TOWARDS IMPROVING RATE-DISTORTION PERFORMANCE OF TRANSFORM-BASED LOSSY COMPRESSION FOR HPC DATASETS

JIALING ZHANG, AEKYEUNG MOON, XIAOYAN ZHUO, SEUNG WOO SON

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING
UNIVERSITY OF MASSACHUSETTS LOWELL

National Science Foundation
WHERE DISCOVERIES BEGIN

OUTLINE

- INTRODUCTION
- ANALYSIS OF TRANSFORM-BASED LOSSY COMPRESSION
- ENERGY COMPACTION BASED COMPRESSION ALGORITHM
- EXPERIMENT EVALUATION
- CONCLUSION

INTRODUCTION

SCIENTIFIC SIMULATION FRAMEWORKS (HPC SIMULATIONS)

- High-performance computing (HPC)
  - Perform research activities through computer modeling, simulation, and analysis.
  - Large volumes of data, reaches exabyte range
    - Periodic checkpointing (checkpoint/restart)
    - Post-simulation data analysis

- File systems in extreme-scale systems
  - Limited storage space and I/O time, e.g.:
    - 170 TB of CESM data is being produced by CMIP5 project
    - n*PB data will be generated for upcoming CMIP6 experiments per entire run
    - Yellowstone (supercomputer): tens of PB of centralized file system and data storage

INTRODUCTION

DATA REDUCTION/DATA COMPRESSION

- Data Reduction/Data Compression
  - Lossless Compression: 100% of data fidelity (fully invertible, 1:1 copy), but not able to achieve appreciable data reduction (GZIP, BZIP, etc.)
  - Lossy Compression: high compression ratio, but error introduced (SZ, ZFP, etc.)

- Floating Point Data Compression (Scientific Simulations)
  - e.g., double precision
  - High precision on reconstructed (decompressed) data
  - Randomness

INTRODUCTION

LOSSY COMPRESSION TECHNIQUES FOR SCIENTIFIC DATA

- Understanding Errors
  - Scientific data can tolerate a certain amount of accuracy loss
  - Errors are inherent in scientific simulations (generated from inaccurate scientific sensors)

- Study by Tao et al.: A comprehensive study of lossy compression on HPC datasets
  - Examined the impact of reduced accuracy on scientific data analysis frameworks.
  - In-depth understanding of the benefits and pitfalls (lossy compression)

ANALYSIS OF TRANSFORM-BASED LOSSY COMPRESSION

TRANSFORM-BASED LOSSY COMPRESSION

- Discrete Data Transform (DCT, HWT, CDF 9/7, etc.)
- Decorrelation (vs. time domain)
- Energy Compaction Property
  - Energy of a signal

$$\varepsilon_f = \sum_{n=1}^{N} |f_n|^2, n = 1, 2, \ldots, N,$$

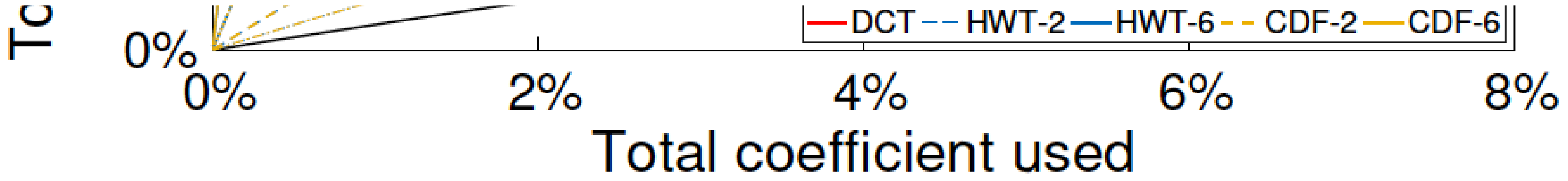
  - Energy Compaction Rate (ECR)

$$ECR_f = \frac{\sum_{n=1}^{M} |f_n|^2}{\sum_{n=1}^{N} |f_n|^2}, n = 1, 2, \ldots, N, M \leq N.$$

