ORNL/LTR-2022/412

The Exploitation of Data Reduction for Visualization

Kenneth Moreland
David Pugmire
Jieyang Chen

August 2022

ORNL IS MANAGED BY UT-BATTELLE LLC FOR THE US DEPARTMENT OF ENERGY

ORNL/LTR-2022/412

Computer Science and Mathematics Division

The Exploitation of Data Reduction for Visualization

Submitted to the ASCR Workshop on Visualization for Scientific Discovery, Decision-Making, & Communication

Kenneth Moreland, David Pugmire, Jieyang Chen

August 2022

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-Battelle LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

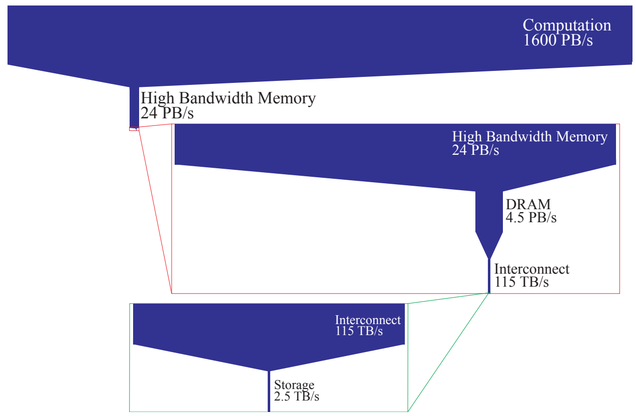

Figure 1. A plot of the relative bandwidth of system components in the Summit supercomputer. (Adapted from Bauer, et al. [3].)

Topics 3c – Extreme scale data (primary), 2a – Uncertainty visualization (secondary)

Challenge

The disparity between the computational speed and storage bandwidth, as demonstrated in Figure 1, is a well known problem that grows with each successive generation. The visualization community is principally responding to this issue by using in situ to reduce which data must be written to storage. However, other communities are taking different, possibly complementary approaches. In particular, data compression is a common general approach to reduce storage demands.

Data compression technologies are typically not designed with post processing in mind. The principal metrics measured are compression ratio, the improved bandwidth to storage, and the error introduced. It is assumed that data is inflated to its full size before any post processing can happen. Although when talking about bandwidth disparities, HPC's dirty little secret is that no part of the memory nor interconnect hardware is increasing at the rate of computation. For example, the Summit supercomputer has a peak computation rate almost 10 times its predecessor, Titan, but only about 4 times the memory, less than twice the aggregate memory bandwidth, and almost no improvement in the interconnect bisection bandwidth [6]. Naively inflating data for post processing does not help with limitations in the memory and interconnect systems.

In practice, scientific data is always compressed with some loss because lossless techniques tend to have poor performance [1]. Unfortunately, any loss adds uncertainty to the data. When aggressive data compression is performed, it is important to be able to measure and represent the uncertainty added. Making the situation more complicated, the uncertainty of the data compression interacts with the uncertainty already present in the data being compressed. The uncertainty of the compression needs to be composed with other sources of uncertainty.

Opportunity

There is an obvious opportunity for transferring data for in situ, in transit, and post processing components. Adoption of data compression will improve data transfer among visualization components and the components they interact with. However, this is just scratching the surface of possibilities.

As disparity between execution and memory bandwidth increases, forward thinking research has speculated the idea of computing when FLOPs are "free." FLOPs are not free, but roofline analysis often shows that memory access can bottleneck computation [7]. Visualization and analysis by its nature performs data-centric operations with low computation to data access ratios. Incorporating data compression in the visualization process could potentially relieve data access.

Compressed representations have structure that could potentially be exploited for more efficient visualization and analysis. Transform-based data reduction techniques such as ZFP and MGARD have been designed for the past decade. One of the advantages of this kind of data reduction technique is that they create an alternative representation of the original data for much more flexible data retrieval and reconstruction. This brings potential opportunities for building more efficient visualization and analysis. For example, ZFP breaks data into blocks that can be streamed across memory spaces [5]. MGARD arranges data in hierarchies of increasing refinement [2]. Such representations can potentially accelerate visualization by reducing not only the amount of data to be transferred but also the amount of computation, since approximation of the original data with reduced degrees of freedom can be constructed. The major advantage over traditional sampling based data reduction is that potentially important information content is prioritized in the process of data retrieval and reconstruction.

Such potential capability has a broader impact on scientific visualization and analysis workflows, which can dramatically help accelerate scientific discovery. For example, accelerated visualization can allow domain scientists to visualize data on a higher spatial-temporal resolution with the same amount of computing resource or enable visualizing large-scale datasets on edge systems near scientific instruments for near real time decision making. However, accelerating visualization through data reduction techniques is not straightforward. Extensive research is necessary for designing transformation methods for data reduction and optimizing visualization algorithms that can effectively take advantage of reduced data representations.

Currently the uncertainty in the data and the error of lossy compression are treated independently. There are opportunities to exploit uncertainty to adjust the aggressiveness of lossy compression. A lossy compression needs only to store the bits that are part of the significant figures. Thus, an uncertain region of data may need less precision in its representation. Likewise, an uncertain region may need less refinement in the analysis.

Timeliness

For the aforementioned reasons and more, DOE has a high interest in data reduction. There was a Data Reduction for Science ASCR Workshop [4] held in January 2021 and a subsequent FOA regarding data reduction later in the year. Consequently, we can expect a significant amount of innovation in data reduction and compression in the coming years. Not investing in the effect of data compression on visualization and analysis will be at best a large missed opportunity. At worst, DOE's data reduction technologies will lead to stored compressed data that cannot be analyzed.

References

[1] Mark Ainsworth, Scott Klasky, and Ben Whitney. Compression using lossless decimation: Analysis and application. *SIAM Journal on Scientific Computing*, 39(4):B732–B757, 2017.

[2] Mark Ainsworth, Ozan Tugluk, Ben Whitney, and Scott Klasky. Multilevel techniques for compression and reduction of scientific data—the multivariate case. *SIAM Journal on Scientific Computing*, 41(2):A1278–A1303, 2019.

[3] Andrew C. Bauer, Hasan Abbasi, James Ahrens, Hank Childs, Berk Geveci, Scott Klasky, Kenneth Moreland, Patrick O'Leary, Venkatram Vishwanath, Brad Whitlock, and E. Wes Bethel. In situ methods, infrastructures, and applications on high performance computing platforms. *Computer Graphics Forum*, 35(3):577–597, June 2016.

[4] Scott Klasky, Jana Thayer, and Habib Najm. Data reduction for science: Brochure from the advanced scientific computing research workshop. USDOE Office of Science (SC), Advanced Scientific Computing Research (ASCR), April 2021.

[5] Peter Lindstrom. Fixed-rate compressed floating-point arrays. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2674–2683, December 2014.

[6] Sudharshan S. Vazhkudai et al. The design, deployment, and evaluation of the CORAL pre-exascale systems. In *SC'18 Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, November 2018.

[7] Samuel Williams, Andrew Waterman, and David Patterson. Roofline: An insightful visual performance model for multicore architectures. *Communications of the ACM*, 52(4):65–76, April 2009.