2017 IEEE International Symposium on Parallel and Distributed Processing with Applications and 2017 IEEE International Conference on Ubiquitous Computing and Communications

An Evaluation Methodology of Cloud Service Quality Based on Queuing Model

Gengzhe Yan
School of Computer Science & Technology
Harbin Institute of Technology
Weihai, China
yangengzhe@163.com

Fanchao Meng, Dianhui Chu
School of Computer Science & Technology
Harbin Institute of Technology
Weihai, China
fcmeng@hit.edu.cn, chudianhui@vip.sina.com

***Abstract*—With the rapid development of Internet technology, online applications have become an indispensable part of government, enterprise and engineering fields. Most enterprises have utilized the advantages of cloud computing and refactored the applications with the service-oriented architecture (SOA) in order to solve the huge number of online users, soaring data and fluctuant load. According to the cloud services quality which evaluated in different cloud computing resources, developer choose the suitable size of the service cluster to meet the needs of enterprise services in the distributed service cluster. It can not only ensure the stable operation of online applications, but also maximize the reduction of basic costs. Therefore, it is important to evaluate the quality of cloud services accurately. This paper proposes an evaluation method of cloud service quality based on queuing model. The M/M/1 and M/M/s queuing models are used to analyze and evaluate the quality of single node service and cluster service respectively. By the method proposed in this paper can not only choose the suitable size of the service cluster for online applications, but also perform dynamic scheduling of cloud resources. Finally, the feasibility and practicability of this method is verified by experiments.**

Keywords-cloud service; queuing model; QoS; SOA

I. INTRODUCTION

With the rapid development of cloud computing, most of the online applications in order to cope with high concurrent load, complex business logic and to ensure their own stability are using a service-oriented architecture (SOA) to achieve. The basic elements of the SOA architecture are services, which are standalone, modular and loosely coupled. Through services are called from each other, dismantling complex online applications into a large service network, and deploying different services independently different virtualization clusters in order to make the overall performance of a strong throughput, processing power and fault tolerance [1]. Furthermore, the online application of the load has a strong volatility, if the cloud service allocated large cloud resources will be a waste, resulting in increased costs; if the distribution of cloud resources is small, it will affect the quality of cloud services, resulting in revenue loss. Therefore, according to the effective evaluation of cloud services quality to reasonably allocate cloud resources has become a key of service-oriented architecture (SOA).

In view of the above problems, this paper proposes a method to evaluate the quality of cloud services. Firstly, the cloud service resource consumption in heterogeneous cloud environment is modeled, and the relationship between cloud quality and cloud resource consumption is described quantitatively. Secondly, analyze the cloud service resource consumption by queuing model, because the online application is also a random service system. Finally, the quality of service (QoS) which can be guaranteed under different cloud resources is obtained. Therefore, the evaluation results of method proposed in this paper can be a reasonable allocation of cloud resources for cloud services, so as to maximize the reduction of base cost and protect the quality of cloud services.

The remainder of the paper is structured as follows. Section II describes the related work. Section III introduced the cloud service quality evaluation framework. Section IV introduces the cloud service quality requirements and resource consumption model. Section V gives the method of cloud service quality evaluation based on queuing model. Section VI validates the proposed method through comparative test analysis. Finally, Section VII draws the concluding remark and summarizes the full paper.

II. RELATED WORK

The application of online applications will consume resources such as CPU, memory and network bandwidth. There are two ways to get the resource consumption metrics online: direct measurement and indirect evaluation. In the traditional server environment, complete measurements of cloud resource consumption by monitoring the CPU, memory, network bandwidth and other direct measurement [2, 3]. However, the cloud computing environment using a large number of virtualization technology, which cannot directly measure its exact value. Therefore, we always use indirect evaluation method to measure the consumption of online applications instead of direct measurement [4].

