Paper review [1]

Yu SU 55323707

1. The Problem

The problem is workload in resource scheduler is always a big challenge for cloud system. Although there have been lots of efficient resource management method, little understand the underlying workload and practical operation demand. Therefore, the paper proposed some analysis on the google trace to figure out some of the characters of the schedulers

2. Challenge

The main challenge is past analysis on the trace often focus on the homogeneous workloads because there is not the diverse workloads issue in most of the previous clusters. While, the paper examined a mix of long-running services, DAG-of-task systems and high-performance computing. There are different challenges for different categories, so it is even harder when all these are combined.

3. Key Insight

There are four types of scheduler the paper analyzed.

First is the machine and workload heterogeneity and variability, because the cluster machines are not homogeneous. And the paper proposed analysis on the workload types, job durations, task shapes and the distributions.

Second is the highly dynamic resource demand and availability, because with the increase of the number of organizational entities under a common computing infrastructure, the more dynamic their aggregated demand and arrival patterns are likely to be. As a result, the paper did analysis on different job submission circumstance, including the crash-loops, small jobs and the evictions.

Third is the predictable but poorly predicted resource needs, because it is hard to predict future resource allocation accurately. And the paper got the statement of usage on the resource requests, such as non-automation, request accuracy and outliers within tasks.

Last is the resource class preferences and constraints because these will cause the lack of utilization and latency, so the paper analyzed the influence factor about the constraints such as constraint-induced scheduling delay and locality.

4. Limitation

I think one of the most limitation for this paper is all the analysis is under the Google trace. That means the result may be biased by the Google trace system instead of a normal result for the cloud system. The second limitation is the paper said this analysis is to provide a guidance for cloud resource scheduler designs, while what the paper actual provided is the specific number under the common sense. This did not solve the real bottleneck for the resource allocation.

5. Future Work

In the future, the analysis could be extended to other platform and balance all the result to propose a more normal result, which could make sure the correctness and scalability of these result.

[1] R. Charles, T. Alexey, R. Ganger, H. Randy, A. Michael (2012). Heterogeneity and dynamicity of clouds at scale: Google trace analysis. *Proceedings of the 3rd ACM Symposium on Cloud Computing Article No.7*