3}}, \quad \mu = 0. \qquad (14)$$

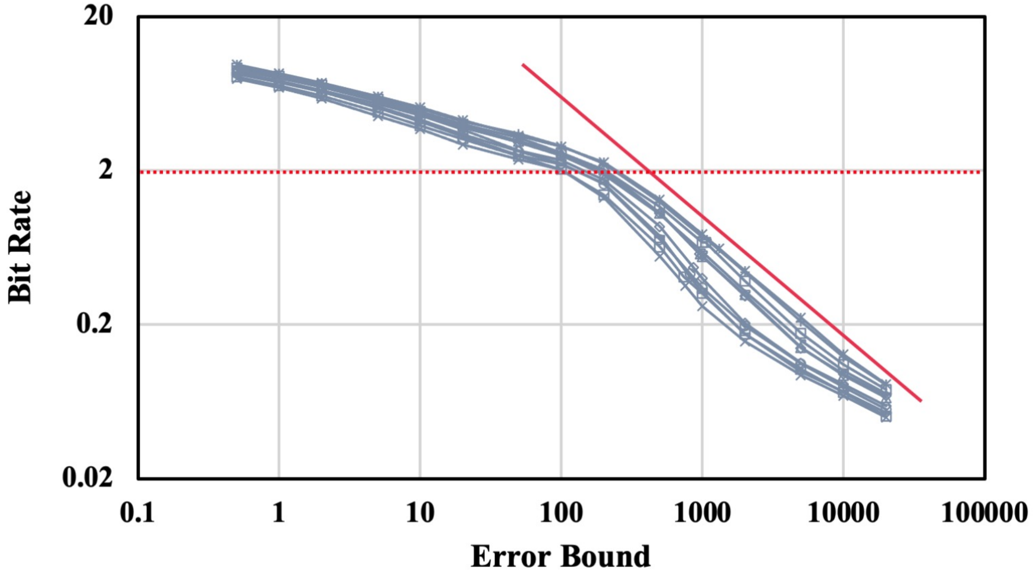
n_{bc}: number of cells near boundary (~81.66)
σ: deviation

Modeling Compression Ratio

- **SZ Lossy Compression**
  - Empirical model
  - Two main sections, critical point at bit rate of 2
  - All partitions forms in similar shape
  - Use mean value as parameter, can also use entropy but result higher performance cost



(a) Relative C_1 Estimation (b) Compression Ratio Consistency

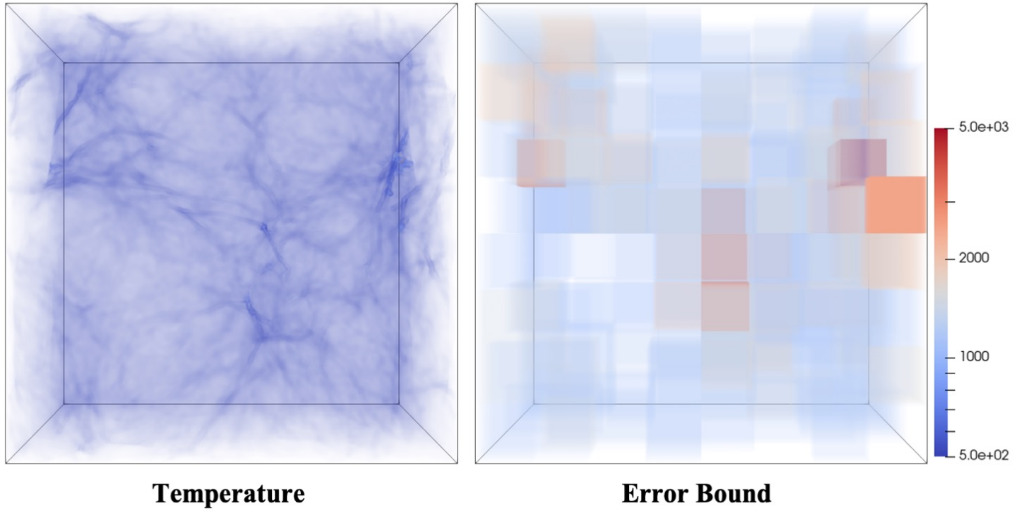


Bit rate with different error bounds using SZ lossy compression. Different lines represent for different partitions.

