Statistical Prediction of Lossy Compression Ratios for 3D Scientific Data

David Krasowska[*] (presenter), Robert Underwood[+], Julie Bessac[+], Jon Calhoun[*], Sheng Di[+], Franck Cappello[+]

[*]Department of Electrical and Computer Engineering at Clemson University

[+]Mathematical and Computer Science Division at Argonne National Laboratory

SC22 Dallas, TX hpc accelerates.

Why use compression in HPC?

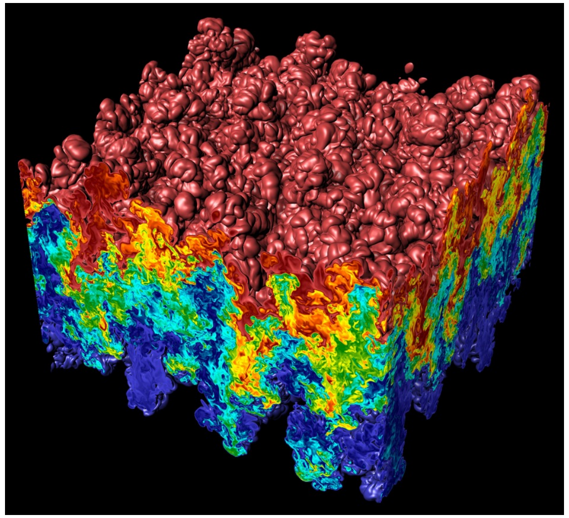
- HPC applications require lots of storage and memory throughput
- Compression allows for larger problem sizes to be ran while accelerating I/O time
- Checkpoint snapshots of an application’s state

Why estimate lossy compression ratios (CRs)?

- Finding the best compressor for the given data
- Accurate estimation enables I/O optimizations
  - Compare different compressors for minimum data size
  - Predict transfer times for lossy data over network links
  - Resource allocation planning
- Next step towards theoretical limit for lossy compressibility

Our contributions

1. Ability to accurately predict compression ratio on 3D scientific data
2. Lower prediction errors than previous attempts
   - <10% error across many compressors and datasets
3. Flexible across compressors, error bounds, and datasets
   - Compressor-free predictors (black-box)
4. Faster than other statistical predictors used in previous models
5. GPU accelerated: 57x speedup compared to CPU implementation



https://wci.llnl.gov/simulation/computer-codes

Estimating lossless compression ratios

- **Entropy**: theoretical lower bound limit of average number of bits needed to code output of source bitstream

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

X_i symbol

- Optimal lossless compressors equal this limit
- No theoretical quantification of lossy compressibility exists

Why is this challenging?

- Compressors have different methods of data reduction
- Need to capture the different notions of:

Correlation | **Entropy** | **Lossyness**

Why is this challenging?

- Compressors have different methods of data reduction
- Need to capture the different notions of:

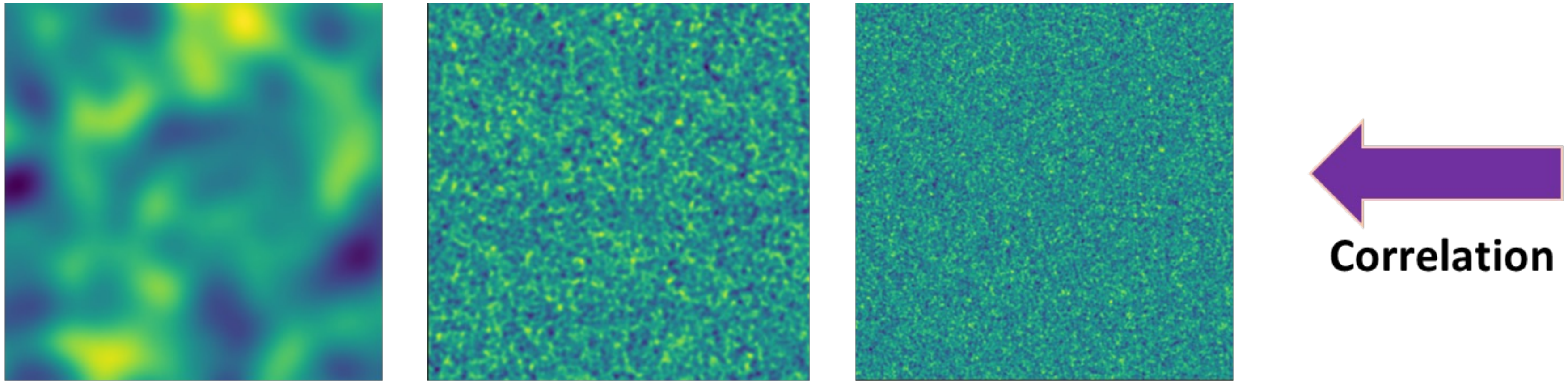


Image:	a = 1	a = 0.1	a = 0.05
SZ 1e-2 abs CR:	37.8	23.7	15.7
SZ 1e-5 abs CR:	5.1	4.9	4.1

