NEW CLOUD BUSINESS MODEL TOWARDS PERFORMANCE

Monika Simjanoska[1], Goran Velkoski[1], Sasko Ristov[2], Marjan Gusev[2]
Ss. Cyril and Methodius University,
Faculty of Information Sciences and Computer Engineering,
Skopje, Macedonia
{m.simjanoska, velkoski.goran}@gmail.com[1], {sashko.ristov, marjan.gushev}@finki.ukim.mk[2]

Abstract *– Cloud service providers (CSPs) offer a pay-per-use pricing model which charges the customers according to acquired resources. The quality assurance of such a model is negotiated between the customers and the CSPs via Service Level Agreements (SLAs). However, SLAs almost never guarantee sustainable performance. In this paper we propose a new CSP pricing model that aims towards performance charging, instead of rental. We believe that this model provides the customers maximum performance gain with minimum monetary costs. The methodology we developed is based on machine learning analysis that promise to provide accurate decisions when offering both performance and cost-effective configurations in the cloud.*

1. INTRODUCTION

Cloud computing is a model for enabling ubiquitous, convenient, on demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction [1]. The illusion of infinite computing resources available on demand and the ability to pay per use of computing resources on a short-term basis as needed [2], offers several advantages, as reduced costs, ability to scale up as the customer requires, etc. [3].

Computing services need to be highly reliable, scalable, and autonomic to support ubiquitous access, dynamic discovery and composability. In particular, consumers can determine the required service level through Quality of Service (QoS) parameters and SLAs [4]. That is, the service provider is required to execute service requests from a customer within negotiated quality of service requirements for a given price [5]. Even though SLAs provide some level of guarantee of how much of the time the server, platform, or application will be available, the performance is almost never discussed [6].

We believe that CSPs must confront the challenge of configuring their offers in maximum performance–minimum monetary costs manner, in order to fulfill customer's expectations of proportionality between the performance and the amount paid for resources. Therefore, in this paper we propose a new machine learning based model which allows the CSPs to classify and map the input load into specific cloud configuration. Hereupon, the customer will be proposed an offer based on his performance expectations. While modeling, we take into account realistic occasions with various web services, variable load, different cloud environments, etc.

The rest of the paper is organized as follows. In Section 2 we present some of the work related to our research. The original methodology we developed is given in Section 3. Finally in Section 4, we exhibit our conclusion and plans for a future application of the introduced methodology.

2. RELATED WORK

In this section we give a brief overview of the latest work related to the use of machine learning techniques in the cloud.

Statistical driven modeling and its application to data intensive workloads is presented in [7]. The authors use statistics to predict resource requirements for cloud computing applications, which can be used for making decisions including job scheduling, resource allocation, and workload management. The authors in [8] propose machine learning approach to predict system performance for future configurations and workloads, and find a control policy that minimizes resource usage while maintaining performance.

Cloud environments benefit from intelligent virtualized resources. Xiong et al. in [9] propose intelligent virtualized resources management solution in their machine learning powered SmartSLA.

Additionally, cloud computing is often hosted on energy consuming data centers. Therefore, energy consumption optimization is another cloud computing and data center issue. The authors in [10] adopt machine learning methods for data center workload, thermal distribution and cooling facilities management.

3. THE METHODOLOGY

In this section we present the methodology we developed, including the methods used for obtaining realistic data necessary for the machine learning approach and the machine learning process itself.

A. Workload Data Simulation

Our paper presents statistical procedure which for particular input (load) suggests the best virtual machine (VM) configuration. In order to obtain reliable suggestion, CSPs must ensure a realistic environment where nothing is homogenous. Thus, an appropriate surrounding would be the one with various cloud settings, hosting web services with different characteristics, loading the servers with variable load, and a tool to measure the performance which involves many parameters as response time, flops, tps, throughput, bandwidth, CPU, RAM, etc.

As an example of a reliable setting we present a case where client-server architecture is deployed on some cloud platform, and a virtual machine manager (VMM) is used to instantiate VM instances. Thereto, CSP chooses

particular application server and operating system for the client-server platform and the VM instances. Simulating various user demands, CSP must define a few cloud environments, all with different number of VMs. After defining the technical framework, CSP needs a software tool for generating load, different number of messages changing in size, and a tool for capturing the performance measurements.

Once the CSP collects enough data from various test cases, it can proceed to applying machine learning analysis.

B. Data Preprocessing

The data obtained from the CSP's testing environment is raw and inclined to noise. The source of noise can be either the phenomenon that the same VM on the same hardware at different times among the other active VMs not always achieves the same performance [11], or some other causes as network latency, network throughput, etc. In order to remove any unexpected variability and picks, the raw data needs to be pre-processed.

Pre-processing is essential for transforming the parameter values into a new space of variables suitable for classification. The first step in eliminating the noise and the outliers from the data is to use appropriate smoothing method. This technique also handles missing values which are not unknown problem when this kind of testing is performed.

Once we cleaned the data, we need to normalize the values to fall in between particular interval. When selecting appropriate normalization method, we take into account that the parameters are measured in different units. Thus, we need a method which will normalize the values, but still preserving all the relationships in the data.

Since the complexity of any classifier depends on the number of inputs, the next step is to handle the dimensionality problem. A correlation-based or entropy-based analysis method can be used to perform attribute relevance analysis and filter out statistically irrelevant or weakly relevant attributes from the descriptive mining process, and thus, finding many interesting relationships among data [12]. Now the data is collection of multidimensional vectors that consist of particular parameters, and according to which different VM configurations can be distinguished.