Modeling Compression Ratio

➢ Optimization Strategy

1. Parameter extraction (estimate compression ratio)
   - Mean value of given partition
   - Mean value of overall dataset
2. Build Rate-Quality Model
   - EB-quality model
   - EB-rate model
3. Per-partition error bound optimization
   - Derivatives of rate-quality curves are balanced for all partitions
4. For baryon density
   - Perform power-spectrum optimization first
   - Perform halo-finder optimization if not satisfied



Fine-grained lossy compression control for different data partitions.

Outline

- **Introduction**
  - Fine-grained lossy compression
  - Rate-quality modeling
- **Background**
  - Cosmology post hoc analysis
  - Lossy compression for scientific data
- **Designs**
  - Modeling error impact on post hoc analysis
  - Modeling compression ratio
  - Optimization strategy
- **Experimental Evaluation**
  - Rate-quality evaluation
  - Performance evaluation

Evaluation

➢ Setup And Dataset

- Evaluation with modified Foresight
  - Open-source toolkit used to evaluate, analyze, and visualize lossy compressors for extreme-scale cosmological simulations
- Nyx dataset provided by the Nyx development team at Lawrence Berkeley National Laboratory (LBNL)
- Experiment platforms
  - Cori system at NERSC
  - Frontera system at TACC

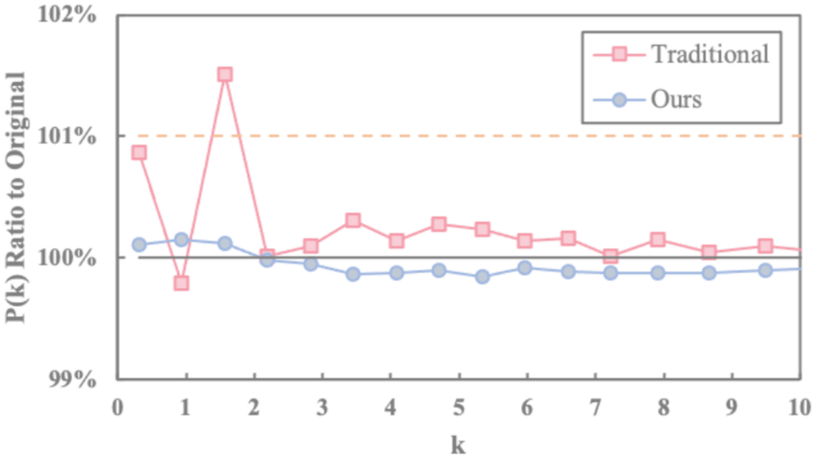
Dimension	Size	Field	Value Range
$512 \times 512 \times 512$	6.6 GB	Baryon Density	$(0,\ 10^5)$
$1024 \times 1024 \times 1024$	52 GB	Dark Matter Density	$(0,\ 10^4)$
$2048 \times 2048 \times 2048$	352 GB	Temperature	$(10^2,\ 10^7)$
		Velocity	$(-10^8,\ 10^8)$

Details of Nyx Dataset Used in Experiments.

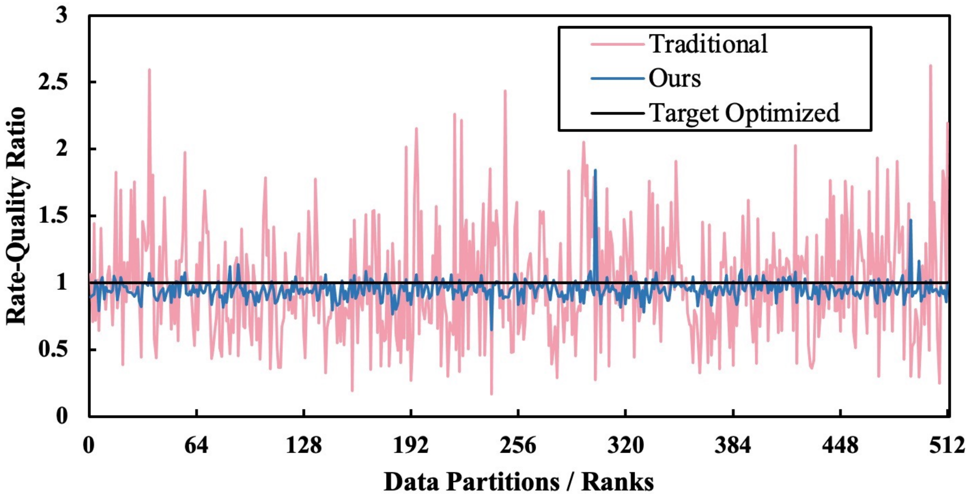
Foresight: https://github.com/lanl/VizAly-Foresight