Introduction

Previous Work

Our Model

Results

Conclusion

Prior work was either inaccurate or slow

- Depend on knowledge of a compressor's design principles
  - High error and relied on many internals of SZ **[Z. Qin]**
  - Improved error but still relied on blocksize **[D. Tao]**
- Rely on trial and error **[R. Underwood]**

Method	Speed	Accuracy
[R. Underwood]	☹	☺
[D. Tao]	☺	☹
[Z. Qin]	☺	☹
Our Method	☺	☺

Our previous work (2D)

- Presented at DRBSD-7 SC’21
- Relied heavily on the variogram
  - Extremely slow relative to modern compressors
- No model of CR based on correlation metrics and error bound

Statistical predictors

Correlation

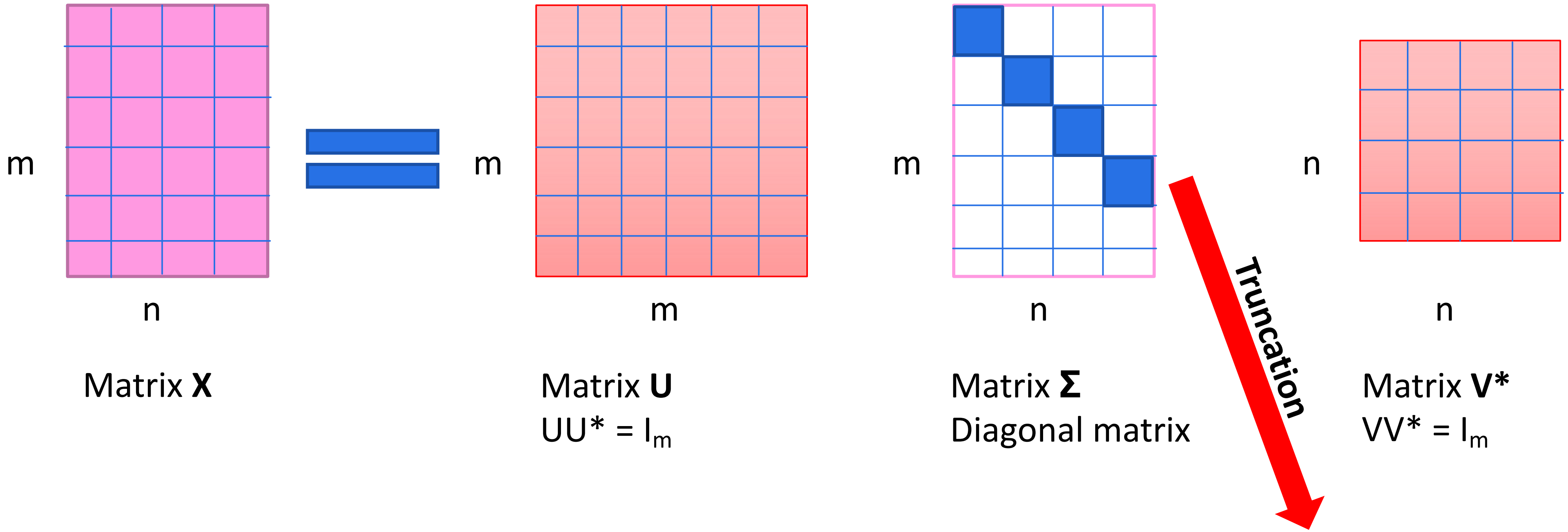
Truncated Singular Value Decomposition (SVD-trunc)

Entropy

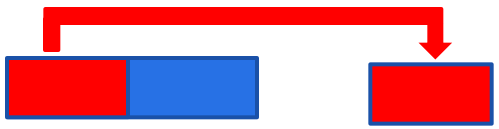
Lossyness

Quantized entropy (q-ent)

