Towards Explainable, Coordinated and Proactive Autoscaling for Microservices and Function Chains

Martin Straesser

3rd International Workshop on Serverless Computing Experience
Introduction

- Automated resource management for cloud applications has always been a crucial aspect for the growth of cloud computing.
- The ever-growing number of customers and newest trends in the domain (e.g., serverless computing) demand novel autoscaling approaches.
- New challenges and opportunities arise regularly (e.g., new architectural styles, programming frameworks, enhanced observability).
Design Goals for Our Autoscaler

- End-to-end metrics of user transactions are in the focus
  - SLOs are given as end-to-end latencies for user transactions (trace-level SLOs)
  - No fine-granular SLOs or performance requirements for every service/endpoint

- Tail Latency-Awareness
  - High latency quantiles are more important than mean latency

- Coordinated
  - Multiple services are scaled in a coordinated way in one scaling cycle
  - Prevents bottleneck shifting, reduces resource wastage
Thank you