IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 24, No. 4, August, 2019, pp. 385-389
Journal homepage: http://iieta.org/journals/isi

Performance and Cost Analysis of Web Application in Elastic Cloud Environment

Suneetha Bulla[1*], Bobba Basaveswara Rao[2]

[1] Computer Science and Engineering, Acharya Nagarjuna University, Guntur 522510, India
[2] Computer Center, Acharya Nagarjuna University, Guntur 522510, India

Corresponding Author: suneethabulla@gmail.com

https://doi.org/10.18280/isi.240404

Received: 4 March 2019
Accepted: 10 June 2019

Keywords:
cloud computing, single class of service, Amazon AWS, e-commerce

ABSTRACT

In the era of cloud computing, the cost analysis of web applications on the cloud is critical to the quality and profit of cloud services. This paper designs an evaluation model for the performance and cost of web applications that provide a single class of services. The model was developed based on the analytical queuing model, and was provided with multiple evaluation metrics. Next, an experiment was carried out on Amazon AWS, a mature commercial cloud platform. The experimental results agree well with the simulation data of the proposed model in both response time and total cost. The research results provide new insights into the development of e-commerce web applications in the environment of cloud computing.

1. INTRODUCTION

In the IT industry, Cloud Computing (CC) is one of the dominant technologies in the real time/online applications and has become one of the fastest growing because of several organizations are moved from local computing infrastructure to cloud infrastructure for reducing the physical resources costs. CC has identified by Gartner as one of the top 10 technologies and declared that CC plays an important role in profits of organizations (www.gartner.com/us/symposium). This is an Internet oriented computing where cloud resources like software's, hardware infrastructure, platform, devices and web services on a pay-as –you-go basis. Customers of the cloud adopt both hardware and software virtual resources from service providers on payment basis as they are utilized instead of, they do not much investment on resources. CC infrastructures provide three types services through centralized data centers and host web applications [1].

NIST (National Institute of Standards and Technology) defined CC is a technology for enabling the on-demand, flexible and global network in and out access to a allocate pool of configurable computing infrastructure like web servers, applications, networks, services and database for storage, which can be allocated and released with the smallest service provider interactions or the management efforts [2]. CC has number of features enable it to provide service to its users effectively. Cloud features include flexibility, on-demand self service provisioning and elasticity (http://www.govtech.com/gt/articles/387269).

To analyze the dynamic nature of CC, to study the scheduling algorithms and QoS metrics with considering SLA are important issues for upcoming research area. These research objectives are improving usability, scalability and availability nature of the cloud application. Because of the elasticity virtual servers are do not left idle, which can reduce costs for the customers while increasing the application speed and also availability of the resources on a cloud. Now a day's E-Commerce playing vital role in both the B2B & B2C global market environment. Technology development shows results to increase in the number of smart mobile phones, network connections, and an online payment tools are leading this growth in online shopping.

Thus Web Application (WA) hosting is major issue for providers and they must be research on the performance and maintenance setup/running cost of the web application. There are many cloud providers like Amazon Web Services (https://aws.amazon.com), Google Cloud Platform (https://cloud.google.com/), IBM Cloud (https://www.ibm.com/cloud/) and Oracle Cloud (https://cloud.oracle.com/home) are given web application hosting services. Amazon AWS is one of the leading service providers, so in this experimental study is considered. Several authors [3-9] are studied the performance analytically by using queuing models and without queuing models. Not much work being to be done to study the performance with experimental test beds and cost point of view. This work fulfills this gap with real time web application hosting on Amazon AWS.

This paper evaluates an analytical model to finds the performance metrics then analyze the cost factor also. This analytical model was verified by the experimental model. This methodology to analyze the performance and cost of web application on single-class type cloud services, that means only one service can be serviced by datacenter. The experiment was conducted on AWS cloud platform. This model considers performance metric and costs resulting from setup and maintenance. The obtained analytical results are validated with the experimental results.

The rest of this paper is formatted as follows, in section 2 discussed about various author's contribution on the web application hosting. The fundamentals and the evaluated analytical queuing model representation are disused in section 3. The proposed experimental architecture model and test-bed design were given in Section 4. The results are depicting in Section 5 and finally section 6 discussed about the conclusions.

2. LITERATURE SURVEY

From the last decade the CC usage is increasing geometric progression. Several authors are published their work on CC performance as well as cost analysis of web application hosting on cloud. The research on this area mainly focuses on two fold, i) analytical models using statistical or queuing models, ii) experimental study also done using Amazon AWS. This section presents these two sections briefly.

