Summary of Difference

October 19, 2018

This paper is the 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 new performance model called *Framework Resources Quantification* (FRQ) model. The proposed model is not only to predict the job completion time, but also to theoretically analyze DAG frameworks. The FRQ models DAG frameworks according to the resource scheduling strategy of each DAG framework by mutiple parameters. Through modeling, the FRQ model can indicate which parameters of the DAG framework are important and which parts of the DAG framework can be optimized. Furthermore, the FRQ models can classify DAG frameworks based on their scheduling stratagies. According to the theoretical analysis results of the FRQ model, we are able to optimize DAG frameworks more accurately.

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

3. We append two parts to the evaluation section. Firstly, we evaluate the FRQ model in both our in-house environment and Amazon EC2 environment. The error between the FRQ's calculated value and the experimental value is below 10%. Secondly, we evaluate the performance of Hadoop MapReduce with SCache. A in the same environment as the conference paper, i.e., 50-node Amazon EC2 cluster. After utilizing pre-fetching of SCache, Hadoop MapReduce with SCache optimize job completion time by up to 15% and an average of 13%.

In summary, we propose a new performance model to provide theoretical basis for performance optimization and improve experiments by adding experiments on Hadoop MapReduce. We believe that the added content makes a 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.