Optimizing Compression Schemes for Parallel Sparse Tensor Algebra

Helen Xu*, Tao B. Schardl†, Michael Pellauer‡, and Joel S. Emer†‡

*Lawrence Berkeley National Laboratory
†Massachusetts Institute of Technology
‡NVIDIA

hjxu@lbl.gov,neboat@mit.edu,mpellauer@nvidia.com,jsemer@mit.edu

This paper studies compression techniques for parallel in-memory sparse tensor algebra. Although one might hope that sufficiently simple compression schemes would generally improve performance by decreasing memory traffic when the computation is memory-bound, we find that applying existing simple compression schemes can lead to performance loss due to the additional computational overhead. To resolve this issue, we introduce a novel algorithm called *byte-opt*, an optimized version of the *byte* format from the Ligra+ graph-processing framework [1] that saves space without sacrificing performance. The byte-opt format takes advantage of per-row structure to speed up decoding without changing the underlying representation from byte.

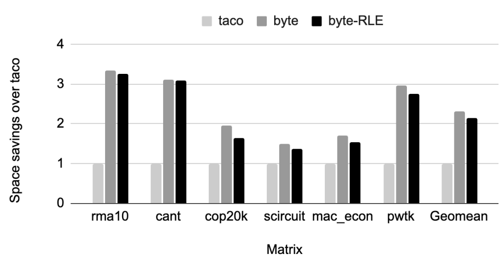

Figure 1: Space savings over original taco CSR.

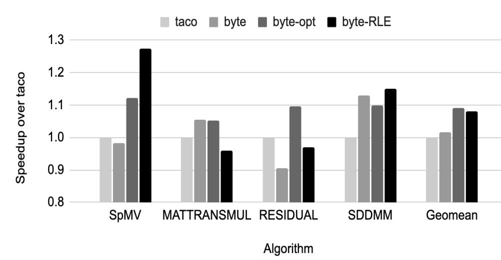

Figure 2: Speedup over original taco CSR when run on 48 hyperthreads.

We evaluate the byte, byte-opt, and *byte-RLE* [1] formats on top of a suite of sparse matrix algorithms generated by taco [2], a compiler for sparse tensor algebra. The byte-RLE format takes advantage of per-row structure, but changes the byte format to improve performance at the cost of space. Figure 1 shows that, on average, the byte and byte-RLE formats are 2.3× and 2.1× smaller, respectively, than CSR. Meanwhile, Figure 2 shows that, although the encoded formats are on average about 1.1× faster than CSR, some algorithms are substantially slower when using byte and byte-RLE compared to CSR. In contrast, algorithms using byte-opt are always faster than the baseline while achieving the same space savings as byte.

References

[1] J. Shun et al., "Smaller and faster: Parallel processing of compressed graphs with Ligra+," in *DCC*, 2015.

[2] F. Kjolstad et al., "The tensor algebra compiler," *OOPSLA*, 2017.

979-8-3503-4795-1/23/$31.00 ©2023 IEEE
DOI 10.1109/DCC55655.2023.00084

Authorized licensed use limited to: Nvidia Corp. Downloaded on August 29,2023 at 14:30:45 UTC from IEEE Xplore. Restrictions apply.

2023 Data Compression Conference (DCC) | 979-8-3503-4795-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/DCC55655.2023.00084