Shaw [8] studied about the performance of WA. This paper concludes by using predictions to avoid problems in the performance and finally he specifies high configured physical infrastructure and supported software's are not producing a good latency. Similarly Samad et al. [9] discussed about performance of WA with respect of response time based on the speed of the server.

Amazon's AWS platform provides a financial substitute for different types of CC research. Garfinkel [10] conducted an experiment on the AWS to calculate the performance of different cloud services. This author taken date throughput and request per second are main parameter in his research.

A queuing model was implemented to study the performance of web service on a cloud datacenter. In that IaaS service is modeled as multiple parallel queues to determine the virtual instances in the centralized datacenter [11]. Hosted a web application in the cloud and implemented analytical model. The author validated analytical model using simulation and concluded there is no significant difference on the performance metrics.

Iosup et al. [3] discussed about the performance analysis results of CC by using for Many-Tasks Scientific Computing. These results are reported by Jackson et al. [4]. Many authors are contributed their research on the elastic cloud and its Challenges and Opportunities are discussed by Kumar Buyya et al. [6]. The performance of web service application, stored data and setup and maintenance costs of the Montage workflow on cloud are detailed by Deelman et al. [7].

3. ANALYTICAL MODEL FOR WEB APPLICATION IN CLOUD COMPUTING

The e-commerce growing day by day, many organizations are showing interest implement their business in the online market also. In this scenario WA is the major source to project business details. Before going to host application have to analyze the performance and setup and maintenance cost of the WA. This paper adobe the Amazon AWS services and implements analytically and experimentally.

Figure 1 show workflow of web application and it is hosted on CC providers like Amazon Web Application Hosting (http://docs.amazonwebservices.com/gettingstarted/latest/wah/web-app-hosting-intro.html?r=1052). This architecture is a combination of different services; those are Elastic Load Balancer (ELB), Elastic Cloud (EC), and RDS database. Work of the ELB is distributes the http load spike or traffic rate to a group of EC2 instances in the elastic computing tier [12, 13] (http://aws.amazon.com/elasticloadbalancing/). VM instances are grouped and registered in elastic cloud to which user's configuring triggers. These triggers are generating dynamically increase EC2 resources based on network or utilization of the cloud. These conditions are check by an AWS CloudWatch service. WA hosted VM instances are run parallel in the centralized data service centers, each VM have queue to provide service to the client requests [14]. The WA architectures have a storage and used as repository for logging and storing data.

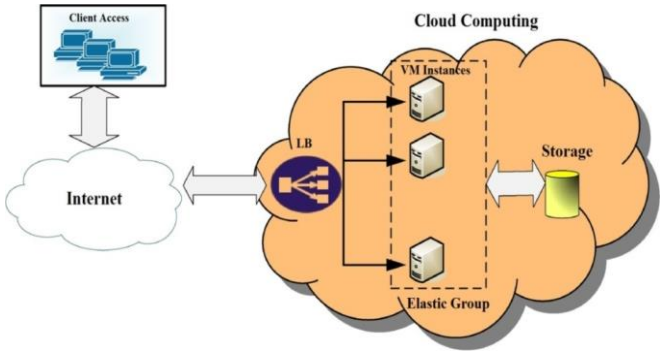

Figure 1. Web application hosting architecture

Figure 2 shows analytical queuing model and it is drawn from Figure 1. This architecture follows model of M/M/1 open queue model, where λ represents the total arrival rate and μ service rate of one instance, S indicates total number of running instances in the elastic cloud. The WA assumed that arrival rate and service rate follows a Poisson distribution, loss probability of the queuing model is zero, the effective arrival rate λ̂ =λ and μi = μ. In the elastic cloud instances are running on the parallel computing.

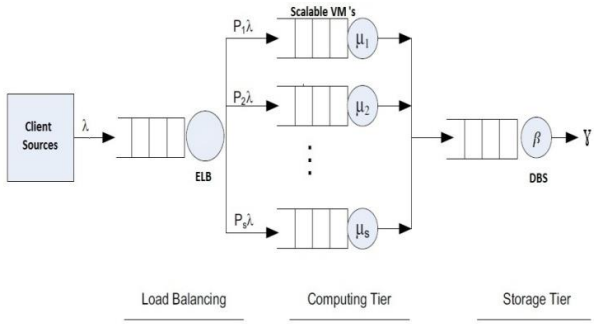

Figure 2. Queuing model of WA

The response time is a one of the important performance metrics. In this paper latency considered as metric to show performance of the WA. The above figure contains open queuing network model with S parallel number of M/M/1 queues of the architecture [15]. Since the ELB share the requests among the S running instances evenly, the probabilities routing matrix will be generated, these are equal to 1/S and where Pi is the routing probability of *i*th VM of the elastic cloud. As a result, each instance arrival rate: $\Lambda_i = \lambda/S$.