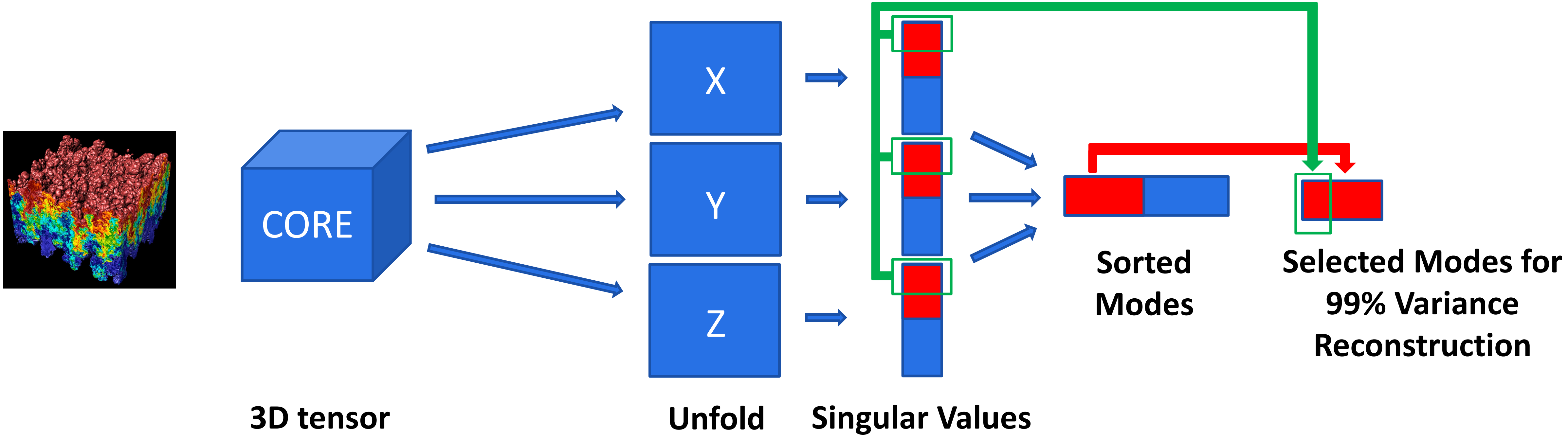
Truncated Singular Value Decomposition (SVD-trunc)



$$X = U\,\Sigma V^*$$

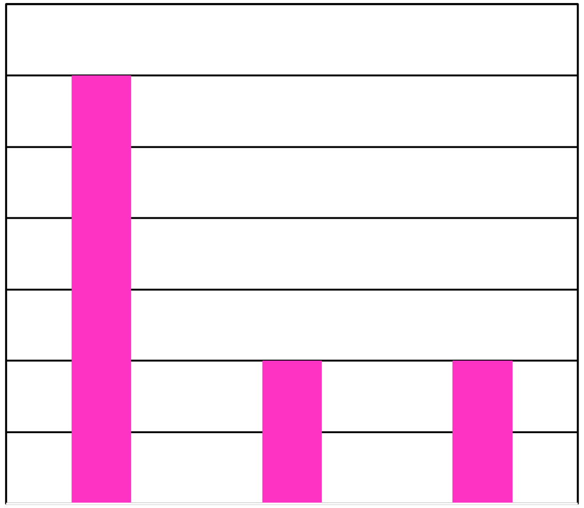


What is the Higher Order SVD (HOSVD)?



What is the quantized entropy (q-ent)?

Original	Quantize $\epsilon = 0.1$
2.31	2.3
2.34	2.3
3.45	3.4
4.61	4.6
2.27	2.3



Probability Density Function

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

$= 2.\ 2553$ average bits / symbol

Our linear regression model

$$\begin{aligned}\log(\text{CR}) \quad &= \quad a + b \times \log(\text{q-ent}) + c \times \log\left(\frac{\text{SVD-trunc}}{\sigma}\right) \\ &\quad +d \times \log(\text{q-ent}) \times \log\left(\frac{\text{SVD-trunc}}{\sigma}\right) + \epsilon,\end{aligned}$$

- Trained on observed CR and statistical predictors
- Least-square techniques to estimate parameters from observed training datasets
- K-fold cross validation to assess without bias

