



UMEÅ UNIVERSITET

Location-aware Resource Allocation in Mobile Edge Clouds

Chanh Nguyen

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för
avläggande av filosofie doktorexamen framläggs till offentligt
försvar i Aula Biologica, Umeå Universitet, den 17 Februari, kl.
13:00.

Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor Maarten van Steen
Digital Society Institute, University of Twente, the Netherlands

Department of Computing Science

Organization
Umeå University
Dept. of Computing Science

Document type
Doctoral thesis

Date of publication
27th January 2021

Author
Chanh Nguyen

Title
Location-aware Resource Allocation for Mobile Edge Clouds

Abstract

Over the last decade, cloud computing has realized the long-held dream of computing as a utility, in which computational and storage services are made available via the Internet to anyone at any time and from anywhere. This has transformed Information Technology (IT) and given rise to new ways of designing and purchasing hardware and software. However, the rapid development of the Internet of Things (IoT) and mobile technology has brought a new wave of disruptive applications and services whose performance requirements are stretching the limits of current cloud computing systems and platforms. In particular, novel large scale mission-critical IoT systems and latency-intolerant applications strictly require very low latency and strong guarantees of privacy, and can generate massive amounts of data that are only of local interest. These requirements are not readily satisfied using modern application deployment strategies that rely on resources from distant large cloud datacenters because they easily cause network congestion and high latency in service delivery.

This has provoked a paradigm shift leading to the emergence of new distributed computing infrastructures known as Mobile Edge Clouds (MECs) in which resource capabilities are widely distributed at the edge of the network, in close proximity to end-users. Experimental studies have validated and quantified many benefits of MECs, which include considerable improvements in response times and enormous reductions in ingress bandwidth demand. However, MECs must cope with several challenges not commonly encountered in traditional cloud systems, including user mobility, hardware heterogeneity, and considerable flexibility in terms of where computing capacity can be used. This makes it especially difficult to analyze, predict, and control resource usage and allocation so as to minimize cost and maximize performance while delivering the expected end-user Quality-of-Service (QoS). Realizing the potential of MECs will thus require the design and development of efficient resource allocation systems that take these factors into consideration.

Since the introduction of the MEC concept, the performance benefits achieved by running MEC-native applications (i.e., applications engineered specifically for MECs) on MECs have been clearly demonstrated. However, the benefits of MECs for non-MEC-native applications (i.e., application not specifically engineered for MECs) are still questioned. This is a fundamental issue that must be explored because it will affect the incentives for service providers and application developers to invest in MECs. To spur the development of MECs, the first part of this thesis presents an extensive investigation of the benefits that MECs can offer to non-MEC-native applications. One class of non-MEC-native applications that could potentially benefit significantly from deployment on an MEC is cloud-native applications, particularly micro-service-based applications with high deployment flexibility. We therefore quantitatively compared the performance of cloud-native applications deployed using resources from cloud datacenters and edge locations. We then developed a network communication profiling tool to identify aspects of these applications that reduce the benefits derived from deployment on MECs, and proposed design improvements that would allow such applications to better exploit MECs' capabilities. The second part of this thesis addresses problems related to resource allocation in highly distributed MECs. First, to overcome challenges arising from the dynamic nature of resource demand in MECs, we used statistical time series models and machine learning techniques to develop two location-aware workload prediction models for EDCs that account for both user mobility and the correlation of workload changes among EDCs in close physical proximity. These models were then utilized to develop an elasticity controller for MECs. In essence, the controller helps MECs to perform resource allocation, i.e. to answer the intertwined questions of what and how many resources should be allocated and when and where they should be deployed. The third part of the thesis focuses on problems relating to the real-time placement of stateful applications on MECs. Specifically, it examines the questions of where to place applications so as to minimize total operating costs while delivering the required end-user QoS and whether the requested applications should be migrated to follow the user's movements. Such questions are easy to pose but intrinsically hard to answer due to the scale and complexity of MEC infrastructures and the stochastic nature of user mobility.

To this end, we first thoroughly modeled the workloads, applications, and infrastructures to be expected in MECs. We then formulated the various costs associated with operating applications, namely the resource cost, migration cost, and service quality degradation cost. Based on our model, we proposed two online application placement algorithms that take these factors into account to minimize the total cost of operating the application. The methods and algorithms proposed in this thesis were evaluated by implementing prototypes on simulated testbeds and conducting experiments using workloads based on real mobility traces. These evaluations showed that the proposed approaches outperformed alternative state-of-the-art approaches and could thus help improve the efficiency of resource allocation in MECs.

Keywords

Mobile Edge Clouds, Edge Data Centers, Location-aware, Resource Allocation, Elasticity, Auto-Scaling, Placement, Stateful Application, Micro-service, Cloud computing

Language
English

ISBN
Print: 978-91-7855-467-6
PDF: 978-91-7855-466-9

ISSN
0348-0542

Number of pages
50 + 5 papers