ANALYSIS OF TRANSFORM-BASED LOSSY COMPRESSION

- Estimation of Energy Compaction on Various Transform
  - Real-world HPC datasets

Transform	Threshold	rlds	mrsos	sedov	cellular	Eddy	Vortex
Original	1/32	6.03	21.65	27.27	6.53	25.91	44.28
	1/64	3.09	11.63	15.50	3.45	16.08	28.12
DCT-II	1/32	99.81	91.36	94.50	99.49	94.78	98.35
	1/64	99.69	88.17	92.06	99.13	89.29	96.93
HWT 5-level	1/32	96.94	33.22	65.91	92.86	36.64	36.01
HWT 6-level	1/64	93.63	17.60	47.87	86.67	18.12	20.19
CDF 9/7 5-level	1/32	98.08	39.17	62.78	91.82	24.76	27.07
CDF 9/7 6-level	1/64	95.83	21.58	44.46	84.47	11.97	15.47



ENERGY COMPACTION BASED COMPRESSION ALGORITHM

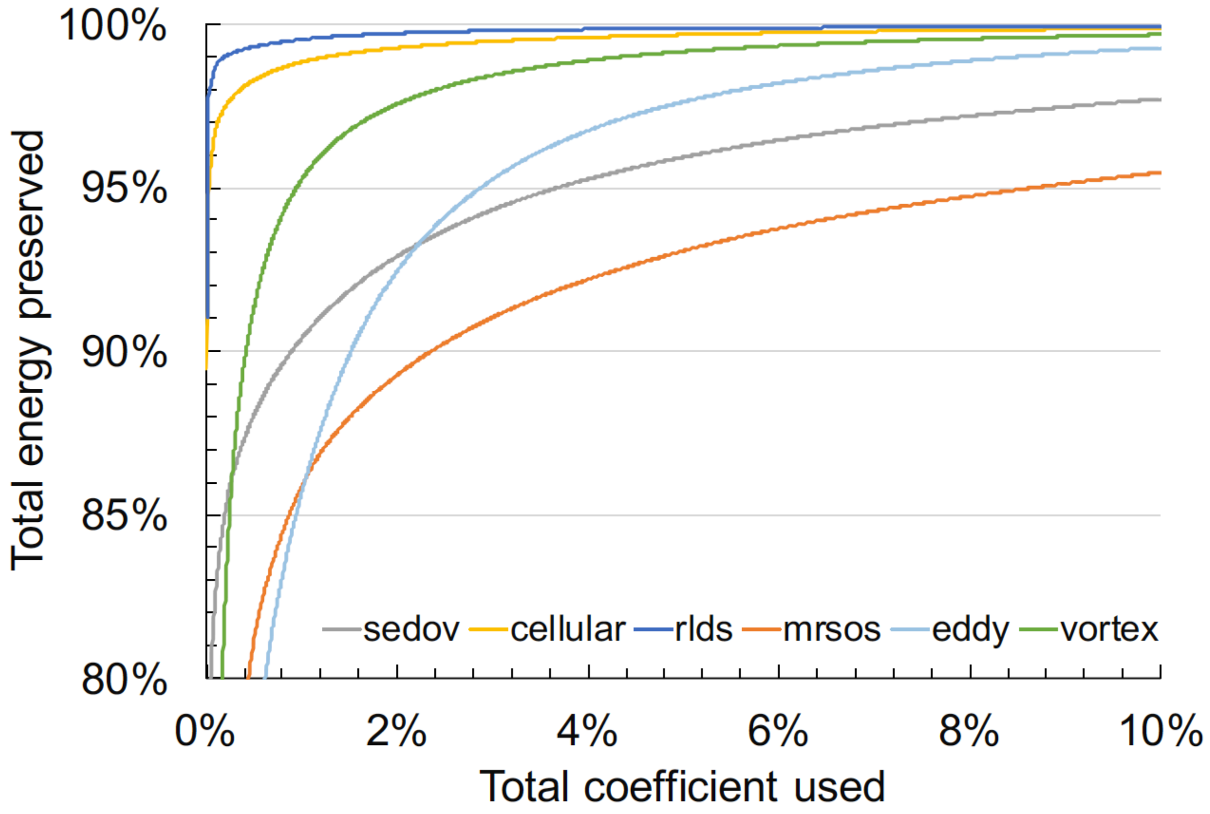
DCT (DISCRETE COSINE TRANSFORM)- BASED LOSSY COMPRESSION

- Block Decomposition with DCT

- Energy Compaction based compression algorithm
  - Compression with a fixed energy compaction rate (ECR)
  - Compression with an optimal energy compaction rate (ECR)

