Student Research Abstract: Towards workload-aware fine-grained control over cloud resources

Amjad Ullah
Division of Computer Science &
Mathematics, University of Stirling, UK
aul@cs.stir.ac.uk

1. PROBLEM AND MOTIVATION

The systems deployed over cloud are subject to unpredictable workload conditions that vary from time to time, e.g. an e-commerce website may face higher workloads than normal during festivals or promotional schemes. In order to maintain the performance of such systems, an efficient elastic resource provisioning strategy is required. However, providing such a strategy that determines the right amount of cloud resources that fulfills the Quality of Service (QoS) demand is a challenging task. Over the period, many proposals have been introduced using techniques like threshold based rules, reinforcement learning and control theory, etc. The existing proposals, however, suffer from issues like lack of expertise to appropriately set the quantitative specification of thresholds, online training time overhead of the algorithm, too specific to work well in particular situation like when there is sudden burst in workload or work well in nominal conditions for stable workload, etc. Moreover, the existing approaches do not address uncertainty. Our proposed framework is a step forward to address the mentioned issues for systems that hold time varying workload conditions.

2. BACKGROUND AND RELATED WORK

Elasticity enriches the applications to dynamically adjust the underlying cloud resources according to the application needs. Threshold based rules is one such technique, which is commonly using by public cloud providers. This technique follows a condition-action pattern, whereas the condition can be derived from system metrics and action specifies the scaling decision. This technique is easy to use; however, they have the following limitations. Firstly, it lacks detailed understanding of the system required to appropriately state the threshold quantities. Secondly, it does not handle uncertainty raised due to unpredictable events such as sudden raise in workload or errors in measurements [1]. Reinforced Learning (RL) is another technique to implement elasticity, which rely on the learning of an auto-scaler that acts as an agent with its environment (application) using trial and error method to figure out the most suitable elastic decision. Such approaches are often criticized for bad performance due to long online training and their inability to cope with sudden burst [2].

SAC 2016, April 04-08, 2016, Pisa, Italy
ACM 978-1-4503-3739-7/16/04.
http://dx.doi.org/10.1145/2851613.2852009

Alternative approach to elasticity includes the use of control theory, which provides a systematic methodology to design elastic feedback controllers that quickly react to disturbances in workload to achieve a target performance. Such controllers usually satisfy a constraint or maintain the output of the system to a desired value [3]. More specifically, an elastic feedback controller maintains the value of a controlled variable (e.g. response time) closed to a reference point by changing the value of a manipulated variable (e.g. virtual machines) [2]. Over the past, the use of feedback controller has been exploited specifically in cloud resource provisioning [2]. The various available feedback controllers can be generally categorized as either fixed or adaptive [2], [4]. Both approaches have their own shortcomings. For example, fixed controllers are criticized of their inability to maintain performance for systems with high workload variability [4], whereas adaptive controllers are blamed for extra computational cost due to online estimation [5] and their inability to handle sudden burst in workload [4]. Moreover, the majority of the existing approaches are based on the use of single model that represents the system behavior in the entire operating period. Such approaches cannot perform well all the time for systems with time varying workloads. This research advocates the use of multi-controller to implement cloud elasticity.

3. APPROACH AND UNIQUENESS

3.1 Proposed methodology

Biological systems have the ability to identify a natural occurred situation and execute the appropriate action in response that results in efficient adaptation of the system to the environment [6]. Our approach is based on a similar idea, i.e. designing multiple controllers and identifying the right one using an intelligent mechanism. Our inspiration comes from the use of multiple controllers for complex systems, e.g. autonomous vehicle systems [7], where each controller is designed specifically for a different component. In cloud, a deployed web system may face various workload behaviors at different time. Thus an efficient elastic system is needed to respond according to the changing behavior of the underlying system. Therefore, the proposed methodology uses multi controller approach, where each controller must be specifically designed for a different operating region and the selection of suitable controller for final decision will be realized at runtime.

Figure 1 represents the architecture of our proposed approach. The key idea is to divide the complexity of the overall system by constructing multiple fixed gain controllers. Designing such a system involves two key challenges: (1) how to partition the system among multiple controllers? (2) How to switch/formulate the final decision? Due to the lack of a standard approach for partitioning the system among sub controllers [7], this research

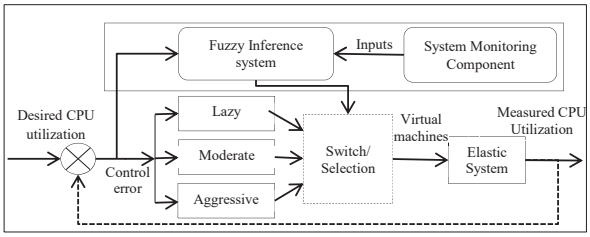

Figure 1: Proposed architecture

realizes the use of expert-oriented distribution of workload intensity into various categories such as low, moderate and high. For each category, a system model is constructed based on which individual controller is designed. The Fuzzy Inference System (FIS) is used as a decision unit that determines the switching mechanism. Two approaches are considered: (1) hard switching - the switching/selection component only allows the suitable controller as per the decision of the FIS; (2) soft switching - the FIS concludes the strength of each controller that determines the contribution of each controller to the final decision. The FIS considers the system behavior including workload intensity, current performance (response time) and control error for decision making. These measurements form inputs of the fuzzy rules, whereas the output is one of the controllers. The following is an example of such rule from the hard switching approach:

IF arrivalRate IS high AND responseTime IS instantaneous AND error IS positive THEN controller IS lazy;

3.2 Uniqueness

Feedback controllers have been exploited for cloud resource provisioning problem. However, to the best of our knowledge, the use of multiple controllers with intelligent fuzzy selection mechanism for cloud resource provisioning problem is novel. The proposed approach has the advantages of both the adaptive and fixed based feedback approaches. Our approach is itself adaptive by adapting to the suitable controller or the contribution of each controller for the current system behavior whereas the design of individual controllers is fixed. Thus, no online estimation is required. Moreover, the use of fuzzy logic helps to incorporate uncertainty and the qualitative selection of the controllers.

4. RESULTS AND CONTRIBUTIONS

As a first phase of the work, we have implemented a prototype using CloudSim [8] in order to demonstrate the suitability of the framework. Currently, the experiments are only performed using hard switching approach. The domain knowledge for the FIS in this phase is extracted from the research carried out in [1]. The details and preliminary results obtained are presented in [9], where the proposed methodology is compared with single/conventional controller approach using two criteria, i.e. the percentage number of SLO violation and the cost. Figure 2 (b) represents an overview of similar results by extending the comparison of the proposed methodology with one commonly used threshold based rules technique called Rightscale.

A real network trace obtained from [10] has been used for this experiment. However, this trace is scaled down to 6000 requests per minute as can be seen from Figure 1 (a). SLO violation and cost in Figure 2 (b) represents the evaluation criteria. SLO violation refers to the phenomenon, where a job request is unable to be completed within a desired time period. This is treated as the

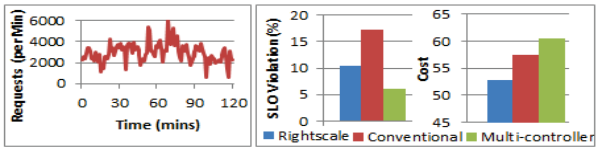

Figure 2: (a) - WITS: Auckland-2 trace (b) - Aggregated results

performance of the system. The results in Figure 2 (b) demonstrate that our multi-controller approach is a bit expensive in terms of cost in comparison to other approaches. However, more importantly, it is much better in performance than the other two approaches. This highlights that the multi-controller approach has higher potential to improve system performance by minimizing the SLO violation. The current focus of the work is twofold. Firstly, the analysis of SASO (Stability, Accuracy, Short settling, Overshoot) properties is in progress for formal evaluation. Moreover, further evaluation of the suitability of the proposed approach using various cloud workload patterns such as big spike, slowly varying and dual phase, etc. will be explored. Secondly, the work on soft switching approach will be carried out.

5. REFERENCES

[1] P. Jamshidi, A. Ahmad, and C. Pahl, "Autonomic resource provisioning for cloud-based software," in *Proceedings of the 9th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2014, pp. 95–104.

[2] T. Lorido-Botran, J. Miguel-Alonso, and J. A. Lozano, "A review of auto-scaling techniques for elastic applications in cloud environments," *J. Grid Comput.*, vol. 12, no. 4, pp. 559–592, 2014.

[3] H. Ghanbari, B. Simmons, M. Litoiu, and G. Iszlai, "Exploring alternative approaches to implement an elasticity policy," in *Cloud Computing (CLOUD), 2011 IEEE International Conference on*, 2011, pp. 716–723.

[4] T. Patikirikorala and A. Colman, "Feedback controllers in the cloud," in *Proceedings of APSEC*, 2010.

[5] T. Patikirikorala, A. Colman, J. Han, and L. Wang, "A multi-model framework to implement self-managing control systems for QoS management," in *Proceedings of the 6th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2011, pp. 218–227.

[6] K. S. Narendra, O. A. Driollet, M. Feiler, and K. George, "Adaptive control using multiple models, switching and tuning," *Int. J. Adapt. Control Signal Process.*, vol. 17, no. 2, pp. 87–102, 2003.

[7] A. Hussain, R. Abdullah, E. Yang, and K. Gurney, "An intelligent multiple-controller framework for the integrated control of autonomous vehicles," in *Advances in Brain Inspired Cognitive Systems*, Springer, 2012, pp. 92–101.

[8] R. N. Calheiros, R. Ranjan, A. Beloglazov, C. A. F. De Rose, and R. Buyya, "CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms," *Softw. Pract. Exp.*, vol. 41, no. 1, pp. 23–50, 2011.

[9] A. Ullah, J. Li, and A. Hussain, "Towards Workload-Aware Cloud Resource Provisioning Using a Novel Multi-Controller Fuzzy Switching Approach," *Int. J. High Perform. Comput. Netw.*, 2015 (accepted).

[10] Wand, "WITS: Auckland II," 2015. [Online]. Available: http://wand.net.nz/wits/auck/2/auckland_ii.php.