Lossy compression for transferring, storing, analyzing huge scientific datasets

Sheng Di, Franck Cappello

MCS division

Argonne National Laboraotry

04/14/2020

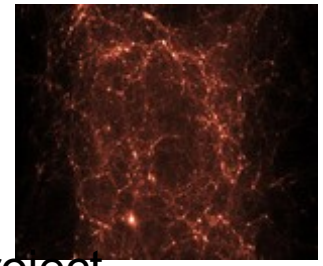
EXASCALE COMPUTING PROJECT

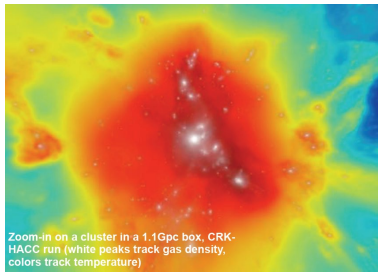
U.S. DEPARTMENT OF ENERGY | Office of Science

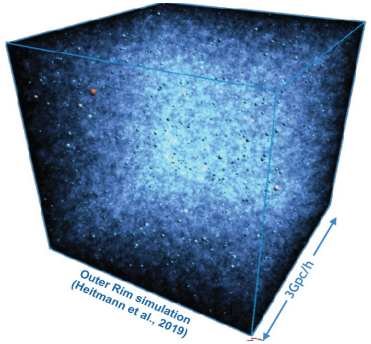
Argonne NATIONAL LABORATORY

Background and Motivation – When the scientific data becomes too big ….

Today's scientific simulations are producing extremely large volumes of data – too large to save, process, and analyze

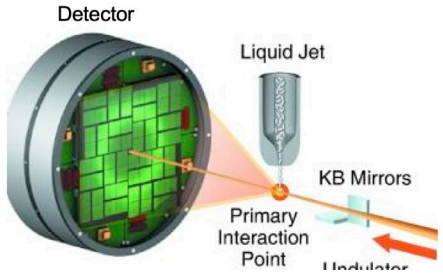






Cosmology simulation:

- A total of **>20PB** of data when simulating 1 trillion of particles (500 snapshots)
- Petascale systems never gave 20PB for one project
- On current file system (1TB/s), storing the 20PB may take 20X10^15 seconds (**5h30**).



Light source data (material science):

- Today: All LCLS-II data detectors per experiment: **250 GB/s**
- Square Kilometer Array (SKA) will generate **300PB/year**; HL-LHC will generate **1EB** in 2026.

Challenges:

Transferring the data through the network or storing the data on file systems become serious performance bottlenecks.

SZ: Error-bounded lossy compression

Argonne NATIONAL LABORATORY

Scientific Achievement

- SZ can significantly reduce the data size from simulations and instruments while respecting user accuracy requirements.
- SZ supports multiple I/O libraries (HDF5, ADIOS, etc) and different parallel models (MPI, multi-core, FPGA, GPU, etc)

Significance and Impact

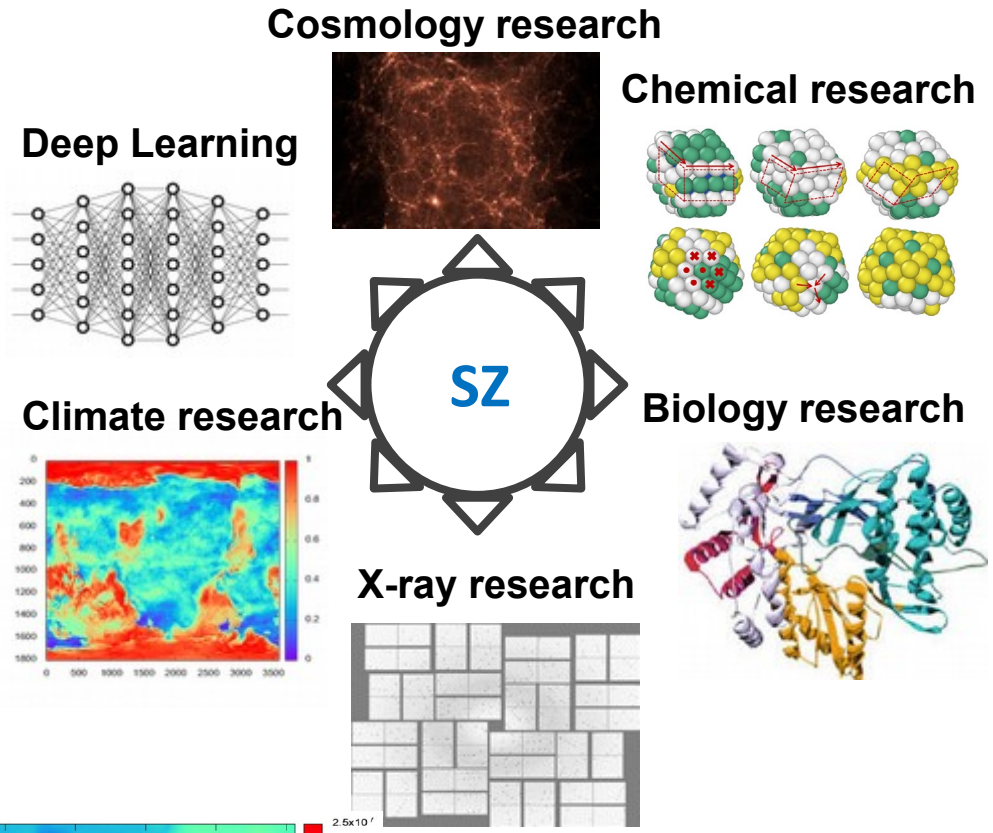
- SZ has been integrated into multiple scientific applications
- SZ has many use-cases: reducing memory & storage footprint, accelerating I/O and computation, reducing streaming intensity, etc.
- SZ has been evaluated/used by 20+ institutes/universities/companies
- More than 400 citations in 2016-2019.