COMPRESSION WITH A FIXED ENERGY COMPACTION RATE

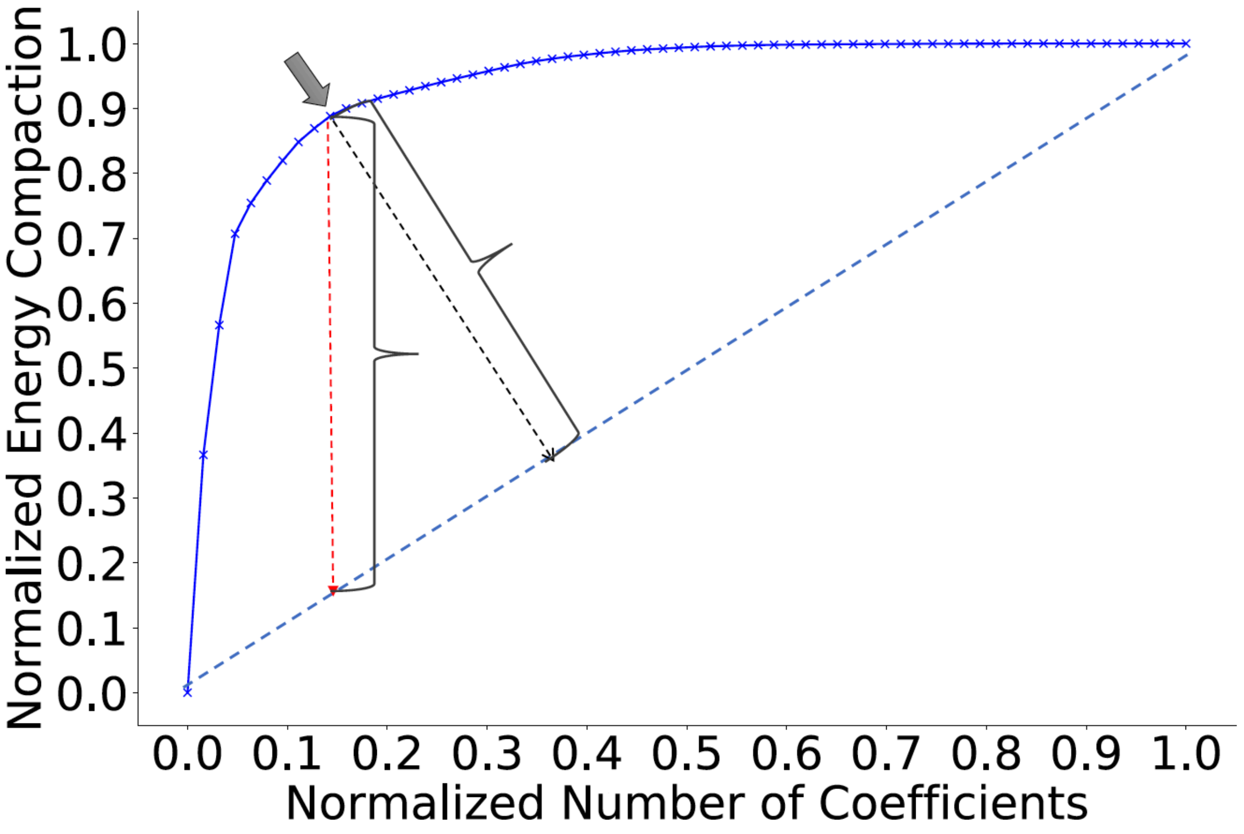
- Fixed Energy Compaction Rate (ECR)
- Find top dominant block coefficients