Evaluation

- **Rate-quality**
  - Higher quality under similar bitrate
  - Lower bitrate (higher ratio) under similar quality
  - Smoother rate-quality balance across all partitions
- **Power Spectrum**



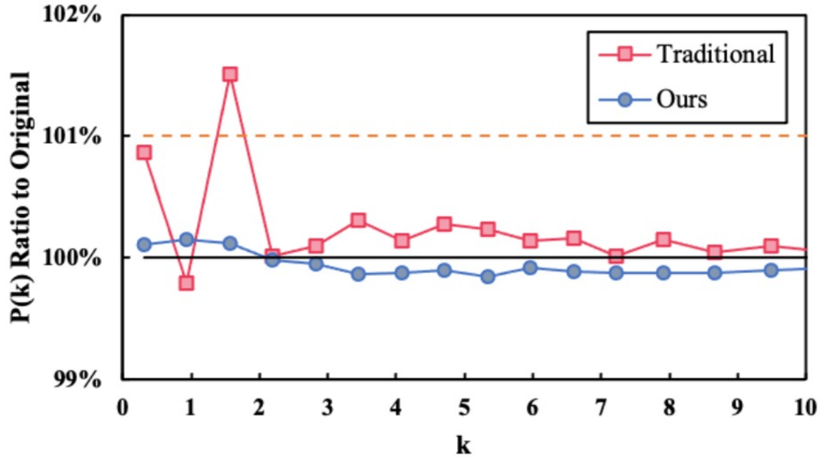
Power spectrum analysis on Nyx's baryon density field.



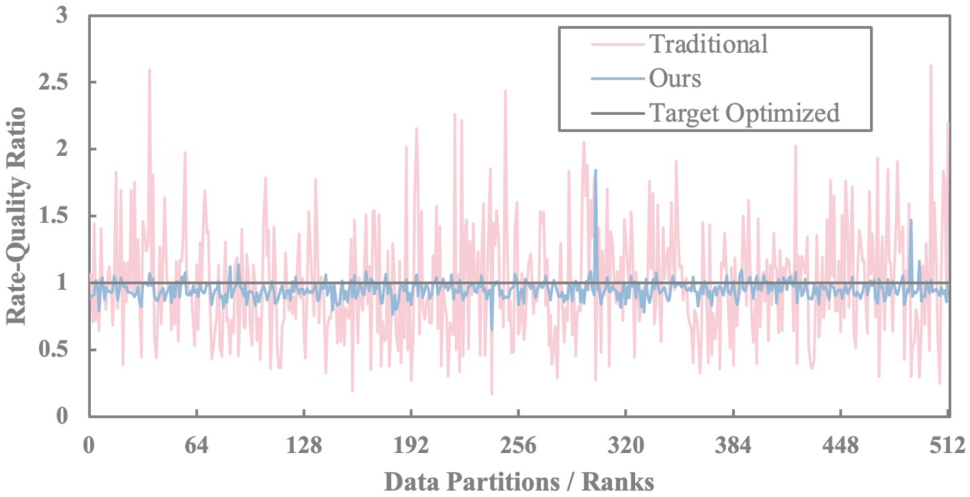
Comparison of bit-quality ratios using traditional and our methods on all partitions.

Evaluation

- **Rate-quality**
  - Higher quality under similar bitrate
  - Lower bitrate (higher ratio) under similar quality
  - Smoother rate-quality balance across all partitions
- **Power Spectrum**



