UMEÅ UNIVERSITET

Edge Orchestration for Latency-Sensitive Applications

Ali Rahmanian

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Hörsal UB.A.240 - Lindellhallen 4, den 29 April, kl. 13:00. Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor Ada Gavrilovska,
Georgia Institute of Technology, Atlanta, GA, USA

Department of Computing Science

Organization
Umeå University
Department of Computing Science

Document type
Doctoral thesis

Date of publication
18 03 2024

Author
Ali Rahmanian

Title
Edge Orchestration for Latency-Sensitive Applications

Abstract
The emerging edge computing infrastructure provides distributed and heterogeneous resources closer to where data is generated and where end-users are located, thereby significantly reducing latency. With the recent advances in telecommunication systems, software architecture, and machine learning, there is a noticeable increase in applications that require processing times within tight latency constraints, i.e. latency-sensitive applications. For instance, numerous video analytics applications, such as traffic control systems, necessitate real-time processing capabilities. Orchestrating such applications at the edge offers numerous advantages, including lower latency, optimized bandwidth utilization, and enhanced scalability. However, despite its potential, effectively managing such latency-sensitive applications at the edge poses several challenges such as constrained compute resources, which holds back the full promise of edge computing.

This thesis proposes approaches to efficiently deploy latency-sensitive applications on the edge infrastructure. It partly addresses general applications with microservice architectures and party addresses the increasingly more important video analytics applications for the edge. To do so, this thesis proposes various application- and system-level solutions aiming to efficiently utilize constrained compute capacity on the edge while meeting prescribed latency constraints. These solutions primarily focus on effective resource management approaches and optimizing incoming workload inputs, considering the constrained compute capacity of edge resources. Additionally, the thesis explores the synergy effects of employing both application- and system-level resource optimization approaches together.

The results demonstrate the effectiveness of the proposed solutions in enhancing the utilization of edge resources for latency-sensitive applications while adhering to application constraints. The proposed resource management solutions, alongside application-level optimization techniques, significantly improve resource efficiency while satisfying application requirements. Our results show that our solutions for microservice architectures significantly improve end-to-end latency by up to 800% while minimizing edge resource usage. Additionally, the results indicate that our application- and system-level optimizations for orchestrating edge resources for video analytics applications can increase the overall throughput by up to 60%. .

Keywords
Edge Computing, Resource Management, Latency-Sensitive Applications, Edge Video Analytics.

Language
English

ISBN
Print: 978-91-8070-350-5
PDF: 978-91-8070-351-2

ISSN
0348-0542

Number of pages
46 + 5 papers