Abstract:

Cloud computing provides a virtualized platform for running various services, including soft real-time applications such as video streaming. To meet application real-time requirements, system resource is often over allocated and results in system underutilization. To solve this problem, we present, ANCHORS, a framework for users to implement any dynamic resource allocation algorithm that utilizes application real-time performance feedback. We then present two resource allocation algorithms that are based on TCP congestion control and PID control respectively. Through experiments, we show that ANCHORS can save more than 50\% CPU utilization while still meeting deadlines 97\% of the time. Finally, we present a stride scheduling based algorithm to maintain application real-time performance when the combined resource demand is greater than system’s available resource. Through experiment, we show that ANCHORS can avoid more than 50\% of deadline misses under oversubscribed resource demand system where no deadlines are met without ANCHORS.