The method of indirect evaluation mainly uses the method of mathematical analysis to estimate the consumption of resources [5]. The commonly used mathematical analysis methods used to estimate the resource consumption are regression analysis, Kalman filter and queuing theory. Regression analysis is an effective tool to study the close relationship between variables, structural states and model predictions by establishing statistical models [6]. Kalman filter is a linear system equation, through the system input and output observation data, the optimal state of the system algorithm [7]. Queuing theory is the study of the phenomenon of stochastic clustering and the mathematical theory and method of stochastic service system work [8].

0-7695-6329-5/17/31.00 ©2017 IEEE
DOI 10.1109/ISPA/IUCC.2017.00142

At present, most of the research on the queuing model is based on the performance model to estimate the resource consumption method [9], or only consider the single node under the SLA standard [10, 11]. It is difficult to ensure the efficiency of operation and distributed largescale application availability. Therefore, this paper proposes a method of queuing theory that uses the queuing theory to map the load, execution time and concurrent user number of the service in the online application system to M/M/1 and M/M/s of the queuing model arrival rate, service time and average queue length. Through the research of the relationship between the queuing theory, according to the rule of these relations constitute the service indicators, respectively, to evaluate the cloud services in a single node and cluster environment quality.

III. CLOUD SERVICE QUALITY EVALUATION FRAMEWORK

The quality of cloud service is the requirement of the performance of online application from the user's opinion. There are three important indexes: load, response time and reliability [12]. Respectively, depicts the rules of the user request to arrive, the cloud service response results of the time interval and the cloud service can correctly respond to the user request probability.

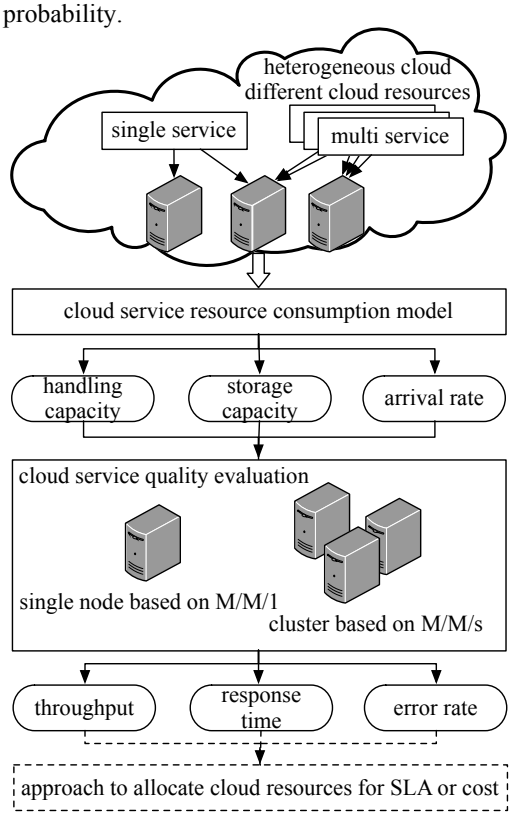


Figure 1. Cloud service quality evaluation framework.

In the heterogeneous public cloud, different hardware environments have different effects on the consumption of resources. Therefore, it is necessary to construct different measurement models for virtual units with different configuration types to ensure the quantitative description of cloud service quality and cloud resource consumption, is the prerequisite and basis for effective cloud resource supply. As is shown in Figure 1, resource consumption parameter is derived by using the cloud service resource consumption model for the single service or multi service, and as input to method of cloud service quality evaluation. Finally, derive parameters which the cloud service quality in the single node or cluster, and then according to the parameters to allocate cloud resources for SLA or cost.

IV. CLOUD SERVICE RESOURCE CONSUMPTION MODEL

In order to be able to characterize the essential characteristics of cloud resource consumption from the service level, this paper propose the concept of handling capacity and storage capacity, which is a high-level abstraction of node computing resources and storage resources. Handling capacity refers to the average number of user requests completed by a virtual unit per unit time; storage capacity is the maximum number of concurrent service requests that a virtual unit can receive [13].