Introduction

Previous Work

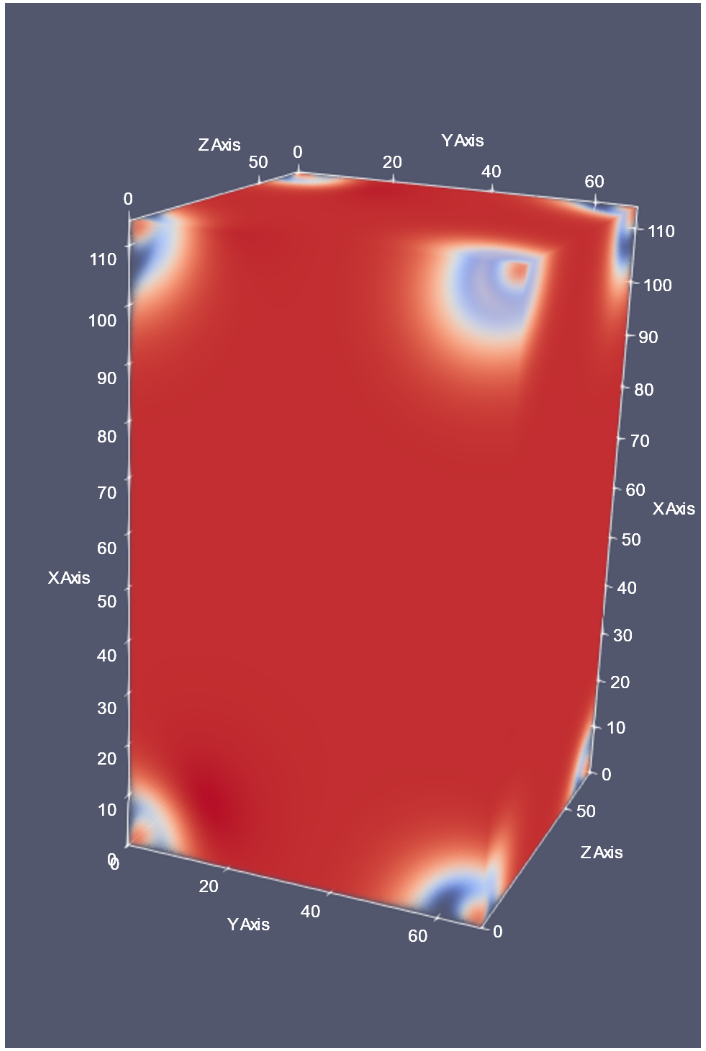
Our Model

Results

Conclusion

Experimental setup for 3D

- 288 3D orbitals from QMCPack were used
  - SDRBENCH benchmark suite
  - Containing structures of atoms, molecules, and solids
- Leading error bounded lossy compressors used
  - SZ, ZFP, MGARD, Bit Grooming, TTHRESH, and more
- Other datasets and results are comparable



QMCPack compression estimation exhibited low 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%

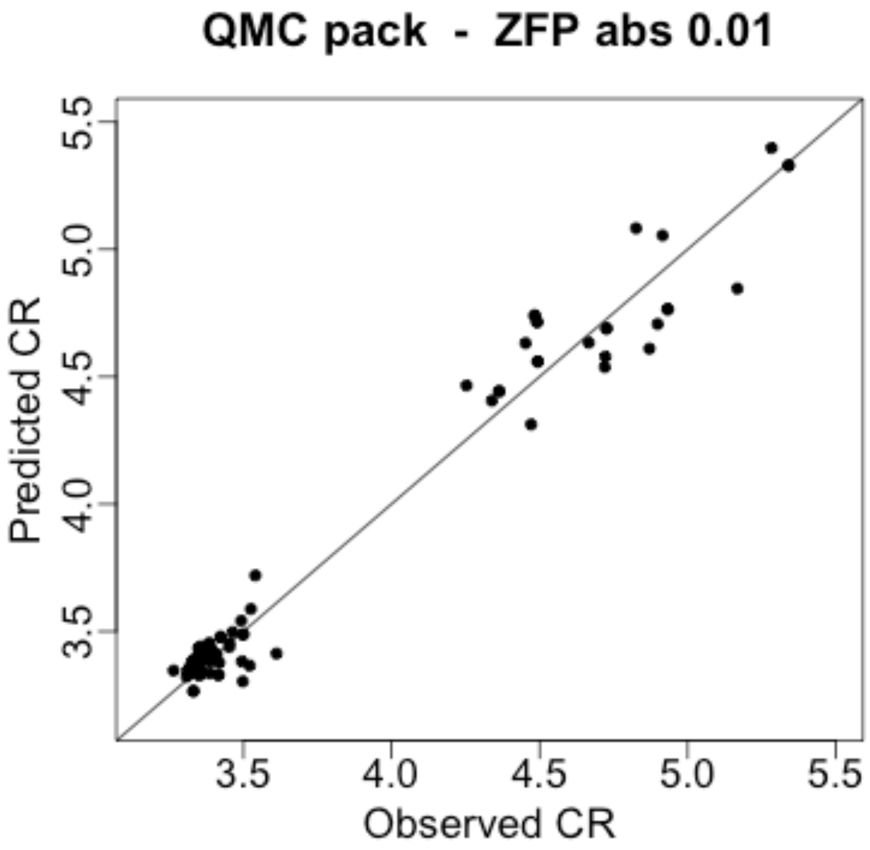
- Predicted CR exhibits low MAPEs (< 7.5%) for SZ2, ZFP, MGARD, and Bit Grooming