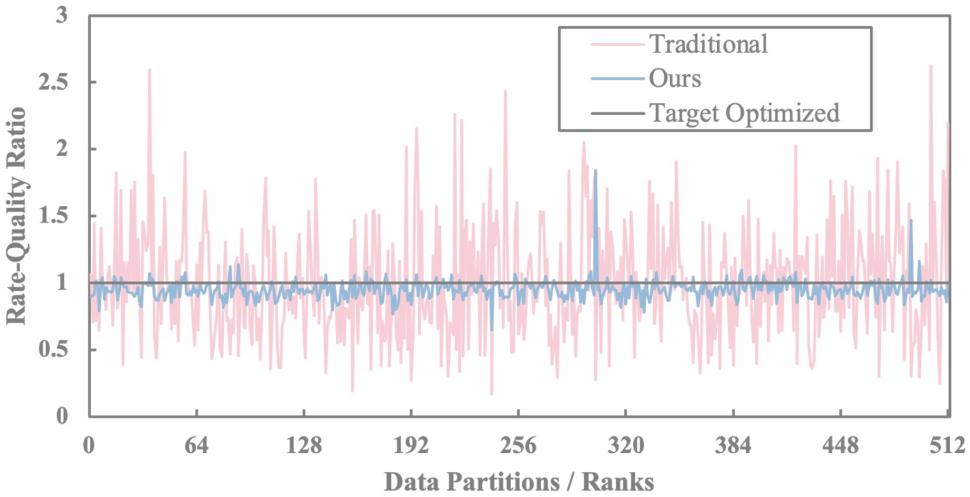
Power spectrum analysis on Nyx's baryon density field.



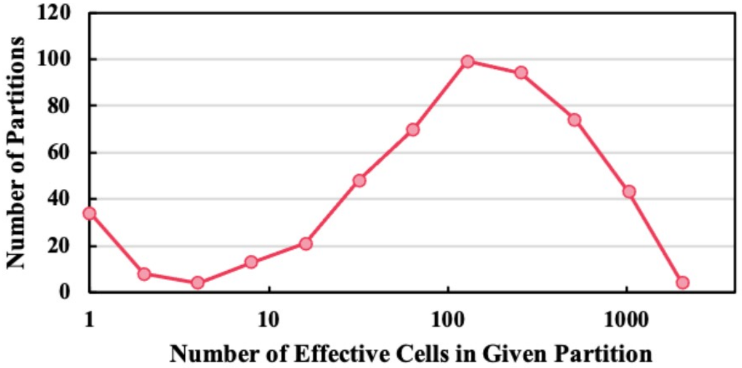
Comparison of bit-quality ratios using traditional and our methods on all partitions.

Evaluation

- **Rate-quality**
  - Higher quality under similar bitrate
  - Lower bitrate (higher ratio) under similar quality
  - Smoother rate-quality balance across all partitions
- **Halo Finder**
  - Number of effective cells varies
  - Result in **higher error bound** than power-spectrum-based optimization



Comparison of bit-quality ratios using traditional and our methods on all partitions.

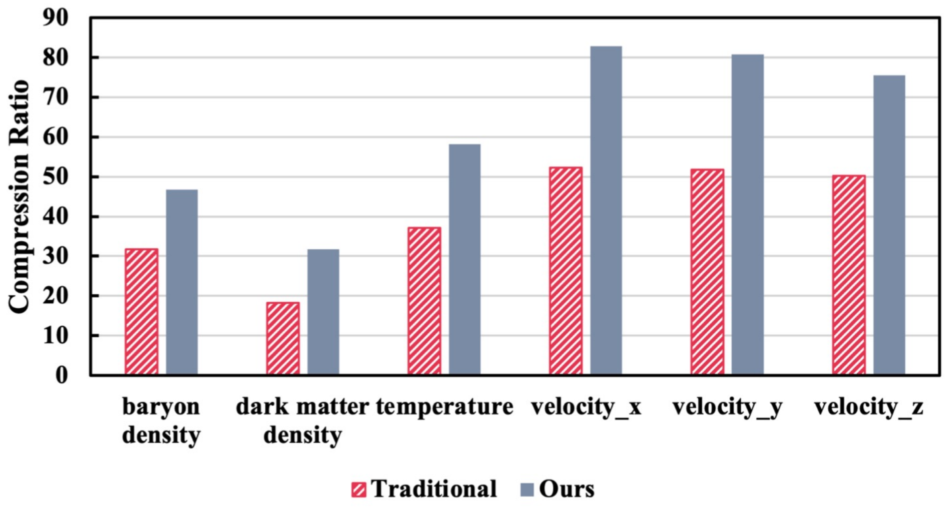


Histogram of effective cell count from all 512 data partitions of 1024×1024×1024 baryon density data.

