A two-tier multi-objective service placement in container-based fog-cloud computing platforms

Javad Dogani, Ali Yazdanpanah, Arash Zare, Farshad Khunjush

DOI: https://link.springer.com/article/10.1007/s10586-023-04183-8

Abstract

The utilization of cloud computing in Internet of Things (IoT) applications has become widespread. However, it presents challenges for latency-sensitive scenarios due to data transmission to the centralized cloud structure, which leads to increased network traffic and service delays. To address this, fog computing has emerged as an intermediary layer between the cloud and IoT, ensuring low-latency interactions. A pivotal challenge within the fog computing paradigm is the service placement problem, involving assigning services to appropriate nodes, which is recognized as NP-hard. Recognizing the intricate nature of service placement, this study introduces a multi-objective optimization approach tailored for dynamic service placement within container-based fog computing environments. Considering multiple objectives is imperative due to the complex interplay of performance metrics in fog computing scenarios. A two-tier resource management framework based on Kubernetes is proposed to manage these diverse yet interrelated objectives effectively. The framework harnesses the power of the multi-objective, non-dominated sorting genetic algorithm II (NSGA-II) to reconcile conflicting objectives and facilitate optimal service placement decisions. Incorporating multi-objective optimization enables a comprehensive evaluation of service placement solutions, ensuring a balanced trade-off between latency, cost-efficiency, and energy consumption. Empirical evaluations demonstrate that the proposed method improves cost, average latency time, and energy consumption by 8% to 36% compared to state-of-the-art methods.

Keywords: Fog computing, service Placement, container-based virtualization, Multi-objective optimization, Kubernetes