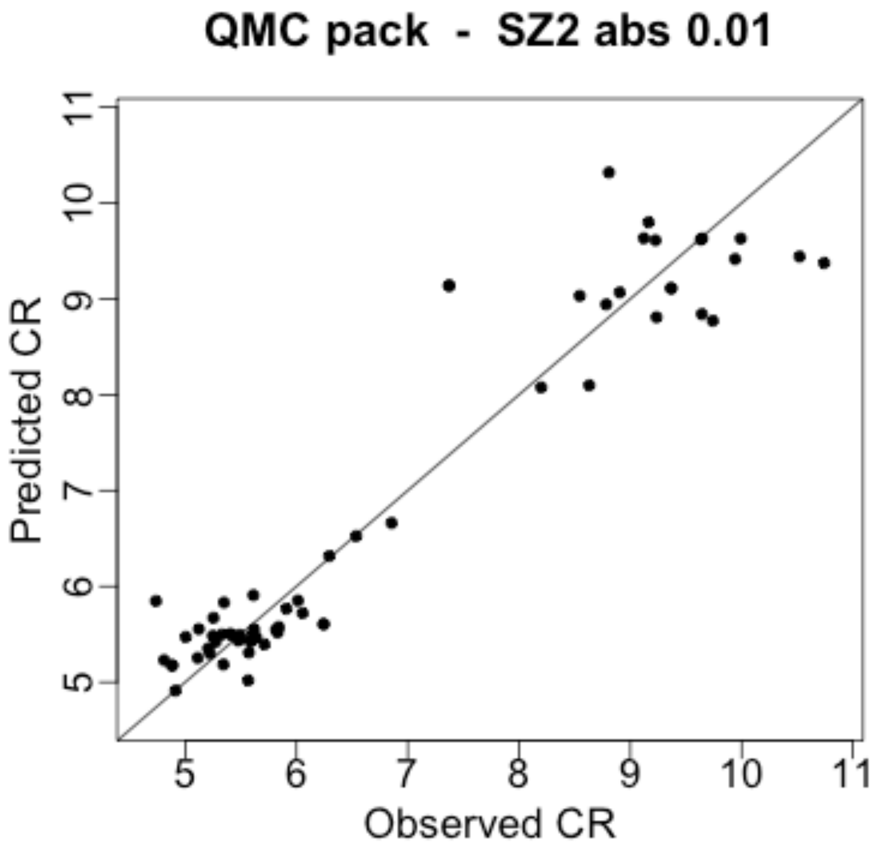
QMCPack compression estimation exhibited low error

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%

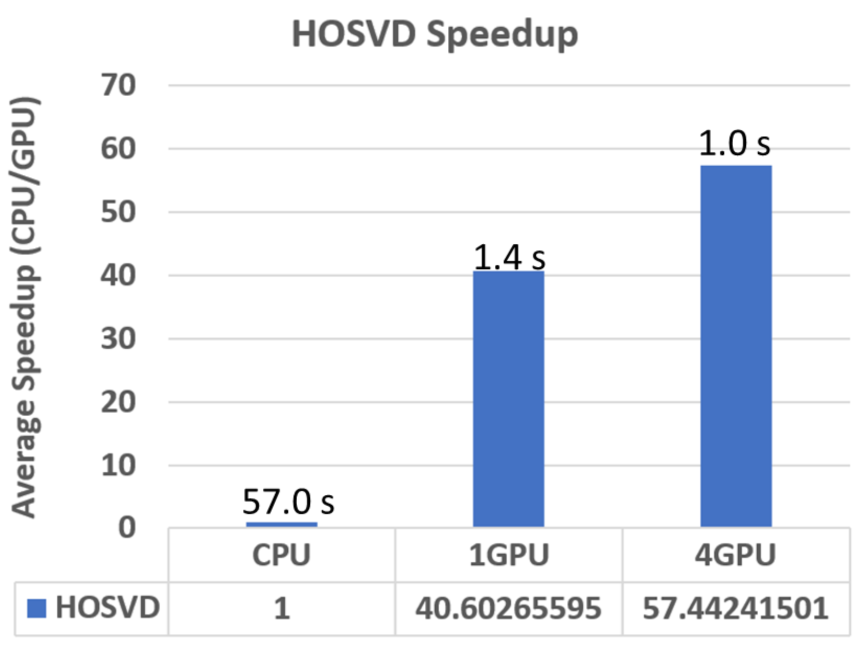
- Predicted CR exhibits low MAPEs (< 7.5%) for SZ2, ZFP, MGARD, and Bit Grooming
- However, TTHRESH produces a higher error

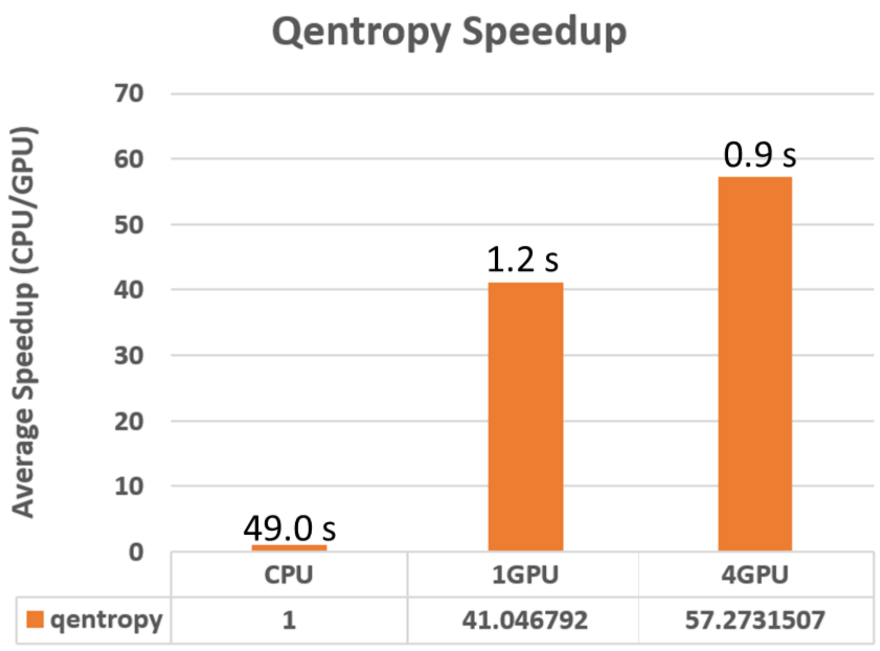
Predictions on cross validation set fit well