A. Single Service Node Resource Consumption Model

Let i is a group of services, let j is a virtual node, let Max_j is the maximum amount of available memory for the virtual node j, T_{ij} is the time i spent on j, M_{ij} is the memory usage of i on j. T_{ij} and M_{ij} in the heterogeneous cloud environment has a different value, but in one cloud environment, can be tested or real-time monitoring of the way to obtain the fixed value of the variable.

In unit time, n_{ij} is the average number of times i executes on j, m_{ij} is the average memory consumed by each execution in Eq. (1).

$$n_{ij} = \frac{1}{\mathrm{E}(\mathrm{T}_{ij})}, m_{ij} = \mathrm{E}(\mathrm{M}_{ij}) \tag{1}$$

Where $\mathrm{E}(\mathrm{T}_{ij})$ is average time to execute i on j. $\mathrm{E}(\mathrm{M}_{ij})$ is average memory usage to execute i on j. According to Eq. (2), calculate the handling capacity μ_{ij} and storage capacity c_{ij}.

$$\mu_{ij} = \tau_{\mathrm{j}} \cdot n_{ij}, c_{ij} = \left\lfloor \theta_j \cdot \frac{Max_j}{m_{ij}} \right\rfloor \tag{2}$$

Where τ_j is handling capacity coefficient which is related to the CPU schema of j. θ_j is storage capacity coefficient which is related to the storage schema of j. Their value can be determined by experiment. μ_{ij} is the average number of times i executed on j in unit time. c_{ij} is the maximum amount of concurrent requests for i executed on j in unit time.

B. Multi Service Node Resource Consumption Model

A multi service node is a virtualized node j that provides more than one different group of services. Let I is a multi service set, and all services in I are deployed to j. according to Eq. (3), the load weight of $i(i \in I)$ can be approximated to $\propto_{ij}$.

$$\propto_{ij}=\frac{n_{ij}}{\sum_{i\in I} n_{ij}} \tag{3}$$

Where n_{ij} is still the average number of times *i* executes on *j*, the value can be calculated by a single service method which mentioned above. According to Eq. (4), calculate the handling capacity μ_j and storage capacity c_j.

$$\mu_j = \frac{\tau_j}{|I|\sum_{i\in I}\frac{\propto_{ij}}{T_{ij}}}, c_j = \frac{\theta_j \cdot Max_j}{|I|\sum_{i\in I}\propto_{ij} m_{ij}} \tag{4}$$

Where $|I|$ is the number of services in *I*.

V. Cloud Service Quality Evaluation Based On Queuing Model

Queuing theory, also known as stochastic service system theory, is a theory of queuing problems. The analysis of the queuing problem is mainly based on the queuing system, and the online application in the cloud environment is also a random service system. Therefore, the queuing theory can be used to analyze the dynamic behavior characteristics of cloud services [14].

The queuing system will always be in a steady state after running for a certain period of time. In the steady state, the characteristics of each indicator are independent of the time of the system. Therefore, this paper mainly studies the indicators of the queuing system when it reaches steady state the relationship between. The mapping between the parameters of the online application and the parameters of the queuing system: the load is mapped to the average capacity of the customer; the number of concurrent service requests is mapped to the average queue length; the response time is mapped to the average stay; the throughput is mapped to absolute throughput; error rate is mapped to loss rate.

A. Evaluation of Single Node Service Quality Based on M/M/1 Model

Let *i* is a group of services, let *j* is a virtual node, the handling capacity μ_{ij} and storage capacity c_{ij} can be calculated from the cloud service resource consumption model. According to real-time monitoring can get the service request load is λ_{ij}, suppose service strength is $\rho = \frac{\lambda_{ij}}{\mu_{ij}}$. Based on the M/M/1/$c_{ij}$ model, the average stay time, loss rate and absolute passing capacity of the queuing system can be obtained when the system is in steady state. Calculate the response time *rt* according to Eq. (5)

$$\mathrm{rt}(\lambda_{ij}) = \frac{1}{\mu_{ij}(1-\rho)} - \frac{c_{ij}\rho^{c_{ij}}}{\mu_{ij}(1-\rho^{c_{ij}})} \tag{5}$$

