Cross-Layer Resource Management for QoS in Serverless Platforms

Abstract

Serverless computing has emerged as a prominent cloud computing paradigm by moving the infrastructure management from developers to platforms while providing automatic resource scaling and fine-grained billing. Despite these advantages, guaranteeing predictable quality of service (QoS) for diverse serverless workloads on shared multi-tenant infrastructures remains a fundamental challenge. Performance degradation arises from bottlenecks at multiple stages of the execution pipeline, including function image retrieval, sandbox initialization, runtime resource contention, and data movement during the execution of a workflow. This thesis systematically investigates these bottlenecks and proposes a set of complementary mechanisms to improve the performance, scalability, and resource efficiency of serverless platforms.

To reduce function startup latency, this thesis proposes FaaSImage, an image management framework that combines layer subsetting, layer fusion, and on-demand file retrieval to minimize image download overheads. To improve the scalability of snapshot-based sandbox initialization, it presents Snapstore, a memory region-aware snapshot deduplication framework that exploits the characteristics of application memory regions to reduce storage overhead while improving deduplication efficiency. To address runtime resource contention, this thesis introduces the notion of the comprehensive latency (CL) as a holistic QoS metric and proposes FaaSCtrl, a reinforcement learning-based resource scheduler that dynamically allocates CPU resources and adjusts process priorities to improve the performance of latency-sensitive applications while preserving fairness for best-effort workloads.

To improve the execution efficiency of serverless workflows, this thesis further proposes Styx, a workflow engine that decouples computation from data movement through predictive input prefetching and asynchronous output handling. By overlapping computation with I/O and reducing the memory provisioning duration of workflow stages, Styx decreases request waiting time, improves resource utilization, and increases workflow throughput. Collectively, the proposed systems optimize successive stages of the serverless execution pipeline, constituting a cross-layer resource management framework that enhances the performance, scalability, and QoS of modern serverless platforms.