GPU acceleration improved performance (57x)

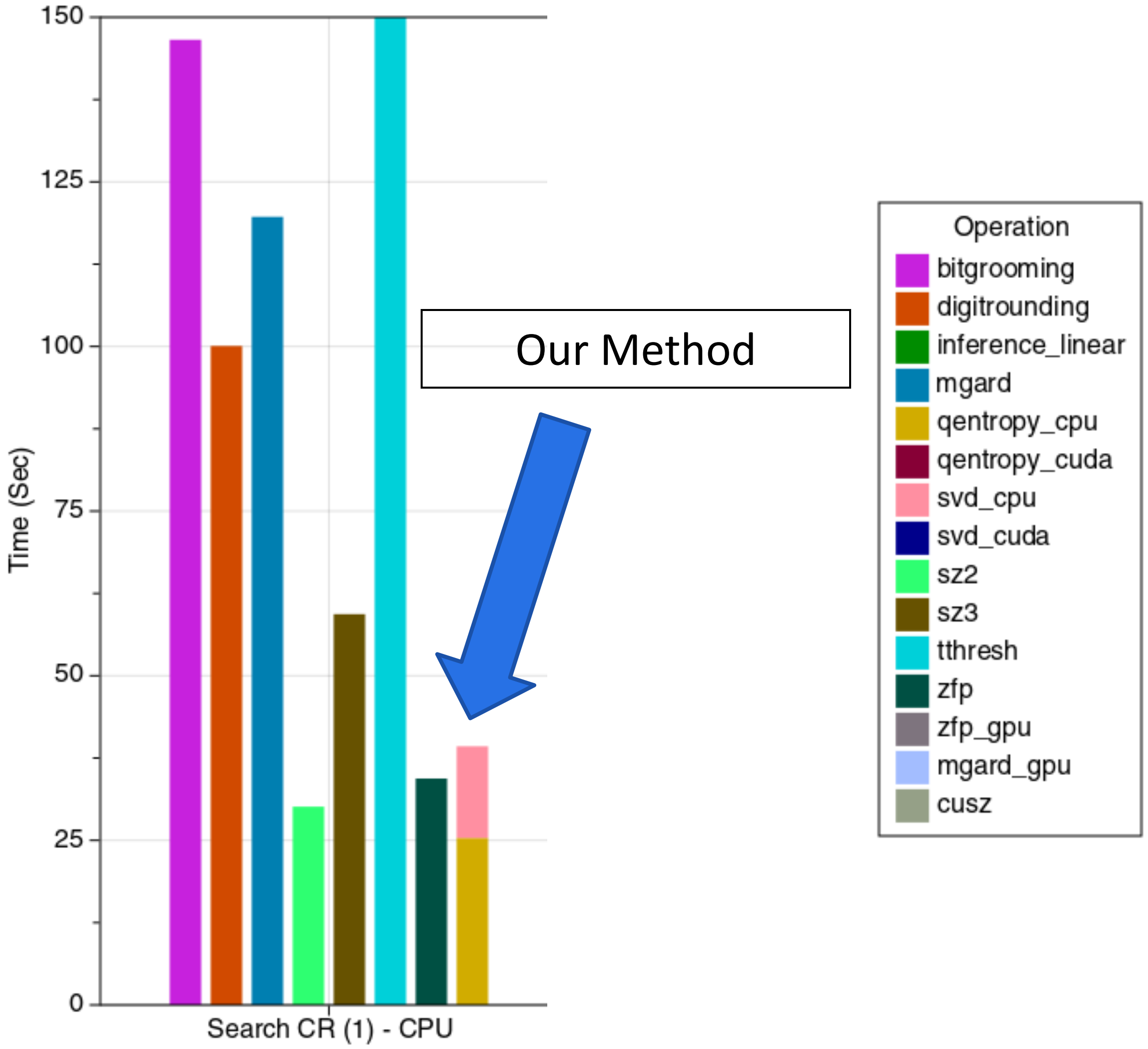




- Average performance of HOSVD and Q-ent on the ***Baryon density*** buffer from the NYX dataset