Calculate the error rate *err* according to Eq. (6)

$$\mathrm{err}(\lambda_{ij}) = \rho^{c_{ij}}\frac{1-\rho}{1-\rho^{c_{ij}}} \tag{6}$$

Calculate the throughput *th* according to Eq. (7)

$$\mathrm{th}(\lambda_{ij}) = \lambda_{ij}\cdot(1-\rho^{c_{ij}}\frac{1-\rho}{1-\rho^{c_{ij}}}) \tag{7}$$

B. Cluster Quality of Service Measurement Based on M/M/s Model

When the service capability of a single node can not meet the needs of users, it uses the technology of cluster and organizes multiple similar nodes into a service cluster. It can be regarded as a queuing model with multiple service stations in service.

Let *i* is a group of services, let *s* is the number of nodes in the cluster, let *j* is a node in the cluster. The handling capacity μ_{ij} and storage capacity c_{ij} can be calculated from the cloud service resource consumption model. According to real-time monitoring can get the service request load is λ_{ij}, suppose service strength is $\rho_s = \frac{\rho}{s} = \frac{\lambda_{ij}}{s\cdot\mu_{ij}}$. Based on the M/M/s/$c_{ij}$ model, the average stay time, loss rate and absolute passing capacity of the queuing system can be obtained when the system is in steady state. Calculate the response time *rt* according to Eq. (8)

$$\mathrm{rt}(\lambda_{ij}, s) = \frac{L_q+(1-P_0)}{\mu_{ij}(1-P_0)} \tag{8}$$

Calculate the error rate *err* according to Eq. (9)

$$\mathrm{err}(\lambda_{ij}, s) = \frac{S^s\rho^{c_{ij}}}{S!}\cdot P_0 \tag{9}$$

Calculate the throughput *th* according to Eq. (10)

$$\mathrm{th}(\lambda_{ij}, s) = \lambda_{ij}\cdot(1-\frac{S^s\rho^{c_{ij}}}{S!}\cdot P_0) \tag{10}$$

Where $P_0 = [\sum_{k=0}^{s}\frac{s^k}{k!} + \frac{s^s\rho(\rho^s-\rho^{c_{ij}})}{s!(1-\rho)}]^{-1}$, and $L_q = \frac{\rho s^s}{s!(1-\rho)^2}\{1-\rho^{c_{ij}-s}[1+(c_{ij}-s)(1-\rho)]\}P_0$.

VI. Experiment

In this paper, the load test method is used to verify the feasibility and effectiveness. A service or multi service test is carried out under different load pressures, and the response time, error rate and throughput of the system under different load levels are collected in real-time by monitoring tools. Finally, compare the real values and calculated values by the establishment of cloud service quality evaluation model in order to verify the feasibility and accuracy of the method proposed in this paper.

A. Experimental Design

We chose the Amazon EC2 virtual machine as a public cloud test platform. Test the single node and the cluster which consists of three nodes, separately. Take the self-developed Java Web service as the test object. The single EC2 parameter used by the test platform is:

Model: t2.small

CPU: 1 core

Memory: 2G

System: Linux

We selected three different kinds of cloud service as a test object. Cloud service A is browsing, static page loading after rendering; cloud service B is query, responsible for retrieval of MySQL database in accordance with the specified key; cloud service C is increase, responsible for inserting a data into the MySQL database. In order to better simulate user requests and evaluate the quality of cloud services, we designed and implemented a cloud service quality evaluate tool framework in Figure 2.

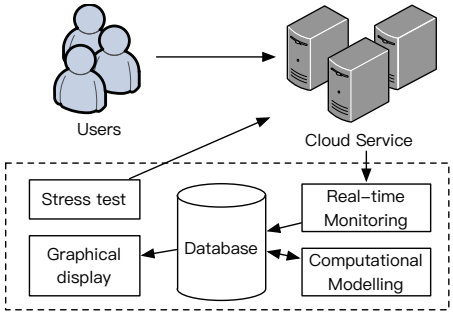


Figure 2. Cloud service quality evaluate tool framework.