COMPRESSION WITH AN OPTIMAL ENERGY COMPACTION RATE

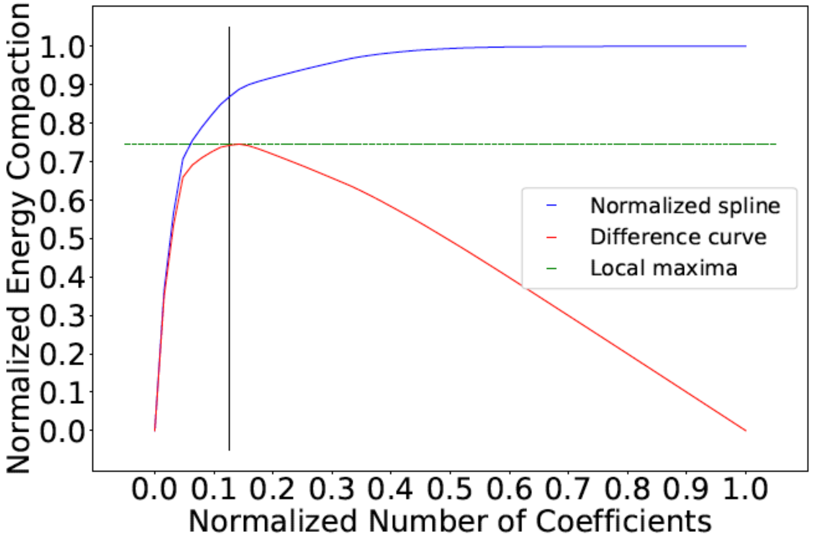
- Spline Fitting
- Kneedle Algorithm
  - Knee-point

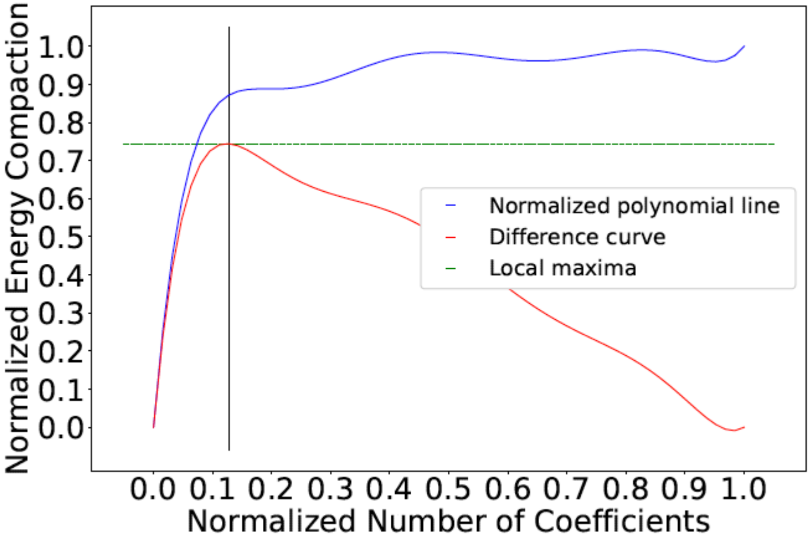
$$K_f(x) = \frac{f''(x)}{(1 + f'(x)^2)^{1.5}},$$