Based on M/M/1 queuing theory the average delay of a client request of each instance can be calculate to be:

$$\bar{T} = \frac{1}{\mu - \Lambda i}$$

Average delay of the network can be calculated by Little's formula $\bar{T} = \frac{1}{\lambda}\sum_{i=1}^{S} \frac{\Lambda i}{\mu - \Lambda i}$.

Assuming that based on the auto scaling of AWS there is no delay to generate new instance in the elastic cloud, the average response time of the request in the queuing model:

$$R = \frac{S}{S\mu - \lambda} \quad (1)$$

If the upper threshold utilization is hundred percent then the total number of servers:

$$S_{required} = \frac{\lambda}{\mu} \quad (2)$$

If the upper threshold utilization is eighty percent then

$$S_{required} = 1.25 * \frac{\lambda}{\mu} + 1 \quad (3)$$

The total setup and maintenance cost can be depicted as follows:

$$COST = (Price_{bw} * \lambda_{GB/s} + Price_{com} * S)\, T \quad (4)$$

where, $Price_{bw}$ is bandwidth cost, $\lambda_{GB/s}$ is arrival network, $Price_{com}$ is computing cost, S is total number of running servers and T is total time.

4. EXPERIMENTAL MODEL FOR WEB APPLICATION IN CLOUD COMPUTING

Figure 3 shows experimental model cloud structure has been taken and designed from most CC infrastructure providers like AWS cloud environment [13]. In this test-bed the main components are EC2, RDS, S3, Route 53 and CloudWatch.

This experiment has been tested US West Origen on Amazon AWS. The results are noted around 03:00 am early morning of UTC time. For this analysis created micro word press EC2 instance in the EC2 console and attached my sql micro database server for storing the network traffic with multiple A-Z basis, this feature clones the database servers to all regions for availability purpose. Auto scalling group was configured with minimum and maximum are 2 and 10. The WA size 580 bytes and it fulfill 100 % utilization of VM when 100 req/sec http loads occur. The ELB was configured with scale up and scale down rules, those are when if the upper threshold value of the elastic cloud exceeds 80 % the it scales up one server similarly when the lower threshold value below 30 % the scale down one server. The Route53 is used to provide DNS name to the web application and S3 was configured for storing the logs of the clients

The experiment starts with two servers and arrival rate increase randomly varying from 200 to 1200 for every 12 minutes. Each request touches the three stages of the architecture. The clients are visit web application through the internet and Route53 and generate the http traffic to the web application. The load balancer distributes the traffic among all available instances in the cloud to provide service. The WA hosted in the scalable cloud if it exceeds the 100 % CPU utilization of the VM then automatically generate a new instance and registered in the load balancer to distribute the requests at the same manner if it is below 30 % of the CPU utilization then delete an instance from load balancer.

The Traffic Rate varied from 200 to 1200 requests per second using http traffic generator. The cost factor been calculated using Eq. (4). In that function computing price is $0.115 as it is taken from Amazon for t2. micro on-demand instances (https://aws.amazon.com/). The bandwidth price of the EC2 instance is $0.01 per GB in/out data transferred based on data transferred “network-in” and “network-out” of Amazon EC2.

