Summary of Difference

October 18, 2018

This paper is extension of our conference paper at PPoPP '18 [1]. This paper makes the following distinct contributions compared with the conference paper.

1. We propose a performance model called *Framework Resources Quantification* (FRQ) model. The FRQ model displays utilization of resources in the time dimension and calculates execution time of computing jobs on different DAG frameworks. We use the FRQ model to assist in analyzing the shuffle process and verify SCache shuffle optimization by mathematics.

2. In the conference paper, we only implemented SCache on Spark. To prove the compatibility of SCache as a cross-framework plug-in, we also implemented SCache on Hadoop MapReduce. Our experiments show that SCache can also optimize the computing performance of Hadoop MapReduce.

3. We append two parts in the evaluation section. Firstly, we verify the FRQ model in both our in-house environment and Amazon EC2 environment. The error between the calculated value and the experimental value is basically below 10%. Secondly, we evaluate the performance of Hadoop MapReduce with SCache in the same environment as the conference paper. After utilizing pre-fetching of SCache, Hadoop MapReduce with SCache optimize job completion time by up to 15% and an average of 13%.

Besides the major contributions, we also revise the paper in many places, including writing and more related work. We believe that the added content makes sufficient contribution to this journal submission.

References

[1] Z. Fu, T. Song, Z. Qi, and H. Guan, "Efficient shuffle management with scache for dag computing frameworks," in *Proceedings of the 23rd ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming*. ACM, 2018, pp. 305–316.