Abstract

Public IaaS clouds have become an essential infrastructure for enterprises and research organizations to run applications/services because the public clouds offer resource diversity, scalability, and elasticity as well as cost efficiency. Predictive resource management systems such as Netflix’ Scryer are developed to efficiently leverage such cloud infrastructures with two interrelated goals: maximizing SLA (Service Level Agreement) satisfaction and minimizing execution cost. However, existing predictive approaches are not sufficient to meet these two goals due to two uncertainties in IaaS clouds -- workload uncertainty and performance uncertainty -- and often show poor performance and adaptability in predicting future workloads and guaranteeing performance SLA. As a result, existing methods incur frequent SLA violations and require high execution cost.

In this presentation, I will talk about two techniques that mitigate these two uncertainties. I will first introduce a novel prediction framework called "CloudInsight." CloudInsight is designed to address the real-world cloud workloads and leverages the combined power of multiple workload predictors. I will discuss how to create an online ensemble model using multiple predictors and how to determine unique contribution (weight) of each predictor for maintaining high prediction accuracy under dynamic workload fluctuations. I then discuss how CloudInsight improves general cloud metrics (e.g., cost efficiency and SLA satisfaction rate) in the resource management. In the second part of the talk, my presentation focuses on a mechanism to guarantee the performance SLA of cloud applications. I will introduce "Orchestra" framework that dynamically controls multiple applications/tasks in the user space and aims at meeting corresponding SLAs for multiple applications on the same cloud instance. I will discuss core mechanisms of Orchestra such as lightweight application performance models as well as runtime optimization/control of shared resource allocation to meet different applications' SLAs. I then discuss how effectively Orchestra guarantees performance SLAs without extra cloud costs.