Spline-Interpolation-Based Error-Bounded Lossy Compression for Scientific Data on GPUs

INDIANA UNIVERSITY BLOOMINGTON

author : Jiannan Tian (IU)
advisor : Dingwen Tao (IU)
Sheng Di (ANL)
Franck Cappello (ANL)

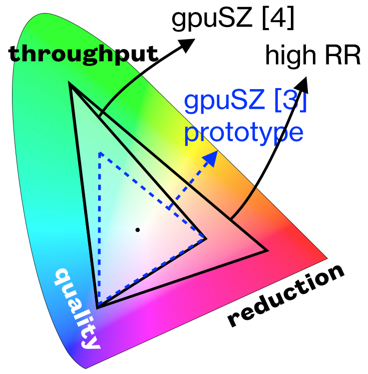
Argonne NATIONAL LABORATORY

Big-Data Scientific Application

application	data scale	to reduce
HACC cosmology simulation	**20 PB** per one-trillion-particle simulation	**10x** in need
CESM climate simulation	20% vs **50%** of h/w budget for storage 2013 vs **2017**	**10x** in need
APS-U High-Energy X-Ray Beams Experiments	hundreds of **PB** brain initiatives	**100x** in need

Data management is a real-world problem to address when we advance in scientific exploration.

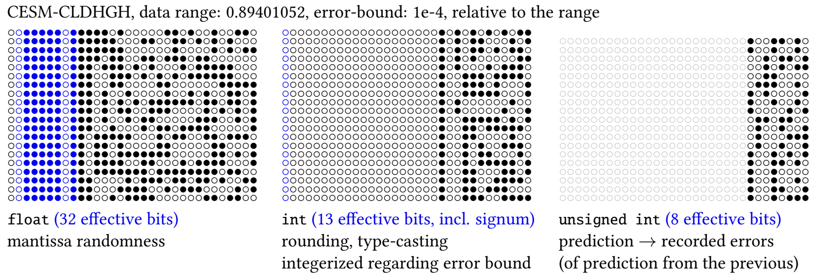
Data Reduction



+ **high-quality** for accurate post-analysis
+ data **reduction rate** at 10x in need
+ high-**throughput** processing to ease I/O and communication pressure