Research Details

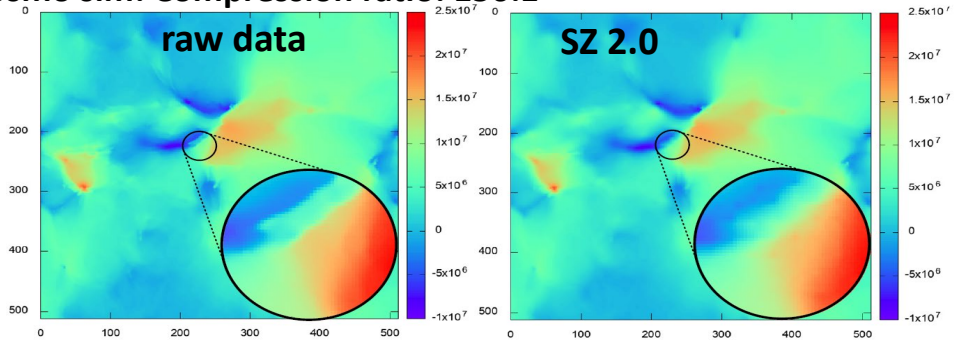
SZ compressor:

https://collab.cels.anl.gov/display/ESR/SZ

30+ papers have been published in prestigious conferences/journals.

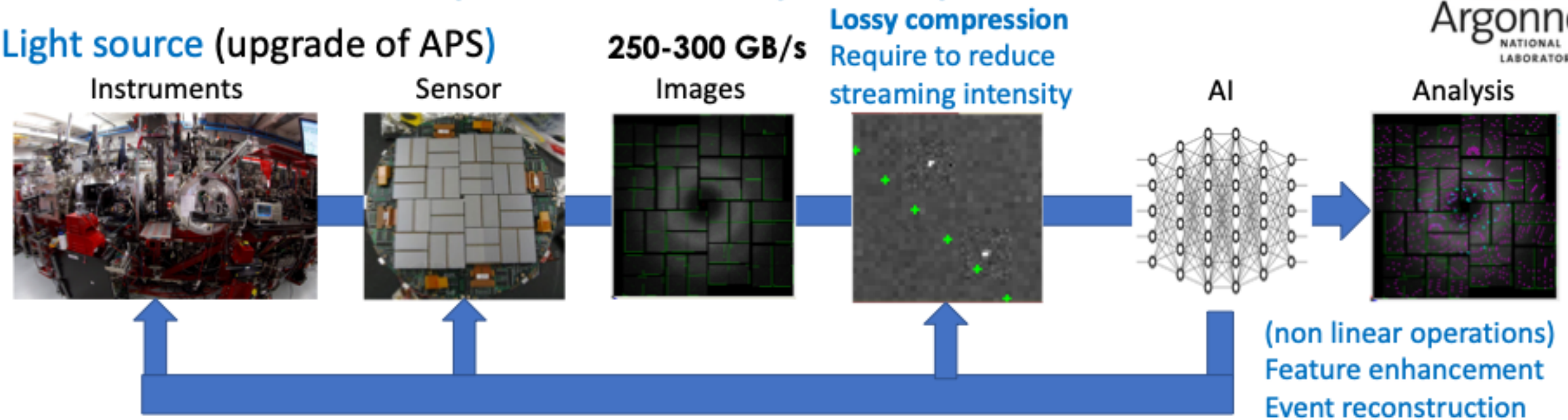


NYX cosmo sim: Compression ratio: 156:1



2 concrete examples of Lossy compression + AI

Light source (upgrade of APS)



ECP CANDLE Drug response problem (DRUG)

To predict drug response based on molecular features of tumor cells and drug descriptors

Ideally:

Drug: molecular structures, interaction, combination, molecular target

Patient genetics: baseline genotype, specific genetics, molecular and cellular tumor properties, etc.

