1. Minimizing Service Latency through Image-Based Microservice Caching and Randomized Request Routing in Mobile Edge Computing

Accession number: 20242416236698
Authors: Sun, Xiao (1); Wang, Desheng (2); Zhang, Weizhe (1); Lou, Guanqing (1); Wang, Jiayin (1); Yadav, Rahul (3)
Author affiliation: (1) School of Cyberspace Science, Harbin Institute of Technology, Harbin, China; (2) School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong, China; (3) College of Computer Science and Technology, Harbin Engineering University, Harbin, China
Source title: IEEE Internet of Things Journal
Abbreviated source title: IEEE Internet Things J.
Issue date: 2024
Publication year: 2024
Pages: 1-1
Language: English
E-ISSN: 23274662
Document type: Article in Press
Publisher: Institute of Electrical and Electronics Engineers Inc.
Abstract: In the context of Mobile Edge Computing (MEC), the traditional method of requesting microservices from a central cloud can result in increased delay for users due to the physical distance between the user and the cloud server. To address this issue, MEC advocates for placing servers closer to the users at the edge of the network. However, this approach is constrained by the storage capacity and computing resources of edge servers. Therefore, it is crucial to devise a strategy for processing user requests that minimizes the average request delay. To address this problem, This paper formulates the microservice caching problem as an image-based microservice placement and task request routing problem. We model the problem as an integer linear programming problem with multi-condition constraints. Considering the limited resources of edge servers, we propose a microservice placement algorithm called approximate algorithm based on randomized task request routing. The proposed algorithm is designed to provide near-optimal solutions in polynomial time, leveraging Chernoff's theorem. Our approach is evaluated through comparisons with two existing algorithms: the image-pull-based microservice cache request algorithm and the greedy-based microservice cache and request routing algorithm. The results demonstrate that our algorithm exhibits superior performance compared to existing methods. IEEE
Main heading: Approximation algorithms
Controlled terms: Computation theory - Edge detection - Integer programming - Internet of things - Job analysis - Mobile edge computing - Network routing - Polynomial approximation
Uncontrolled terms: Edge server - Image edge detection - Image-based - Microservice architecture - Microservice cache - Routings - Service latency - Task analysis - Task request routing
Classification code: 721.1 Computer Theory, Includes Formal Logic, Automata Theory, Switching Theory, Programming Theory - 722.3 Data Communication, Equipment and Techniques - 722.4 Digital Computers and Systems - 723 Computer Software, Data Handling and Applications - 921 Mathematics - 921.5 Optimization Techniques - 921.6 Numerical Methods
DOI: 10.1109/JIOT.2024.3410546
Database: Compendex
Data Provider: Engineering Village