COMPRESSION WITH AN OPTIMAL ENERGY COMPACTION RATE

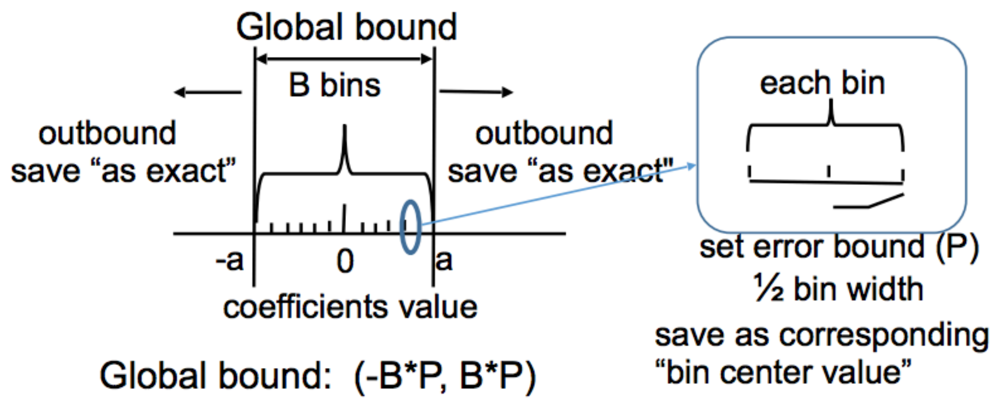
- Spline Fitting
  - 1D interpolation
  - Polynomial interpolation





APPROXIMATION & ENCODING

- Adjustable error-controlled quantizer
  - Equal width binning
  - Approximate as bin's center value
  - Fixed error bound



- Encoding
  - Dominant coefficient
  - Approximation: bin center values
  - Bin indices (and QT)
  - Lossless add-on

EVALUATION (EXPERIMENTAL SETUP & APPLICATIONS)

- Conduct on Massachusetts Green High Performance Computing Cluster (MGHPCC)
- 6 real-world scientific datasets