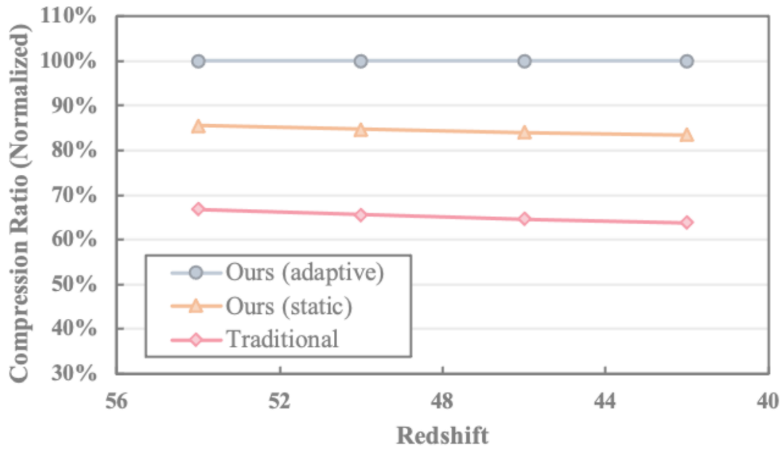
Evaluation

➢ **Compression Ratio Improvement**

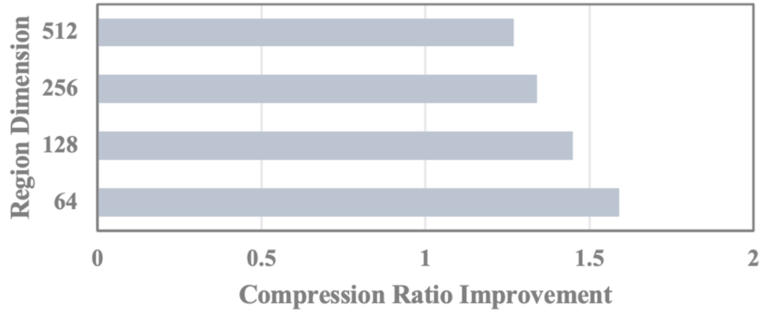
- Up to 1.73x, 1.56x overall improvement
- Capable across time steps
- Works best with smaller partitions
- Capable across simulation with different resolutions



Compression ratio comparison between our and traditional methods on all 6 Nyx fields.



Compression ratio comparison between our and traditional methods on multiple redshifts' data using baryon density field.

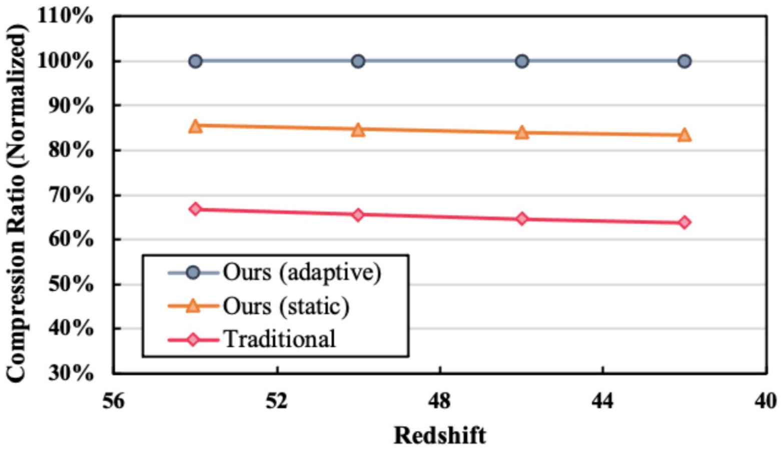


Compression ratio improvement with different partition sizes.

