QoE- and System-aware Management for Critical Microservice at the Mobile Edge

Yenchia Yu

Edge computing targets to bring computation and data storage closer to end users, thereby reducing communication latency and improving application responsiveness. However, with the growing prevalence of complex, latency-sensitive, and mission-critical applications deployed on high mobility users, such as Unmanned Aerial Vehicles (UAVs), simple edge offloading strategies struggle to satisfy the stringent user requirements without significant resource overprovisioning, which is often impractical in resource-limited edge system. This thesis investigates QoE-aware and system-aware management of critical microservices at the mobile edge, addressing both application-level quality guarantees and infrastructure-level resource efficiency.

We first propose an application-aware management approach, focusing on offloading a critical deep neural network (DNN) from a UAV to the edge in the form of a microservice. A split-computing paradigm is adopted, in which a lightweight DNN head executes on the UAV, while the computationally intensive tail runs at the edge. To overcome the prohibitive bandwidth cost associated with transmitting intermediate activation tensors, we introduce the CoTeD framework, which dynamically compresses and reconstructs such tensors without requiring model retraining. CoTeD explicitly trades off wireless bandwidth consumption against inference accuracy and latency. Experimental results demonstrate that CoTeD reduces radio traffic by up to 90% while maintaining inference success rates of at least 90% under time-varying network conditions.

Then we investigate an application-agnostic approach to manage the offloaded edge microservices, focusing on stateful microservice migration and dynamic resource allocation. To enable seamless service relocation, we first introduce COAT, a network architecture that enhances traditional stateful migration by enabling transparent transport-layer connection migration. Building on this capability, we propose the MOSE framework, which autonomously configures and executes stateful microservice migration, reducing migration overhead by up to 77% while satisfying both network- and application-level requirements. To support accurate and efficient orchestration, we further develop the PAM model, which precisely characterizes migration duration and service downtime by capturing real-world processing over-

heads neglected by prior models, achieving up to a 99% reduction in estimation error. Finally, we address the joint deployment, resource allocation, and migration of composite applications at the edge. We formulate the MAP (Multi-microservice Application Placement) problem and propose the STEP heuristic, which efficiently manages both stateless and stateful microservices, exploits service shareability, and adapts service quality to available resources. Large-scale Kubernetes-based experiments demonstrate that STEP nearly doubles the edge application quality while reducing deployment costs by up to 50% and radio resource usage by up to 15%.