Once the raw data is pre-processed, it is prepared for the classification process.

C. Machine Learning Approach

Since the data is labeled, i.e. for each input vector we know the desired output, we can use supervised machine learning techniques. When using a supervised learning technique, the classifier learns from a training data. The training data needs to be carefully selected so that the overfitting classification issue must be avoided. If overfitting occurs, that means that the classifier got to very biased to the training data, and the classifier's accuracy decreases when new unknown data arrives. Thus, the cross-validations technique seems to be appropriate when choosing the training and the testing set [13].

When choosing appropriate classification technique, we must be aware of the multiclass problem. Since we aimed to choose the parameters that make best distinction among classes, we assume that our problem is linearly separable. Therefore, neural networks tend to perform much better when dealing with multidimensions and continuous features [14]. If we believe that the data is poor, we may additionally try the bootstrapping method which multiplies the data few times and acts like more experiments are performed. It supposes to improve the accuracy and the stability of the classifier. After the classifier is trained, it is ready to classify new unknown load inputs. The neural networks process is described in Fig. 1.

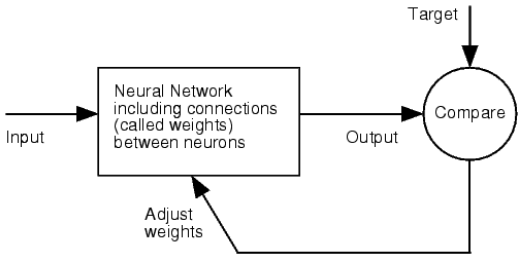

Figure 1. Neural Network classification process [15]

In addition, it is very important to add one more classification level which classifies the loads into two more classes, which we refer to as high demand, and economy class. The first needs more VMs to ensure better performance, and the second follows the lower cost.

We believe that the methodology we defined is very useful and is beneficial for both the customers and the CSPs.

4. CONCLUSION

Recently, cloud service providers offer pay-per-usage pricing model, which charges the customers according to the amount of rented resources. We believe that for customers' contentment it is very important cloud service providers to provide a guarantee of sustainable performance, which unfortunately misses in the Service Level Agreements.

This intrigued us to think of developing a new model which will focus on charging according to performance, instead of according to rented VMs. Therefore, in this paper we propose a new methodology, which intends to allow the cloud service providers to predict the cloud configuration needed for a certain load. We believe that in a well formed testing environment we can collect enough data that can be used in an accurate prediction of real demands. As we defined the surroundings under which reliable data is obtained, we proceeded to develop a new methodology based on machine learning techniques. Assuming the nature of our problem, we carefully picked up set of techniques, which we expect to lead to a very precise classification of what cloud configuration is most likely to satisfy the customer needs for performance. Moreover, our approach aims to make a good trade-off between the high demanding configuration that requires high performance, and the economy one that ensures low cost.

In our future work we will implement this methodology on real data, and we will improve the eventual weaknesses.

REFERENCE

[1] P. Mell and T. Grance, "The NIST definition of cloud computing," National Institute of Standards and Technology, Information Technology Laboratory, Sep. 2011.

[2] Fox, R. Griffith et al., "Above the clouds: A Berkeley view of cloud computing," Dept. Electrical Eng. and Comput. Sciences, University of California, Berkeley, Tech. Rep. UCB/EECS, vol. 28, 2009.

[3] R. Grossman, "The case for cloud computing," IT professional, vol. 11, no. 2, pp. 23–27, 2009.

[4] R. Buyya, C. S. Yeo and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," High Performance Computing and Communications, 2008. HPCC'08. 10th IEEE International Conference on.

[5] K. Xiong and H. Perros, "Service performance and analysis in cloud computing," Services-I, 2009 World Conference on, IEEE.

[6] D. Durkee, "Why cloud computing will never be free," Queue, vol. 8, no. 4, p. 20, 2010.

[7] A. S. Ganapathi, Y. Chen, A. Fox, R. Katz and D. Patterson, "Statistics-driven workload modeling for the cloud," Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on.

[8] P. Bodik, R. Griffith, C. Sutton, A. Fox, M. Jordan and D. Patterson, "Statistical machine learning makes automatic control practical for internet datacenters," Proceedings of the 2009 conference on Hot topics in cloud computing, HotCloud.

[9] Xiong, Pengcheng, Yun Chi, Shenghuo Zhu, Hyun Jin Moon, Calton Pu, and Hakan Hacigumus. "Intelligent management of virtualized resources for database systems in cloud environment." In Data Engineering (ICDE), 2011 IEEE 27th International Conference on, pp. 87-98. IEEE, 2011.

[10] Chen, H., Kesavan, M., Schwan, K., Gavrilovska, A., Kumar, P., & Joshi, Y., "Spatially-Aware Optimization of Energy Consumption in Consolidated Data Center Systems," 2011 Proceedings of InterPACK, Portland, OR.

[11] Y. Koh, R. Knauerhase, P. Brett, M. Bowman, Z. Wen, and C. Pu, "An analysis of performance interference effects in virtual environments," in Performance Analysis of Systems Software, 2007. ISPASS 2007. IEEE International Symposium on, pp. 200 –209.

[12] E. Alpaydin, "Introduction to Machine Learning," MIT Press, Cambridge, MA, 2010.

[13] J. Han and M. Kamber, "Data Mining, Second Edition," University of Illinois, Elsevier, 2006.

[14] S. B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," 2007.

[15] http://matlab.izmiran.ru/help/toolbox/nnet/01_nnbla.gif