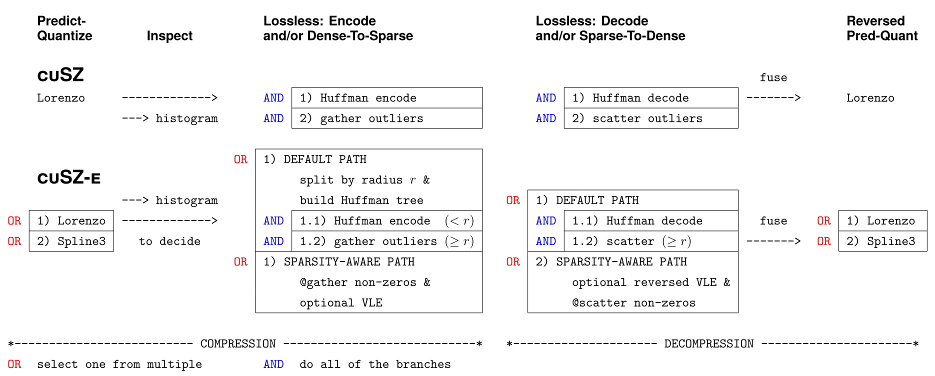
SZ [1, 2] Lossy Compression Essence

CESM-CLDHGH, data range: 0.89401052, error-bound: 1e-4, relative to the range



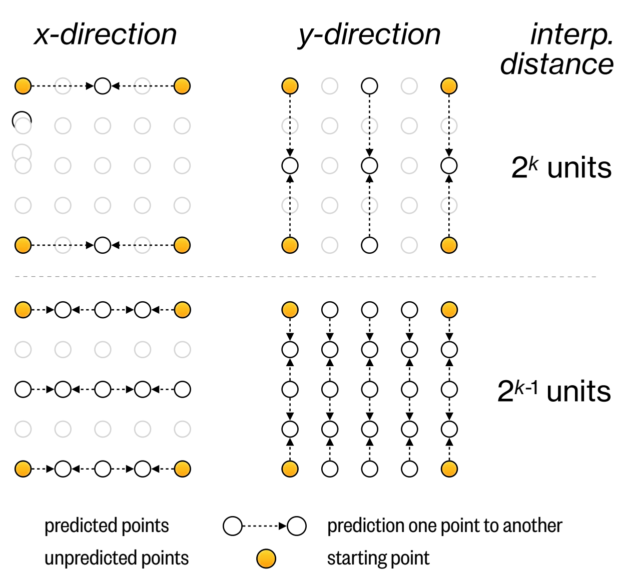
SZ, while guaranteeing the error-boundness, performs bit-level randomness elimination to increase the compressibility.

To lower bit randomness: prediction-based SZ.



When the prediction does well and results in highly sparse error-quantization code, gather (& scatter) are conducted to boost the compression ratio.

Interpolative Prediction Data-Access Pattern



We use 2D interpolation for conceptual demonstration; the prediction direction alters each stage.

An interpolation iteration contains *ndim* stages. Stages feature altering interpolation direction. When an iteration finishes, the interpolation distances shrinks by a factor of 2.

Starting points are distant from each other and beyond GPU thread blocks can handle; they become **anchor points** and are saved directly.

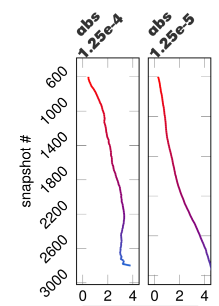
Preliminary Evaluation (Seismic Data)

Kernel Throughput

GB/s	compression
A100	253.5 (1.09x)
V100	232.7 (1.00x)
	decompression
A100	335.8 (1.39x)
V100	242.3 (1.00x)

Need further in-depth optimization.

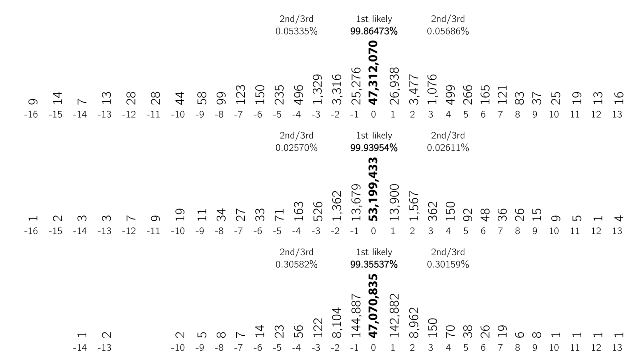
PSNR Advantage over Default



Spline interpolation outperforms the default predictor (Lorenzo).

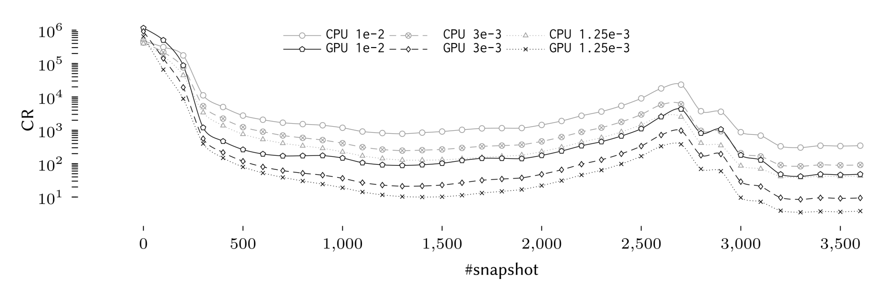
Dataset: RTM Simulation #600 to #3000.