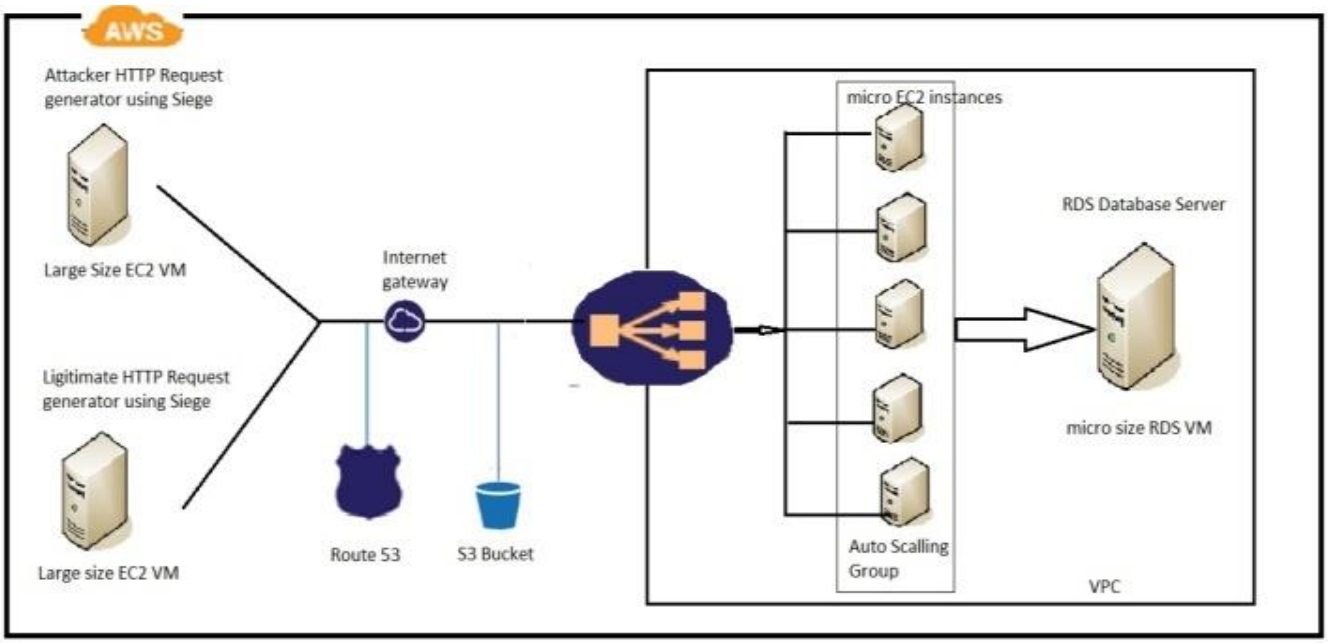

Figure 3. Experimental test-bed for hosting web application in the Amazon AWS

5. RESULTS

Figure 4 depicts the experimental and analytical evaluation of the response time of user requests. The obtained results show that when the load of the traffic increases, the corresponding response time also increases. This experiment used scalability of elasticity to increase instances when the high load spike occurs. Based on the elasticity of the cloud,

the letancy does not affect more when the high traffic rate occurs. The analytical results show same upward increment whenever traffic increases without any fluctuations whereas empirical results follow same type of increasing pattern with fluctuations.

Figure 5 shows the costs behavior of web application deploys on cloud for different user requests for both analytical and experimental results. When the Traffic rate increases then the computing resources and the network utilization also increases. Thus the cost of the infrastructure and instances are increased. Both analytical and experimental results are very close. The analytical results show same upward increment whenever traffic increases with fluctuations whereas empirical results follow same type of increasing pattern with fluctuations.

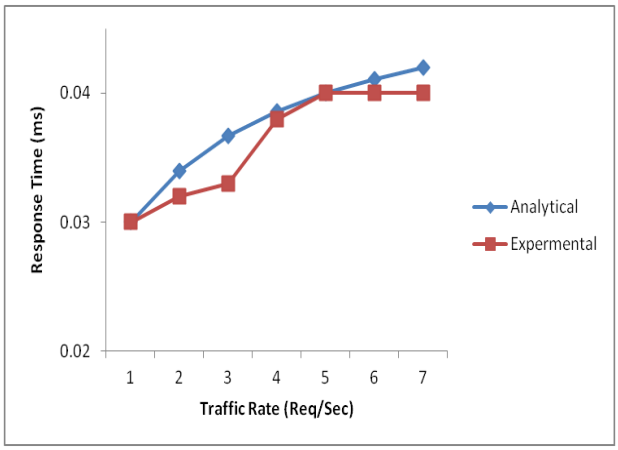

Figure 4. Response time of the client request

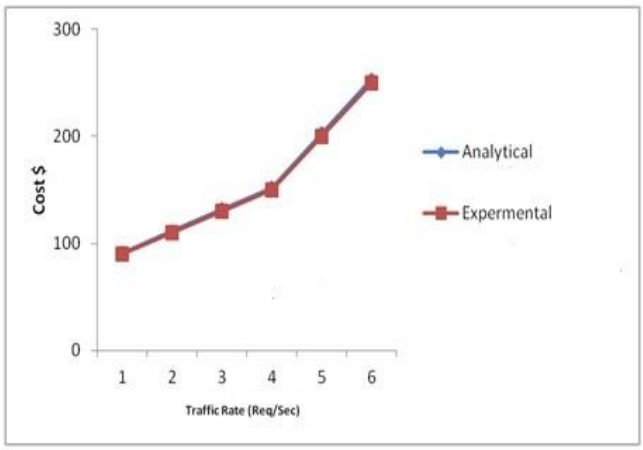

Figure 5. Total cost of the cloud setup

6. CONCLUSION

The paper evaluated analytical model to analyze the performance and cost of the web application hosting on the CC. these analytical results are supported or enhanced with obtained experimental results. The experiment was conducted on Amazon AWS. The obtained experimental test-bed results have various performance metrics have been agreed with the analytical results, with maximum error of 0.1 %. There is no significant difference between these two results. As per the results, the traffic rate has a notable effect on the performance of the cloud services, such as end-to-end response time affected with unacceptable delay of the user's requests. By observing the results to find that both computing and bandwidth costs go to high when the traffic rate increases. As a future work, to study the attacks activity on the elastic cloud by analytical queuing model and mitigation of various attacks an Experimental test-bed.

