Quality-of-Service-Aware Scheduling in Heterogeneous Data centers with Paragon

Abstract:

Large-scale datacenters host tens of thousands of diverse applications each day. However, performance is degraded by interference between colocated workloads and the difficulty of matching applications to one of the many hardware platforms available, violating the quality of service (QoS) guarantees that many cloud workloads require. Thus, the authors present Paragon, an online and scalable datacenter scheduler that is aware of heterogeneity and interference. Paragon is derived from robust analytical methods. Instead of profiling each application in detail, it leverages information the system already has about applications it has previously seen. It uses collaborative filtering techniques to quickly and accurately classify an unknown, incoming workload with respect to heterogeneity and interference by identifying similarities to previously scheduled applications. The classification allows Paragon to greedily schedule applications in a manner that minimizes interference and maximizes server utilization. Paragon scales to tens of thousands of servers with marginal scheduling overheads. The authors evaluated Paragon with many workload scenarios, on both small and large-scale systems, including 1,000 servers on Amazon Elastic Compute Cloud (Amazon EC2). For a 2,500-workload scenario, Paragon preserves performance constraints for 91 percent of applications, while significantly improving utilization. In comparison, a baseline least-loaded scheduler only provides similar guarantees for 3 percent of workloads. The differences are more striking during high load when resource efficiency is more critical.