Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Deployment and Management of Microservices at the Network Edge

Madhura Adeppady

Abstract

Edge computing represents a paradigm shift in computing, where processing and storage resources are distributed towards the network edge, in close proximity to the data source. Unlike traditional cloud, edge computing can meet low latency and high bandwidth requirements of time- and mission-critical applications of 5G and next-generation mobile networks. Applications deployed at the edge increasingly adopt microservice architecture enabling monolithic applications to be decomposed into multiple loosely coupled microservices (MSs) implemented through containers. Such MSs are then independently developed, distributed, and maintained on the edge nodes. Although edge computing seems to be promising, effectively deploying and managing containerized MSs on resource-constrained edge nodes is challenging. This thesis focuses on addressing these challenges by proposing various algorithmic approaches.

To increase resource utilization, multiple MSs are often placed on the same server, and isolation among them is provided by running the associated containers on dedicated cores. Interference arises among such co-located MSs on the same server as they share and compete for memory resources. Such interference can result in severe throughput degradation, violating the Quality of Service (QoS) offered to the users. We propose iPlace, an interference-aware MS placement algorithm, that clusters together MSs competing for resources as diverse as possible, and hence, interfering as little as possible. Additionally, clustering enables batch deployment of MSs associated with a cluster, hence reducing the deployment time compared to the sequential deployment. Compared to the state-of-the-art schemes, iPlace uses 21-92% fewer servers proving to be highly scalable. Further, by deploying MSs in parallel using Kubernetes, iPlace reduces the deployment time by 69% compared to the state-of-the-art solutions.

Due to high resource elasticity promises and the ability to handle short-lived MSs effectively, there has been considerable interest in adopting the serverless computing model originally proposed for the cloud at edge nodes. Unfortunately, the high startup latency of the containers is a critical issue in serverless edge computing as it significantly impacts the responsiveness of the MSs. Advancements in virtualization techniques proposed several container states, such as warm, pre-warm, etc., which can be cached at the edge nodes to reduce the startup latency for future requests. Nonetheless, the resource overheads of these container states limit their practicality. In this work, first, we characterize the resource overheads of these container states and their effectiveness in reducing startup latency. To dynamically provision edge resources according to user demands, we propose the Always in Warm (AiW) algorithm for the orchestrator. By leveraging a multi-queueing system for waiting MS requests, AiW can balance resource overheads of warm containers and performance tradeoffs by reusing the existing containers, and invoking cold-starts only when necessary. A high container reusable probability of AiW can reduce the energy consumption of the active servers. To further explore energy minimization opportunities, we introduce COME, a two-timescale framework comprising orchestrator running our proposed AiW algorithm for container provisioning and the Dynamic Server Provisioner (DSP) for dynamically activating/deactivating servers in response to AiW's decisions on request scheduling. Extensive performance evaluation demonstrates AiW's close match to the optimum and COME's significant reduction in power consumption by 22–64% compared to its alternatives.