Error-Quantization Distribution



GPU Spline (if w/o anchor) outperforms others in error-quantization centrality.

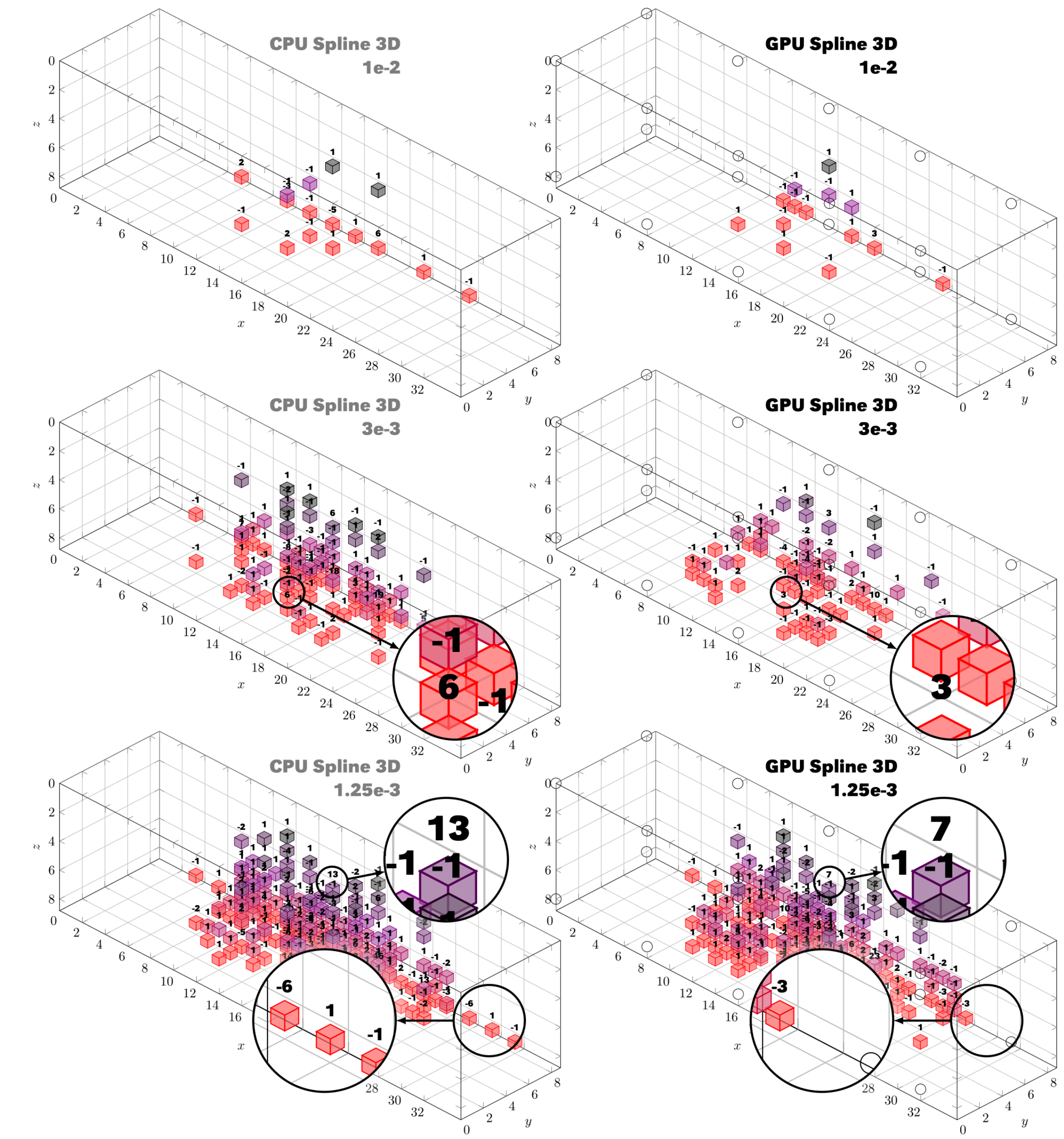
Compression Ratio Throughput 3,000 RTM Snapshots



