Abstract

Network Functions Virtualization, Edge Computing, and Microservices architectures are pivotal concepts to attain unprecedented performance and flexibility in next-generation mobile networks. Nevertheless, the rapid proliferation of AI-driven network services for latency- and mission- critical use cases introduces significant challenges, including maintaining continuous proximity to mobile end users, ensuring efficient data coordination and sharing, as well as minimizing the overall network energy footprint. This thesis sheds light on these challenges by focusing on three fundamental aspects of edge service orchestration and management.

First, we tackle stateful migration as a core solution for supporting latency-sensitive microservices at the edge while ensuring a satisfying experience for mobile end users. To this end, we introduce COAT, a novel, agnostic, and connection-aware migration process, and derive PAM, an analytical model designed to accurately predict the fundamental migration KPIs. Building on these, we propose MOSE, an innovative framework that efficiently implements and orchestrates the migration process by effectively fulfilling both network and application KPI targets. Next, we shift our focus to alternative approaches that enhance the migration process while ensuring uninterrupted service operation. We do so by tackling distributed data coordination and sharing while addressing the inherent challenges of distributed microservices architectures. To this end, we propose and compare various architectures and implementations of a holistic edge platform that regulates the collection and usage of network- and context- related information. To facilitate realistic emulation of diverse edge environments, we develop PACE, a highly configurable, scalable, microservice-based emulation framework that reproduces a broad range of interaction patterns, load dynamics, and traffic scenarios. Finally, we integrate these solutions within the practical context of Open RAN, introducing CORMO-RAN, a data-driven edge orchestrator that jointly optimizes compute node activation and microservice migration strategies so as to minimize the overall RAN energy footprint while ensuring uninterrupted service operation.

Our extensive experimental evaluation, leveraging real-world microservices across a spectrum of practical use cases — ranging from UAV autopilots, multi-object tracking, and DRL-based RAN schedulers — demonstrates the effectiveness of our

contributions. Results show that PAM predicts migration KPIs values with an error up to 99.7% smaller than state-of-the-art models, while MOSE, compared to existing solutions, greatly improves migration performance, achieving up to a 77% decrease of the service disruption duration upon migration. Further, we demonstrate that PACE effectively reveals fundamental trade-offs in varying edge platform architectures and implementations, and proves to be a valuable tool to support the design of robust and efficient microservices coordination and data sharing mechanisms across heterogeneous edge scenarios. Lastly, we prove that CORMO-RAN effectively orchestrates edge nodes activation and service placement by optimally balancing service availability, scalability, and energy-efficiency, yielding up to a 64% reduction of the edge cluster energy consumption compared to existing approaches.

By addressing these critical research challenges and demonstrating the effectiveness of our solutions in practical, real-world scenarios, this thesis advances the state of the art in mobility-aware orchestration and management of edge services, paving the way for more efficient, resilient, and sustainable next-generation mobile networks.