Experimental steps:

a) Interface pressure test: for the service interface to be tested using the pressure test module, the interface to varying degrees of testing, respectively, record the corresponding response time, error rate, throughput and other parameters.

b) Real-time monitoring: monitoring the cloud service interface, through real-time monitoring module to collect cloud service runtime indicators such as response time, error rate, throughput and other parameters, and deposited in the database.

c) Computational modeling: query the operating parameters of a cloud service from the database, using the cloud service quality evaluate method proposed in this paper to carry out modeling and evaluation, and the results stored in the database.

d) Graphical display: in order to more intuitive display of the accuracy of the algorithm, graphical display module from the database to queries modeling data, through the evaluation of the value and the true value of the comparison, the use of graphics to show the difference.

B. Experimental Results and Analysis

Using the above test tools, the self-developed java web service is tested, the single service and multi-service through the stress test method, in a single machine and cluster environment for cloud service quality assessment, and the evaluation value and the true value in a graphical way.

The parameters of the cloud service resource evaluation model are: processing capacity coefficient $\tau = 3.86$; available memory is $D = 128$M; capacity coefficient $\theta = 1.44$; each kind of cloud service performance parameters of the service are shown in Table 1.

TABLE I. PERFORMANCE OF CLOUD SERVICES

	Type	***Response***	***Memory***
Service A	Browse	21(ms)	12.94(k)
Service B	Query	80(ms)	189.71(k)
Service C	Insert	97(ms)	47.24(k)

Select cloud service B as a single service, to test and verify in a single node and the cluster which consists of three nodes. Compare throughput, response time and error rate which calculated by proposed method and actual measurement, to verify the availability of the proposed method for single service. The experimental results are shown in Figure 3.

Select cloud service A, cloud service B and cloud service C combinations for multi-service, to test and verify in a single node and the cluster which consists of three nodes. Compare throughput, response time and error rate which calculated by proposed method and actual measurement, to verify the availability of the proposed method for multi service. The experimental results are shown in Figure 4.

As shown in the Figure 3 and Figure 4, experimental result proves the ability of the method which this paper proposed cloud service quality evaluation based on queuing model by comparing the calculated and the actual values. In the single cloud service, each virtual node only to provide a service capability, and test the ability to provide computing power; in the multi cloud service, each virtual node at the same time to provide three different kinds of service capabilities, the same time each service have the same amount of concurrency, and test the general ability to provide computing power.

As can be got from Figure 3, for the single cloud service in the single node or cluster, the value of the evaluation is well fitted with the actual value, and the result is more accurate; as can be got from Figure 4, for multi cloud service in the single node or cluster, the evaluation value can reflect the changing trend of the real value .Therefore, this shows that the proposed method can better evaluate the quality of cloud services in single-service or multi-service on single nodes or clusters, reflect the trend of throughput, error rate and response time, and give the accurately predictions.

VII. CONCLUSION

In this paper, we have proposed a method of cloud service quality evaluation based on queueing model for the online SOA architecture distributed system. The goal is to evaluate the cloud service quality reasonably, through the modeling of resource consumption of cloud service, in order to maximize the reduction of basic costs and plan the use of cloud resources to ensure the stable and reliable operation of cloud services.

In this paper, the M/M/1 and M/M/s queuing models are used to analyze and evaluate the quality of single node service and cluster service respectively. It can not only reflect the trend of change accurately, but also predict throughput, error rate and response time under different concurrency in the single service or multi service. Based on the evaluation model proposed in this paper, we can not only guarantee the quality of cloud services, but also evaluate the cloud computing resources consumed by cloud services and realize the goal of green cloud computing.

ACKNOWLEDGMENT

This work is supported by the Key R&D Plan Project of China (No. 2014BA07B02), the Natural Science Foundation of Shandong Province (ZR2015FM006), the Major Science & Technology Specific Project of Shandong Province (No. 2015ZDXX0201B02) and the Key R&D plan Project of Shandong Province (GG201703130116).

