Guaranteeing Quality of Service (QoS) in an (edge-)cloud environment is one of the biggest open problems in the field of cloud computing. Currently, deployment of cloud applications is managed by cloud orchestration systems, such as Kubernetes. These systems make deployment of applications in cloud easier than ever, offering their users the benefits of availability, scalability and resilience. However, at the moment they are not capable of optimizing the deployment of cloud applications with respect to performance QoS metrics, such as response time and throughput.

The thesis proposes an approach that provides probabilistic guarantees on the performance QoS metrics in an (edge-)cloud environment. The approach is based on assessing the performance of cloud applications and subsequently controlling their deployment in a way that the applications are deployed only in the environments in which their performance does not violate their QoS requirements. The thesis also presents a proof-of-concept implementation of that approach. The implementation verifies the effectiveness of the approach and will serve for further research.