Initial formulation drug response problem

10 year problem target: 10PB training set

Research Opportunities and Challenges

Many open research questions and high impact research opportunities:

1. How to optimize the compression for specific use-cases (storing data, transferring data, etc.);
   - Challenge: diverse requirements in different use-cases
2. Compression quality/performance depends on parameters/settings
   - Challenge: Autotuning parameters of compression is a non-convex optimization problem.
3. How to assess the impact of data distortion to user's analysis
   - Challenge: No standard/criterion because of diverse applications, metrics, analysis.
4. How to control data distortion to respond to user requirements
   - Challenge: Hard to establish a link between user analysis and acceptable data distortion
5. How to acclerate lossy compressors on GPUs, FPGA and ASICs
   - Challenge: advanced high-performance compression pipelines are quite complex

Selected papers published recently

Argonne NATIONAL LABORATORY

- J. Tian, S. Di, C, Zhang, Xin Liang, S, Jin, D, Cheng, D, Tao, and F. Cappello, "waveSZ: A Hardware-Algorithm Co-Design of Efficient Lossy Compression for Scientific Data", Proceedings of the 25th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (**ACM PPoPP2020**), San Diego, California, USA, February 22-26, 2020.
- R. Underwood, S. Di, J. Calhoun, F. Cappello, "FRaZ: A Generic High-Fidelity Fixed-Ratio Lossy Compression Framework for Scientific Floating-point Data", in Proceedings of the 34th IEEE International Parallel and Distributed Symposium (**IEEE IPDPS2020**), New Orleans, LA, May 18-22, 2020.
- X. Liang, S. Di, D. Tao, S. Li, B. Nicolae, Z. Chen, F. Cappello, "Improving Performance of Data Dumping with Lossy Compression for Scientific Simulation", in **IEEE CLUSTER2019**, 2019.
- X. Wu, S. Di, E. Maitreyee Dasgupta, F. Cappello, Y. Alexeev, H. Finkel, F. T. Chong, "Full State Quantum Circuit Simulation by Using Data Compression", in IEEE/ACM 30th The International Conference for High Performance computing, Networking, Storage and Analysis (**IEEE/ACM SC2019**), 2019.
- X. Liang, S. Di, S. Li, D. Tao, B. Nicolae, Z. Chen, F. Cappello, "Significantly Improving Lossy Compression Quality based on An Optimized Hybrid Prediction Model", in IEEE/ACM 30th The International Conference for High Performance computing, Networking, Storage and Analysis (**IEEE/ACM SC2019**), 2019.
- D. Tao, S. Di, X. Liang, Z. Chen and F. Cappello. Optimization of Fault Tolerance for Iterative Methods with Lossy Checkpointing. in 27th ACM Symposium on High-Performance Parallel and Distributed Computing (**ACM HPDC2018**), 2018.
- X. Liang, S. Di, D. Tao, S. Li, S. Li, H. Guo, Z. Chen, F. Cappello, "Error-Controlled Lossy Compression Optimized for High Compression Ratios of Scientific Datasets", in **IEEE Bigdata2018**, 2018.
- X. Liang, S. Di, D. Tao, Z. Chen, F. Cappello, "Efficient Transformation Scheme for Lossy Data Compression with Point-wise Relative Error Bound", in **IEEE CLUSTER 2018**. (**best paper**)
- A. M. Gok, S. Di, Y. Alexeev, D. Tao, V. Mironov, F. Cappello, "PaSTRI: Error-bounded Lossy Compression for Two-Electron Integrals in Quantum Chemistry", in **IEEE CLUSTER 2018**, 2018. (**best paper**)
- S. Li, S. Di, X. Liang, Z. Chen, F. Cappello, "Optimizing Lossy Compression with Adjacent Snapshots for N-body Simulation", in IEEE Bigdata2018, 2018.
- D. Tao, S. Di, Z. Chen, and F. Cappello. In-Depth Exploration of Single-Snapshot Lossy Compression Techniques for N-Body Simulations. Proceedings of the 2017 IEEE International Conference on Big Data (**IEEE BigData2017**), Boston, MA, USA, December 11 - 14, 2017.
- 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. **IEEE IPDPS17**, May 2017
- D. Tao, S. Di, H. Guo, F. Cappello, Z-checker: A Framework for Assessing Lossy Compression of Scientific Data, International Journal of High Performance Computing Applications, **IJHPCA**, Sage Publishing, forthcoming, 2017.
- S. Di, F. Cappello. Optimizing Error-Bounded Lossy Compression for Hard-to-Compress HPC Data. in IEEE Transactions on Parallel and Distributed Computing **IEEE TPDS,** 2017**.**
- S. Di, F. Cappello. Fast Error-Bounded Lossy HPC Data Compression with SZ, **IEEE IPDPS16**, 2016**.**