REFERENCES

[1] Ma Zhiyi, Chen Hongjie. A Service-Oriented Ar-chitecture Reference Model[J]. Chinese Journal of Computers, 2006, 29(7): 1011-1019.

[2] Singh S, Chana I. QoS-aware autonomic resource management in cloud computing: a systematic review[J]. ACM Computing Surveys (CSUR), 2016, 48(3): 42.

[3] Nassiffe R, Camponogara E, Lima G, et al. Optimising QoS in adaptive real-time systems with energy constraint varying CPU frequency[J]. International Journal of Embedded Systems, 2016, 8(5-6): 368-379.

[4] Calheiros R N, Ranjan R, Beloglazov A, et al. CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms[J]. Software: Practice and experience, 2011, 41(1): 23-50.

[5] Armbrust M, Fox A, Griffith R, et al. A view of cloud computing[J]. Communications of the ACM, 2010, 53(4): 50-58.

[6] Islam S, Keung J, Lee K, et al. Empirical prediction models for adaptive resource provisioning in the cloud[J]. Future Generation Computer Systems, 2012, 28(1): 155-162.

[7] Han R, Ghanem M M, Guo L, et al. Enabling cost-aware and adaptive elasticity of multi-tier cloud applications[J]. Future Generation Computer Systems, 2014, 32: 82-98.

[8] Liu Sai，Li Xurong，Wan Linrui. Computing Clouds Resources Pool Model Research Based on Queue Theory[J]. Computer Technology and Development, 2012, 22(12): 87-89.

[9] Vilaplana J, Solsona F, Teixidó I, et al. A queuing theory model for cloud computing[J]. The Journal of Supercomputing, 2014, 69(1): 492-507.

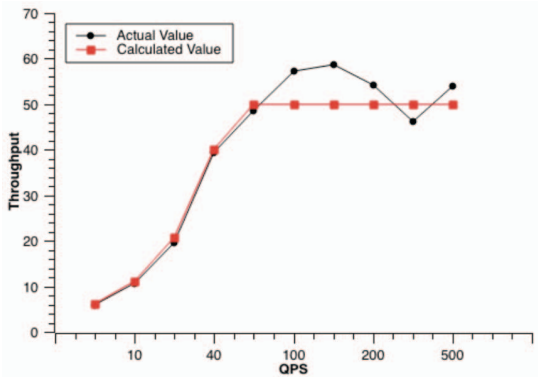
[10] Son S, Choi H H, Oh B T, et al. Cloud SLA relationships in multi-cloud environment: models and practices[C]//Proceedings of the 8th International Conference on Computer Modeling and Simulation. ACM, 2017: 1-6.

[11] Menasce D A. Computing Missing Service Demand Parameters for Performance Models[C]//Int. CMG Conference. 2008: 241-248.

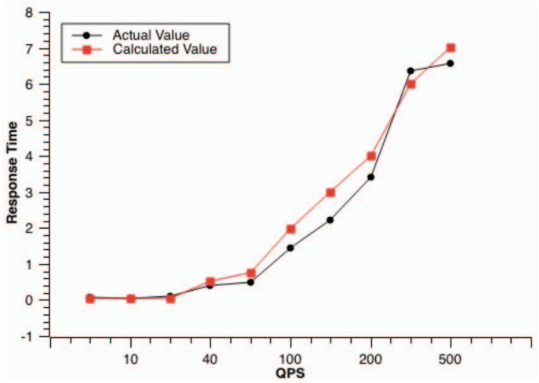
[12] Ye Shiyang, Zhang Wenbo, Zhong Hua. Sla-Oriented Virtual Resources Scheduling In Cloud Computing Environment[J]. Computer Applications and Soft-ware, 2015, 32(4): 11-17.

[13] Zhao Hongwei, Shen Derong, Tian Liwei. Research on Resources Forecasting and Scheduling Method in Cloud Computing Environment[J]. Journal of Chinese Computer Systems, 2016, 37(4): 659-663.