Statistical predictor reuse speeds up compressor comparisons

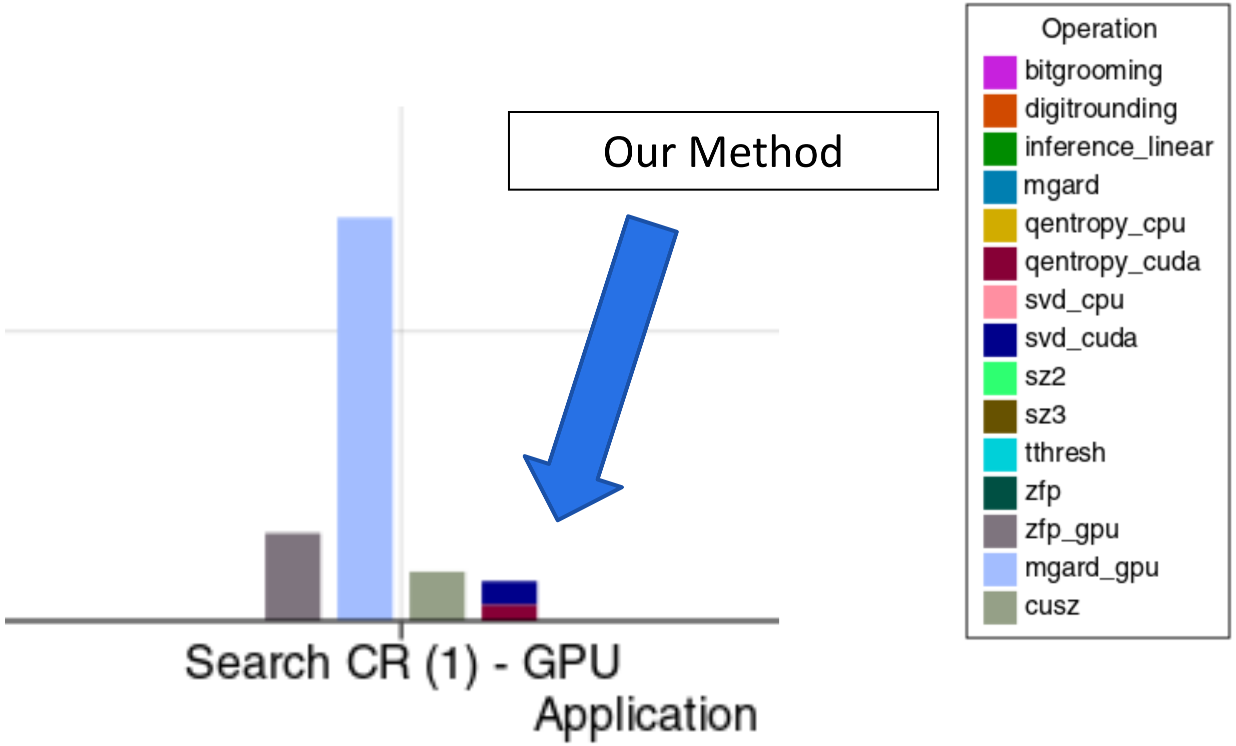
Time used to find compressor matching a CR