REFERENCES

[1] Kavis, M.J. (2014). Architecting the Cloud: Design Decision for Cloud Computing Service Models (SAAS, PAAS and IAAS). Wiley India Private Limited, 2014 edition.

[2] Mell, P., Grance, T. (2011). The NIST Definition of Cloud Computing. NIST, Special Publication 800-145. https://doi.org/10.6028/NIST.SP.800-145

[3] Iosup, A., Ostermann, S., Nezih Yigitbasi, M., Prodan, R., Fahringer, T., Epema, D. (2011). Performance analysis of cloud computing services for many-tasks scientific computing. IEEE Transaction on Parallel and Distributed Systems, 22(6): 931-945. https://doi.org/10.1109/TPDS.2011.66

[4] Jackson, K.R., Ramakrishnan, L., Muriki, K., Canon, S., Cholia, S., Shalf, J., Wasserman, H.J., Wright, N.J. (2010). Performance analysis of high performance computing applications on the amazon web services cloud. Proc. IEEE Second International Conference on Cloud Computing Technologies Science, Indianapolis, IN, USA. https://doi.org/10.1109/CloudCom.2010.69

[5] Stantchev, V. (2009). Performance evaluation of cloud computing offerings. Proc. 2009 Third International Conference on Advanced Engineering Computing and Application in Sciences, Sliema, Malta, pp. 187-192. https://doi.org/10.1109/ADVCOMP.2009.36

[6] Kumar Buyya, R., Ranjan, R., Calheiros1, R.N. (2009). Modeling and simulation of scalable cloud computing environments and the CloudSimToolkit: Challenges and opportunities. Proc. of International Conference on High Performance Computing & Simulation, Leipzig, Germany, pp. 1-11. https://doi.org/10.1109/HPCSIM.2009.5192685

[7] Deelman, E., Singh, G., Livny, M., Berriman, B., Good, J. (2008). The cost of doing science on the cloud: the montage example. Proceedings of the 2008 ACM/IEEE conference on Supercomputing. IEEE Press, Austin, TX, USA, pp. 1-12. https://doi.org/10.1109/SC.2008.5217932

[8] Shaw, J. (2000). Web application performance testing - a case study of an on-line learning application. BT Technology Journal, 18(2): 79-86. https://doi.org/10.1023/A:1026732502654

[9] Samad, H., Hanizan, S.H., Din, R., Murad, R., Tahir, A. (2018). Performance evaluation of web application server based on request bit per second and transfer rate parameters. Journal of Physics: Conference Series, 1018 012007. https://doi.org/10.1088/1742-6596/1018/1/012007

[10] Garfinkel, S.L. (2007). An evaluation of Amazon's grid computing services: EC2, S3 and SQS. Technical Report TR-08-07, Harvard University.

[11] Addamani, S., Basu, A. (2013). Performance analysis of web applications on IaaS cloud computing platform. International Journal of Computer Applications, 64(15).

https://doi.org/10.5120/10707-5668
[12] Buyya, R., Ranjan, R., Calheiros, R.N. (2010). InterCloud: Utility-oriented federation of cloud computing environments for scaling of application services. In: The 10th International Conference on Algorithms and Architectures for Parallel Processing, Busan, Korea, pp. 13-31. https://doi.org/10.1007/978-3-642-13119-6_2
[13] Bellenger, D., Bertram, J., Budina, A., Koschel, A., Pfänder, B., Serowy, C. (2011). Scaling in cloud environment. In: Proceedings of the 15th WSEAS International Conference on Computers, Wisconsin, pp. 145-150.
[14] Idziorek, J. (2010). Discrete event simulation model for analysis of horizontal scaling in the cloud computing model. In: Proceedings of the 2010 Winter Simulation Conference, Baltimore, MD, USA, pp. 3004-3014. https://doi.org/10.1109/WSC.2010.5678994
[15] Morein, W.G., Stavrou, A., Cook, D.L., Keromytis, A.D., Misra, V., Rubenstein, D. (2003). Using graphic turing tests to counter automated DDOS attacks against web servers. CCS '03 Proceedings of the 10th ACM Conference on Computer and Communications Security, pp. 8-19. https://doi.org/10.1145/948109.948114