Paper review [1]

Yu SU 55323707

1. The Problem

In these years, to reduce the file system IO overhead, the traditional approach is using caching. However, the caching cannot improve the performance of write operator because of the replication mechanism, which is essential in the distributed system. And the paper proposed an in-memory storage system called Tachyon, which could realize the high throughput operation with the good fault tolerance.

2. Challenge

There are two main challenges. First, when there is a long-running job, it is hard for Tachyon to recompute it in a bound recovery time, because the storage layer is independent to the job, it is hard to set periodic checkpointing. Second, it is how to allocate resources for recomputations, which means that there ought to be a fair priority mechanism to make sure the system could keep efficiency with the failure.

3. Key Insight

The first key insight is the special architecture of the system. Tachyon consists of lineage and persistence, two layers, which provide two kinds of approach for storage, with lineage or not. In detail, there is a workflow manager in the master to track the lineage information, compute checkpoint order and interact with resource manager. For workers, they could interact with the data at memory speeds by the RAMdisk.

Focusing on each parts, the paper proposed a checkpoint algorithm called edge algorithm, which enable asynchronously checkpoint in the background without stalling writes. In detail, it provides bounded recomputation time. And to reduce the extra overhead when do checkpointing, the algorithm could only checkpoint the hot files but avoid checkpointing temporary files.

Another key sight is the resource allocation strategy of Tachyon. Within this strategy on the scheduler, the system provides priority compatibility, which could make sure the important jobs will complete. And they also give resource sharing to increase the utilization of resource. What's more, it could also avoid cascading recomputation, which means the recomputation will consider about the data dependencies and avoid the recursive job launching.

4. Limitation

One of the limitations is the Tachyon focus on the memory management and does not fit the CPU or network intensive jobs. What's more, although the Tachyon give a new resource allocation strategy, it still cannot jump out the restriction of memory size. It looks like a caching system without a good hierarchical storage.

5. Future Work

Although the Tachyon is great itself, it lacks good storage abstraction and common file API, so I think Tachyon could be combined with the Resilient Distributed Dataset. Because the biggest problem of RDD is the memory management problem, which is just the Tachyon is good at. What's more, with Tachyon, the RDD could leave the memory management part out and increase the computation stability.

[1] L. Haoyuan, G. Ali, Z. Matei, S. Scott, S. Ion (2014). Tachyon: Reliable, Memory Speed Storage for Cluster Computing Frameworks. *Proceedings of the ACM Symposium on Cloud Computing – SOCC 14. doi:10.1145/2670979.2670985*