  - rlds, mrsos, cellular, sedov, vortex, eddy

Code	Description
FLASH	Physics Solvers implicit with AMR
CMIP5	Coupled Model Intercomparison Project World Climate
Nek5000	High-order Solver for fluid dynamics

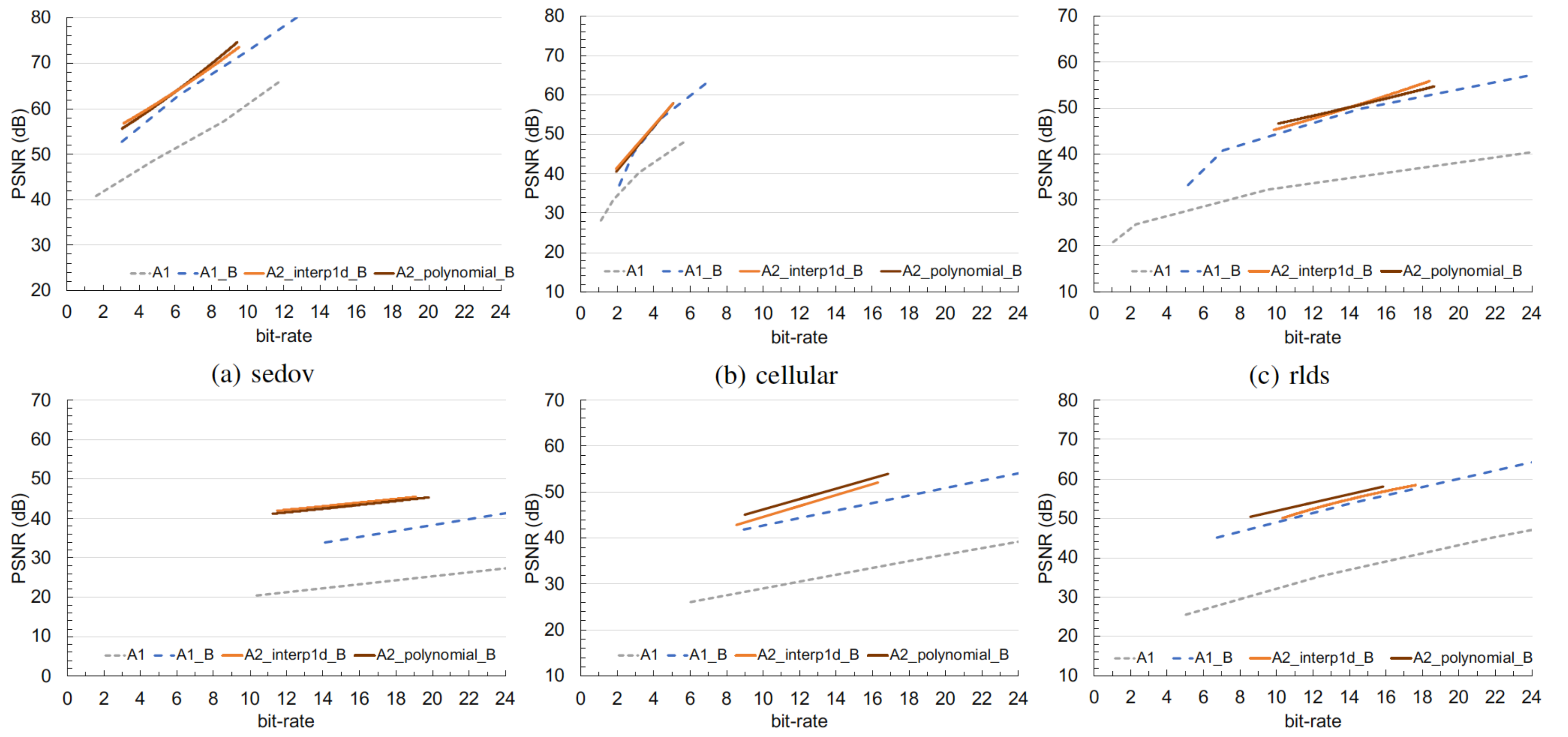
Code	Dataset	Value Range	Avg Value	Entropy	Dimension
FLASH [21]	sedov cellular	4.2385 $2.6482E^7$	1.0000 $2.2083E^7$	4.9702 4.1190	31040*154 32768*295
CMIP5 [22]	rlds mrsos	361.2303 44.5000	285.8844 7.6916	7.2106 4.4864	12960*100 12960*100
Nek5000 [23]	eddy vortex	4.8345 0.0550	$3.2366E^{-8}$ 0.0017	7.6047 7.5797	16384*999 37024*99

