A Forecasting Methodology for Workload Forecasting in Cloud Systems

Francisco J. Baldán[1], Sergio Ramírez-Gallego[1], Christoph Bergmeir[2], Francisco Herrera[1], and José M. Benítez[1]

[1]Dept. of Computer Science and Artificial Intelligence, University of Granada, CITIC-UGR, Granada 18071, Spain.
[2]Faculty of Information Technology, Monash University, Melbourne, Australia.
fjbaldan@decsai.ugr.es, sramirez@decsai.ugr.es, christoph.bergmeir@monash.edu, herrera@decsai.ugr.es, J.M.Benitez@decsai.ugr.es

Abstract. This is a summary of our article accepted in IEEE Transactions on Cloud Computing [1] to be part of the MultiConference CAEPIA'16 KeyWorks.

Keywords: Cloud Computing, Elasticity, Workload forecasting, Machine learning, Time series forecasting

Summary

Cloud Computing (see, e.g., [3] for an introduction) is an essential paradigm of computing services based on the "elasticity" property, where available resources are adapted efficiently to different workloads over time. A user of a cloud platform will not need to acquire all the resources initially, but requires a variable amount of resources which will be provided or released by the provider in a dynamic fashion, according to the actual demand. Thus, resources are provided in an elastic manner [5]. In elastic platforms, the forecasting component can be considered by far the most important element and the differentiating factor when comparing such systems, with workload forecasting one of the problems to solve if we want to achieve a truly elastic system.

There are multiple related work in the literature on the elastic cloud systems topic and how these proposals address the specific problem of workload forecasting through a wide range of forecasting techniques [2]. When properly addressed the cloud workload forecasting problem becomes a really interesting case study. As there is no general methodology in the literature that addresses this problem analytically and from a time series forecasting perspective (even less so in the cloud field), we propose a combination of these tools based on a state-of-the-art forecasting methodology which we have enhanced with some elements, such as: a specific cost function, statistical tests, visual analysis, etc.

The insights obtained from this analysis are used to detect the asymmetrical nature of the Cloud system Workload Forecasting (CWF) problem and to find

Actas de la XVII Conferencia de la Asociación Española para la Inteligencia Artificial, pp. 765-766

the best forecasting model from the viewpoint of the current state of the art in time series forecasting.

As there is no general framework or methodology in the literature that addresses the CWF problem analytically from a time series point of view to analyze and predict time series data described in Hyndman and Athanasopoulos [4], we propose a combination of tools based on a state-of-the-art forecasting methodology, which we have enhanced and completed. We include, e.g., non-linear and ML forecasting procedures and statistical tests for linearity to determine when to use these more complex procedures. We can summarize this methodology in the following 3 steps: first, we visualize the time series, and analyze Auto-Correlation Function (ACF) and Partial Auto-Correlation Function (PACF) plots. Second, we perform a non-seasonal study, constructing ETS and ARIMA models as a first approach, and using the obtained results to build other regression models. Finally, we perform a similar study but focused on seasonality. This methodology comprises a deeper study through the analysis and the modeling of workload series to better understand the intrinsic properties of the series.

Since the error in CWF is clearly asymmetric we have proposed a new measure to assess the cost of under-provisioning and over-provisioning. These measures allow for a better visualization of the performance of the elastic provisioning module of cloud platforms. From an operational point of view the most interesting forecast is a short-time horizon, so we focus on this. To show the feasibility of this methodology, we apply it to several realistic workload datasets from different datacenters. The results show that the analyzed series are non-linear in nature, and that no seasonal patterns can be found. Moreover, on the analyzed datasets, the penalty cost as usually included in the SLA (Service Level Agreement) can be reduced to a 30% on an average.

Acknowledgments

This work was partially supported by the Spanish Ministry of Science and Technology under projects TIN2011-28488, TIN2013-47210, and the Andalusian Research Plans P11-TIC-7765, P10-TIC-6858 and P12-TIC-2958.

References

1. F. J. Baldán, S. Ramírez-Gallego, C. Bergmeir, F. Herrera, and J. M. Benítez. A forecasting methodology for workload forecasting in cloud systems. *IEEE Transactions on Cloud Computing*, **Accepted on 05-May-2016**.
2. G. Box and G. Jenkins. *Time series analysis: Forecasting and control*. Holden-Day, 1970.
3. R. Buyya, J. Broberg, and A. M. Goscinski. *Cloud Computing Principles and Paradigms*. Wiley Publishing, 2011.
4. R. Hyndman and G. Athanasopoulos. Forecasting: Principles and practice, 2013.
5. K. Konstanteli, T. Cucinotta, K. Psychas, and T. Varvarigou. Elastic admission control for federated cloud services. *IEEE Transactions on Cloud Computing*, 2(3):348–361, 2014.