Compression Ratio & Quality Across {GPU, CPU}-Spline and Default (Lorenzo)

		GPU Spline Prediction			CPU Spline Prediction			GPU Lorenzo Prediction	
eb	quant.	freq.	CR	quality	freq.	CR	quality	CR	quality
1.25e-3	-1	156,221 0.2935%	G 67.0		243,936 0.5149%	G 35.4			
	0	52,834,426 99.2538%	GA 59.3	PSNR 68.7	46,707,654 98.5888%	G+VLE 70.9	PSNR 69.7	G+VLE (lower)	PSNR 63.6
	-1	152,777 0.2870%	GA+VLE 118.5	corr. 0.9979	235,588 0.4973%	(upper limit•) 397.5	corr. 0.9983		corr. 0.9937
3.00e-3	-1	61,234 0.1150%	G 177.0		112,857 0.2382%	G 82.4			
	0	53,081,227 99.7175%	GA 131.5	PSNR 63.9	47,088,719 99.3931%	G+VLE 164.8	PSNR 63.6	G+VLE (lower)	PSNR 58.0
	-1	61,518 0.1156%	GA+VLE 263.0	corr. 0.9937	112,969 0.2385%	(upper limit•) 938.8	corr. 0.9932		corr. 0.9771
1.00e-2	-1	13,679 0.0257%	G 827.0		25,276 0.0534%	G 369.2			
	0	53,199,433 99.9395%	GA 316.2	PSNR 59.1	47,312,070 99.8646%	G+VLE 738.4	PSNR 56.2	G+VLE 77.6	PSNR 53.9
	-1	13,900 0.0261%	GA+VLE 632.5	corr. 0.9809	26,938 0.0569%	(upper limit•) 3606.7	corr. 0.962		corr. 0.9390

Impact on Error-Quantization



The color palette is used to increased the sense of space description. The larger z value, the more purple.

Interpolating 32x8x8 data chunk. With anchor points, the error quantizations are smaller in amplitude throughout different error bounds. Anchor points are used for performance concern: fitting subproblem size to GPU hardware architecture.

Acknowledgment

This R&D was supported by the Exascale Computing Project (ECP), Project Number: 17-SC-20-SC, a collaborative effort of two DOE organizations – the Office of Science and the National Nuclear Security Administration, responsible for the planning and preparation of a capable exascale ecosystem. This repository was based upon work supported by the U.S. Department of Energy, Office of Science, under contract DEAC02-06CH11357, and also supported by the National Science Foundation under Grants SHF-1617488, SHF-1619253, OAC-2003709, OAC-1948447/2034169, and OAC-2003624.

Reference

[1] S. Di and F. Cappello, "Fast error-bounded lossy HPC data compression with SZ," in *2016 IEEE International Parallel and Distributed Processing Symposium*, Chicago, IL, USA: IEEE, 2016, pp. 730–739.
[2] D. Tao, S. Di, Z. Chen, and F. Cappello, "Significantly improving lossy compression for scientific data sets based on multidimensional prediction and error-controlled quantization," in *2017 IEEE International Parallel and Distributed Processing Symposium*, Orlando, FL, USA: IEEE, 2017, pp. 1129–1139.
[4] J. Tian *et al.*, "cuSZ: An efficient gpu-based error-bounded lossy compression framework for scientific data," in *Proceedings of the ACM International Conference on Parallel Architectures and Compilation Techniques*, 2020, pp. 3–15.
[5] J. Tian *et al.*, "Optimizing Error-Bounded Lossy Compression for Scientific Data on GPUs." Proceedings of the 2021 IEEE International Conference on Cluster Computing, (Virtual Event) Portland, OR, September 7–10, 2021.