Statistical predictor reuse speeds up compressor comparisons

Time used to find compressor matching a CR

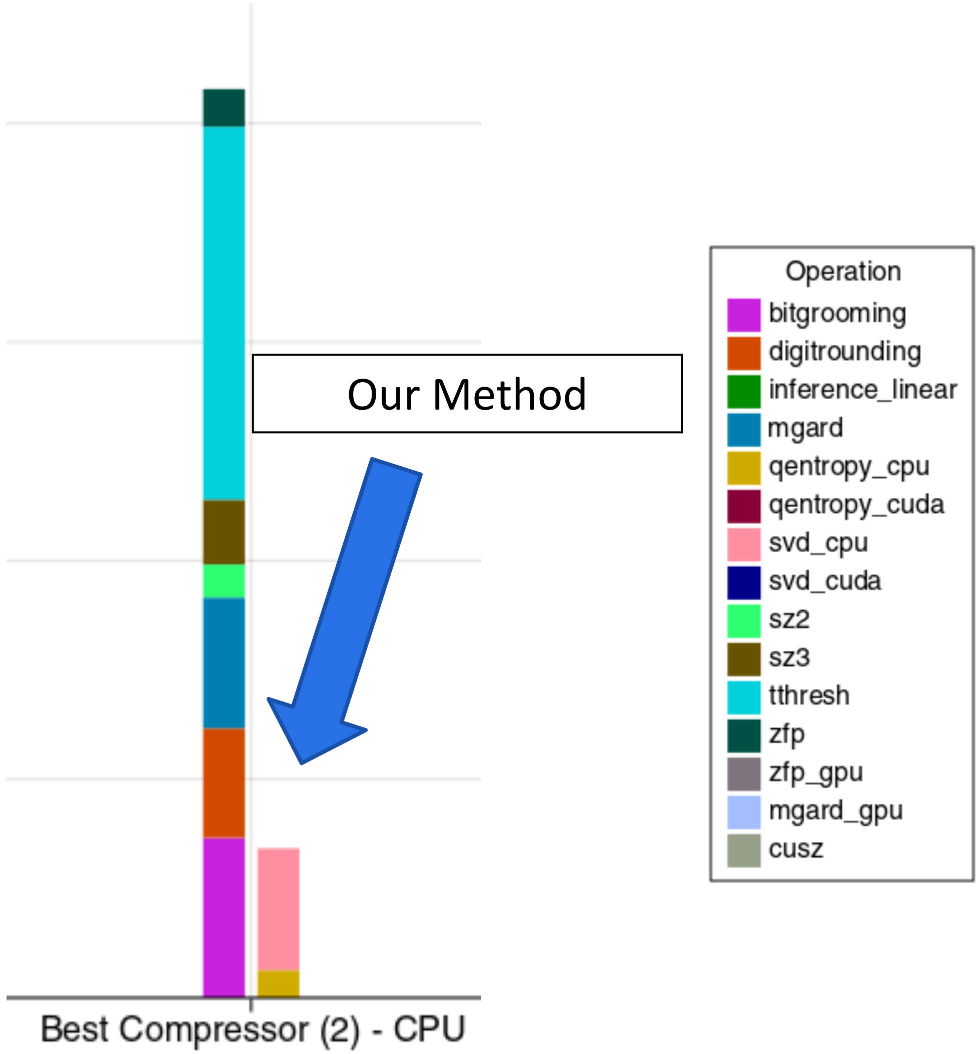
Our Method



Statistical predictor reuse speeds up compressor comparisons

Time used to find best compressor (highest CR)

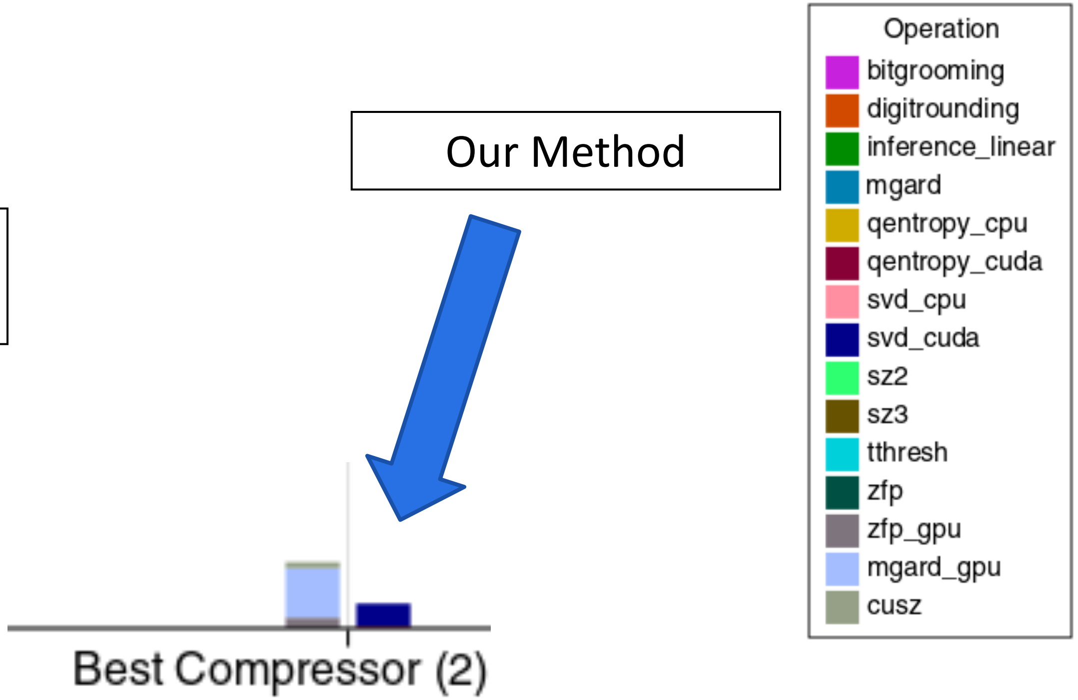
Our Method



Statistical predictor reuse speeds up compressor comparisons

Time used to find best compressor (highest CR)

Our Method



Introduction

Previous Work

Our Model

Results

Conclusion

Conclusions

1. Ability to accurately predict CRs for 3D scientific datasets
2. Flexible across compressors, error bounds, and datasets
   - Compressor-free statistical predictors
3. Statistical predictor reuse allows for comparison of different compressors to find largest CR
4. Performance speedup
   - Different predictors (variogram vs SVD)
   - Software methodology (OptZconfig vs regression model)
   - Hardware (CPU vs GPU)
5. Next step towards theoretical quantification of lossy compressibility

Future work

- Sampling-based approaches to reduce computational costs
  - Generate training samples from blocks of 3D tensor data
  - Estimate CR using the samples and our predictors
- Training free model for estimation

SC22
Dallas, TX | hpc accelerates.

QUESTIONS?

LINKTR.EE/KRASOW
DKRASOW@CLEMSON.EDU

CONTACT ME

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

Argonne NATIONAL LABORATORY

CLEMSON UNIVERSITY