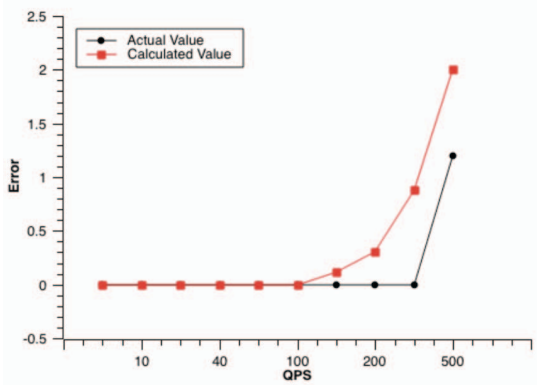
[14] Tan Yiming, Zeng GuoSun, Wang Wei. Policy of Energy Optimal Management for Cloud Computing Platform with Stochastic Tasks[J]. Journal of Software,2012,(02):266-278.



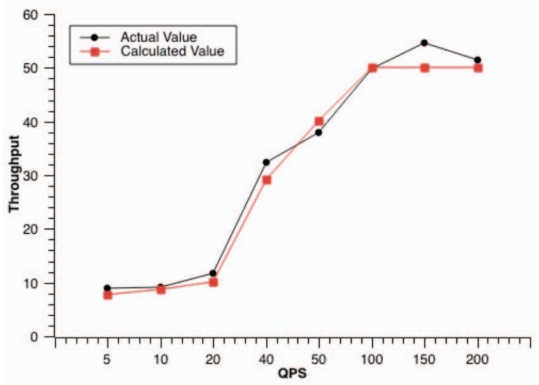
(a) Throughput in the single node



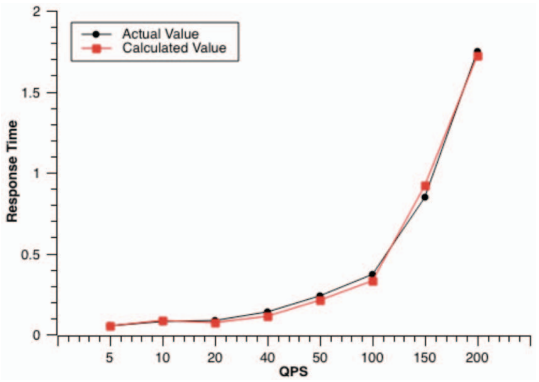
(b) Response time in the single node



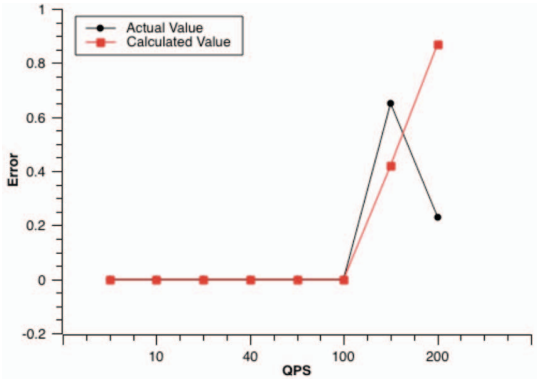
(c) Error rate in the single node



(d) Throughput in the cluster

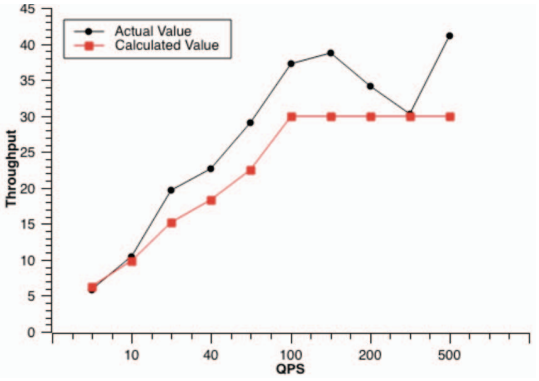


(e) Response time in the cluster

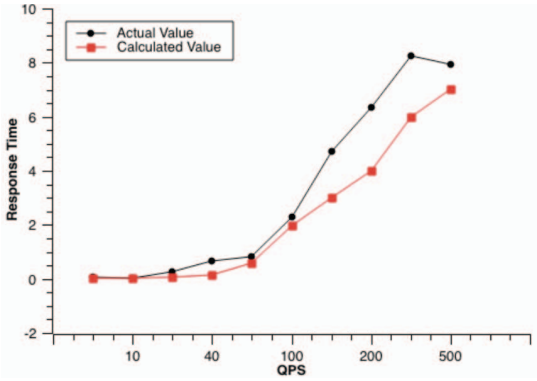