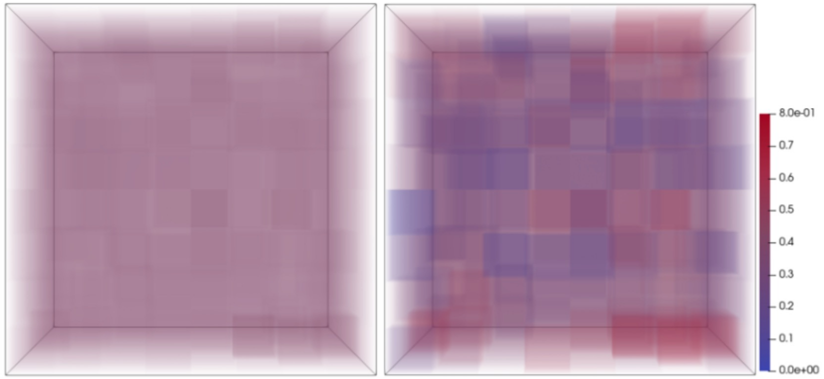
Evaluation

➢ Compression Ratio Improvement

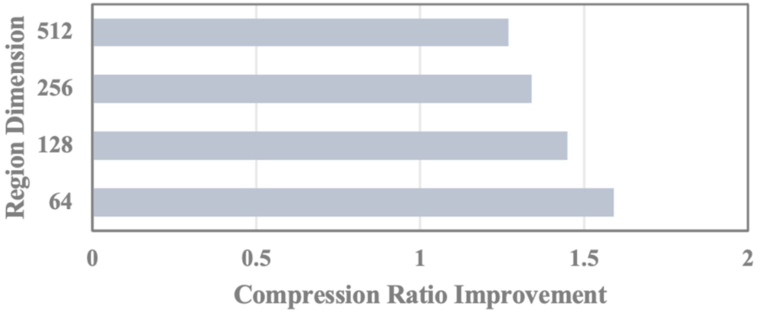
- Up to 1.73x, 1.56x overall improvement
- Capable across time steps
- Works best with smaller partitions
- Capable across simulation with different resolutions



Compression ratio comparison between our and traditional methods on multiple redshifts' data using baryon density field.



Comparison of our optimized error bounds on the data with larger redshift (left, early in simulation) and the data with lower redshift (right, late in simulation)



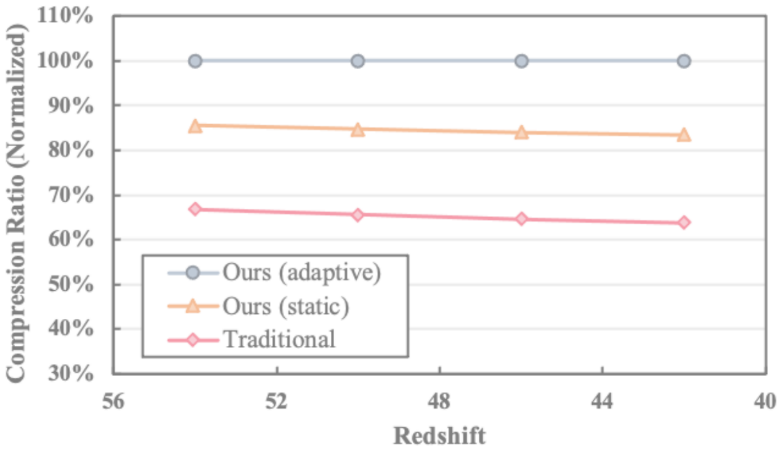
Compression ratio improvement with different partition sizes.

Evaluation

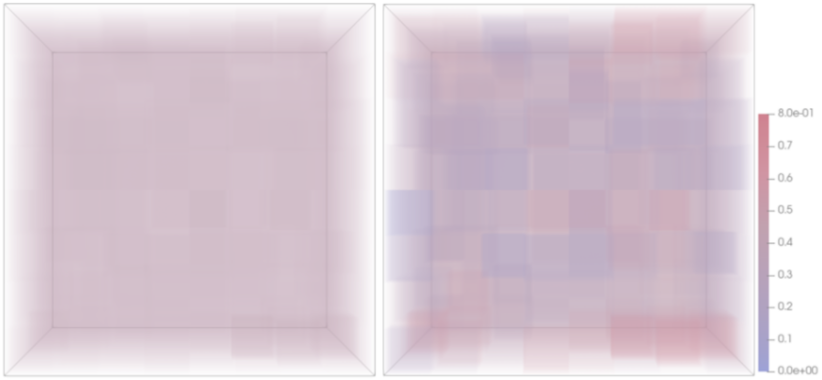
Los Alamos NATIONAL LABORATORY EST. 1943 · WASHINGTON STATE UNIVERSITY