EVALUATION (SCHEMES & METRICS)

- Evaluation Schemes
  - **A1**: compression with fixed energy compaction rate.
  - **A2**: compression with optimal energy compaction rate.
  - **A2_interp1d**: A2 using 1D interpolation.
  - **A2_polynomial**: A2 using polynomial interpolation.
  - **A1_B**: A1 with equal-width-binning.
  - **A2_interp1d_B** & **A2_polynomial_B**: A2_interp1d and A2_polynomial with equal-width-binning, respectively.
- Rate-Distortion
  - Bit-rate (smaller bit-rate represents higher compression ratio)
  - Distortion: PSNR (higher PSNR represents less error)

EVALUATION

COMPARISON BETWEEN FIXED AND OPTIMAL ENERGY COMPACTION RATES



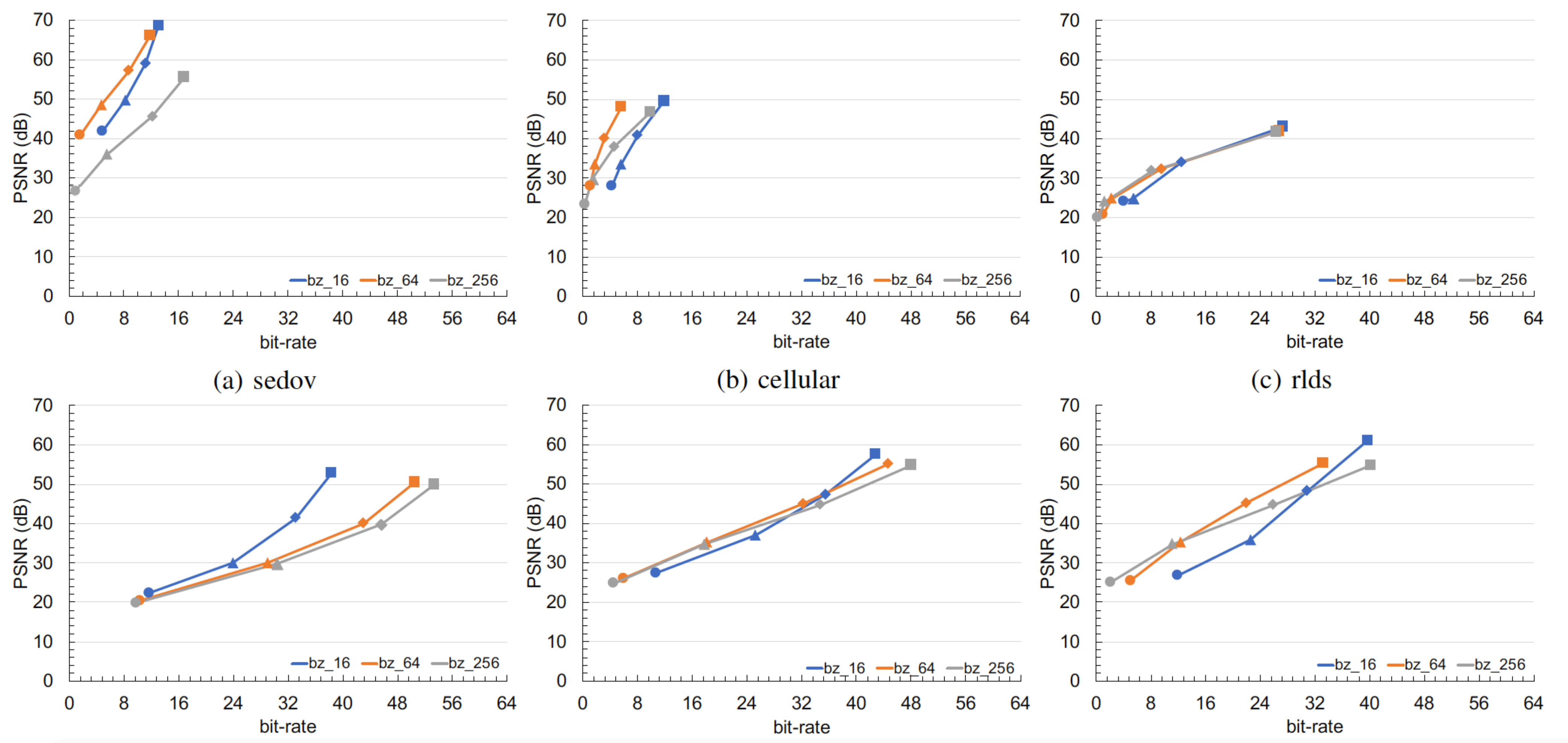
(a) sedov (b) cellular (c) rlds

(d) mrsos (e) eddy (f) vortex

- A1→A1_B
  PSNR↑ 15.8dB
- A2_interp1d_B→A1_B
  PSNR ↑2.46dB
  (mrsos, 6.49dB)
- 5.5 bit (~99% Energy)
- 6.67 bit (~97% Energy)

EVALUATION

RATE-DISTORTION AND BLOCK SIZE



(a) sedov (b) cellular (c) rlds

(d) mrsos (e) eddy (f) vortex

EVALUATION

SPLINE FITTING

Total number of coefficients (on average) used in each block

Algorithm	sedov	cellular	rlds	mrsos	eddy	vortex
Interp1d	3.929	6.844	9.356	11.172	9.597	11.925
Polynomial	3.853	6.449	9.557	10.956	9.182	10.122

Average energy compaction rate (%)

Algorithm	sedov	cellular	rlds	mrsos	eddy	vortex
Interp1d	98.44	99.59	99.99	99.99	89.83	93.55
Polynomial	98.44	99.59	99.99	99.99	90.12	91.08

CONCLUSION

- We analyze different transforms by exploiting their energy compaction property. By finding an optimal energy compaction rate based on our knee detection algorithm, our compression technique can acquire the best trade-off solution between information loss and compression rate.

- Specially, on average, only 6.67 bits are required to preserve an optimal energy compaction rate on our evaluated datasets. Our knee detection algorithm improves the distortion in terms of peak signal-to-noise ratio by 2.46 dB on average.

THANK YOU

JIALING_ZHANG@STUDENT.UML.EDU