(f) Error rate in the cluster

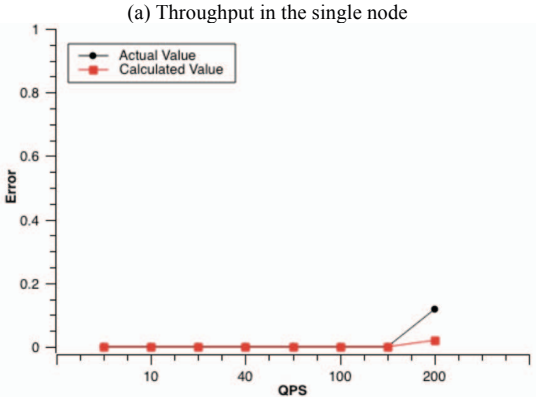
Figure 3. Experimental results of single cloud service in single node and cluster.



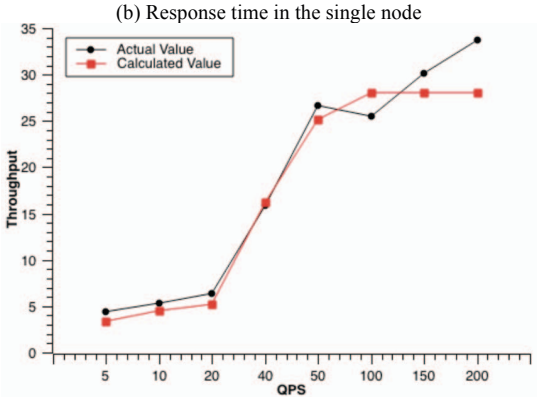
(a) Throughput in the single node



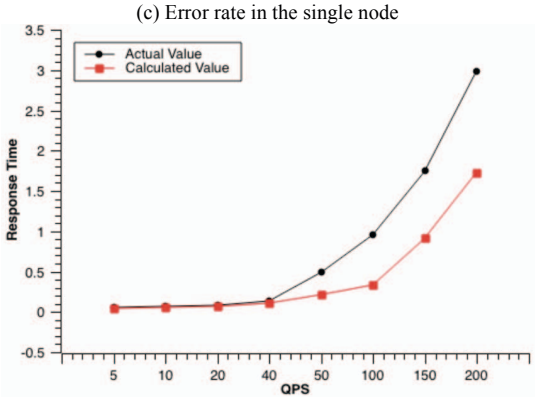
(b) Response time in the single node



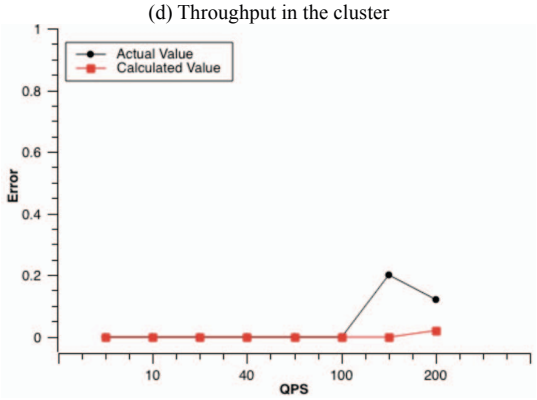
(c) Error rate in the single node



(d) Throughput in the cluster



(e) Response time in the cluster



(f) Error rate in the cluster

Figure 4. Experimental results of multi cloud service in single node and cluster.