Workload Characterization, Controller Design and Performance Evaluation for Cloud Capacity Autoscaling

Ahmed Ali-Eldin Hassan

Department of Computing Science
Umeå University
Sweden

Akademisk avhandling
för avläggande av teknologie doktorsexamen, som med vederbörligt tillstånd av Rektor vid Umeå universitet framlägges till offentlig granskning Fredag den 2:e Oktober 2015, 14:00, i N360, Naturvetarhuset, Umeå universitet. Avhandlingen kommer att försvaras på engelska. Fakultetsopponent är Dr. Prashant Shenoy, Department of Computer Science, University of Massachusetts, Amherst, USA.

Dissertation
for the degree of Doctor of Philosophy, as authorised by the Vice-Chancellor at Umeå University, will be publicly defended on Friday 2nd of October 2015, at 14:00, in N360, Naturvetarhuset building, Umeå universitet. Faculty opponent is Dr. Prashant Shenoy, Department of Computer Science, University of Massachusetts, Amherst, USA.

Abstract
This thesis studies cloud capacity auto-scaling, or how to provision and release resources to a service running in the cloud based on its actual demand using an automatic controller. As the performance of server systems depends on the system design, the system implementation, and the workloads the system is subjected to, we focus on these aspects with respect to designing auto-scaling algorithms. Towards this goal, we design and implement two auto-scaling algorithms for cloud infrastructures. The algorithms predict the future load for an application running in the cloud. We discuss the different approaches to designing an auto-scaler combining reactive and proactive control methods, and to be able to handle long running requests, e.g., tasks running for longer than the actuation interval, in a cloud. We compare the performance of our algorithms with state-of-the-art auto-scalers and evaluate the controllers' performance with a set of workloads. As any controller is designed with an assumption on the operating conditions and system dynamics, the performance of an auto-scaler varies with different workloads. In order to better understand the workload dynamics and evolution, we analyze a 6-years long workload trace of the sixth most popular Internet website. In addition, we analyze a workload from one of the largest Video-on-Demand streaming services in Sweden. We discuss the popularity of objects served by the two services, the spikes in the two workloads, and the invariants in the workloads. We also introduce, a measure for the disorder in a workload, i.e., the amount of burstiness. The measure is based on Sample Entropy, an empirical statistic used in biomedical signal processing to characterize biomedical signals. The introduced measure can be used to characterize the workloads based on their burstiness profiles. We compare our introduced measure with the literature on quantifying burstiness in a server workload, and show the advantages of our introduced measure. To better understand the tradeoffs between using different auto-scalers with different workloads, we design a framework to compare auto-scalers and give probabilistic guarantees on the performance in worst-case scenarios. Using different evaluation criteria and more than 700 workload traces, we compare six state-of-the-art auto-scalers that we believe represent the development of the field in the past 8 years. Knowing that the auto-scalers' performance depends on the workloads, we design a workload analysis and classification tool that assigns a workload to its most suitable elasticity controller out of a set of implemented controllers. The tool has two main components; an analyzer, and a classifier. The analyzer analyzes a workload and feeds the analysis results to the classifier. The classifier assigns a workload to the most suitable elasticity controller based on the workload characteristics and a set of predefined business level objectives. The tool is evaluated with a set of collected real workloads, and a set of generated synthetic workloads. Our evaluation results shows that the tool can help a cloud provider to improve the QoS provided to the customers.

Keywords
cloud computing, autoscaling, workloads, performance modeling, controller design.

Document name	**Language**	**Date of issue**	**Pages**
Doctoral thesis	English	September 11th, 2015	23 + 7 papers
ISBN	**ISSN**	**UMINF**	
978-91-7601-330-4	0348-0542	15.09	
Department	**School**	**Address**	
Computing Science	Umeå University	SE-901 87 Umeå	