➢ Compression Ratio Improvement

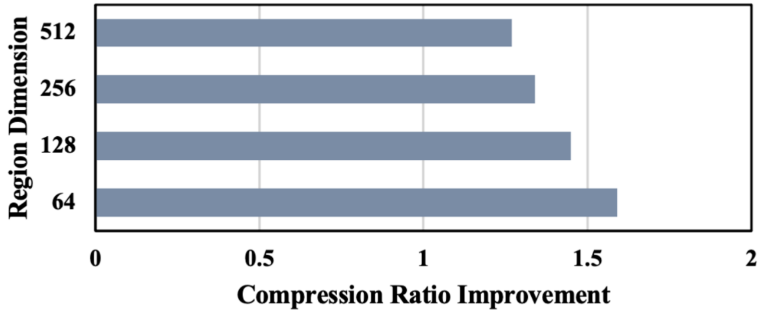
- Up to 1.73x, 1.56x overall improvement
- Capable across time steps
- Works best with smaller partitions
- Capable across simulation with different resolutions



Compression ratio comparison between our and traditional methods on multiple redshifts' data using baryon density field.



Comparison of our optimized error bounds on the data with larger redshift (left, early in simulation) and the data with lower redshift (right, late in simulation)



Compression ratio improvement with different partition sizes.

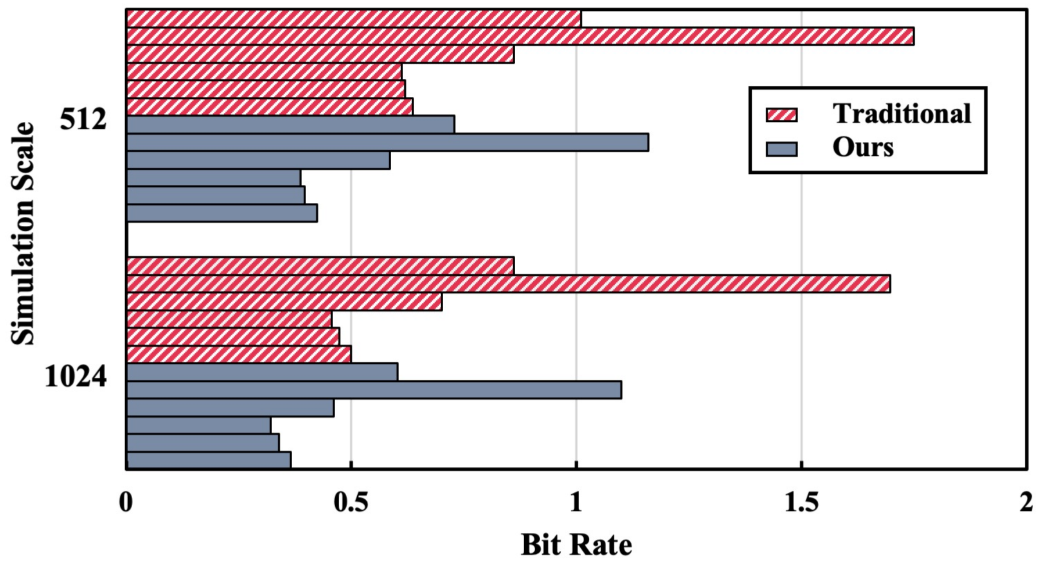
Evaluation

➢ Compression Ratio Improvement

- Up to 1.73x, 1.56x overall improvement
- Capable across time steps
- Works best with smaller partitions
- Capable across simulation with different resolutions

➢ Performance

- Little overhead, as low as 1%
- Only extract features of mean value and effective cell count



Compression efficiency improvement with different data sizes.

Conclusion

- **Our Approach**
  - **Adaptive approach** to select **feasible error-bound combinations** for **different partitions**
  - **Theoretical-based models** to efficiently estimate:
    - **Loss of post hoc post-analysis** caused by lossy compression error
    - **Compression ratio** based on each partition's data feature
  - efficient **optimization strategy** to determine the best-fit configuration
  - Improve the compression ratio by up to **73%** (with only **1% performance overhead**)
- **Future Work**
  - Applications and metrics able to establish theoretical analysis for post-hoc analysis

Thank you!

Any questions are welcome!

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

Argonne NATIONAL LABORATORY
NSF
ECP EXASCALE COMPUTING PROJECT
TACC