ISSN: 2321-7782 (Online)
Impact Factor: 6.047

Volume 4, Issue 10, October 2016

International Journal of Advance Research in Computer Science and Management Studies

Research Article / Survey Paper / Case Study

Available online at: www.ijarcsms.com

Ensuring an Efficient and Reliable Quality of Service in Cloud Computing

Anusha Pothuri[1]
M.Tech Scholar
St.Marys group Of Institutions Guntur
Chebrole(V&M),Guntur(dt)
Andhra Pradesh – India

Subhani Shaik[2]
Associate Professor
St.Marys group Of Institutions Guntur
Chebrole(V&M),Guntur(dt)
Andhra Pradesh – India

***Abstract:** As an effective and efficient way to provide computing resources and services to customers on demand, cloud computing has become more and more popular. From cloud service providers' perspective, profit is one of the most important considerations, and it is mainly determined by the configuration of a cloud service platform under given market demand. However, a single long-term renting scheme is usually adopted to configure a cloud platform, which cannot guarantee the service quality but leads to serious resource waste. In this paper, a double resource renting scheme is designed firstly in which short-term renting and long-term renting are combined aiming at the existing issues. This double renting scheme can effectively guarantee the quality of service of all requests and reduce the resource waste greatly. Secondly, a service system is considered as an M/M/m+D queuing model and the performance indicators that affect the profit of our double renting scheme are analyzed, e.g., the average charge, the ratio of requests that need temporary servers, and so forth. Thirdly, a profit maximization problem is formulated for the double renting scheme and the optimized configuration of a cloud platform is obtained by solving the profit maximization problem. Finally, a series of calculations are conducted to compare the profit of our proposed scheme with that of the single renting scheme. The results show that our scheme can not only guarantee the service quality of all requests, but also obtain more profit than the latter.*

***Keywords:** Cloud computing, guaranteed service quality, multi server system, profit maximization, queuing model, service-level agreement, waiting time.*

I. Introduction

As an effective and efficient way to consolidate computing resources and computing services, clouding computing has become more and more popular [1]. Cloud computing centralizes management of resources and services, and delivers hosted services over the Internet. The hardware, software, databases, information, and all resources are concentrated and provided to consumers on-demand [2]. Cloud computing turns information technology into ordinary commodities and utilities by the the pay-per-use pricing model [3, 4, 5]. In a cloud computing environment, there are always three tiers, i.e., infrastructure providers, services providers, and customers (see Fig. 1 and its elaboration in Section 3.1). An infrastructure provider maintains the basic hardware and software facilities. A service provider rents resources from the infrastructure providers and provides services to customers. A customer submits its request to a service provider and pays for it based on the amount and the quality of the provided service [6]. In this paper, we aim at researching the multiserver configuration of a service provider such that its profit is maximized. To configure a cloud service platform, a service provider usually adopts a single renting scheme. That's to say, the servers in the service system are all long-term rented. Because of the limited number of servers, some of the incoming service requests cannot be processed immediately. So they are first inserted into a queue until they can handled by any available server. However, the waiting time of the service requests cannot be too long. In order to satisfy quality-of-service requirements, the waiting time of each incoming service request should be limited within a certain range, which is determined by a service-level agreement (SLA). If the quality of service is guaranteed, the service is fully charged, otherwise, the service provider serves the request for free as a penalty of low quality. To obtain higher revenue, a service provider should rent more servers from the infrastructure providers or scale up the server execution speed to ensure that more service requests are processed with high service quality. Even though the cloud has greatly simplified the capacity provisioning process, it poses several novel challenges in the area of Quality-of-Service (QoS) management. QoS denotes the levels of performance, reliability, and availability offered by an application and by the platform or infrastructure that hosts it[a]. QoS is fundamental for cloud users, who expect providers to deliver the advertised quality characteristics, and for cloud providers, who need to find the right tradeoffs between QoS levels and operational costs. However, finding optimal tradeoff is a difficult decision problem, often exacerbated by the presence of service level agreements (SLAs) specifying QoS targets and economical penalties associated to SLA violations [3].While QoS properties have received constant attention well before the advent of cloud computing, performance heterogeneity and resource isolation mechanisms of cloud platforms have significantly complicated QoS analysis, prediction, and assurance. This is prompting several researchers to investigate automated QoS management methods that can leverage the high programmability of hardware and software resources in the cloud [4]. This paper aims at supporting these efforts by providing a survey of the state of the art of QoS modeling approaches applicable to cloud computing and by describing their initial application to cloud resource management. Quality of Service (QoS) is a broad topic in Distributed Systems and is most often referred to as the resource reservation control mechanisms in place to guarantee a certain level of performance and availability of a service. The scope of this paper is primarily concerned with the management and performance of resources such as processors memory, storage and networks in Cloud Computing. A defined QoS is not just limited to guarantees of performance and availability and can cover other aspects of service quality, which are outside the scope of this pap er, such as security and dependability. The problems surrounding resource reservation are non-trivial for all but the most basic best effort guarantees and the problems behind resource capacity planning are non-deterministic polynomial-time hard to solve. In this paper, we propose a novel renting scheme for service providers, which not only can satisfy quality-of-service requirements, but also can obtain more profit.

Our contributions in this paper can be summarized as follows.

- A novel double renting scheme is proposed for service providers. It combines long-term renting with short-term renting, which can not only satisfy quality-of-service requirements under the varying system workload, but also reduce the resource waste greatly.
- A multi server system adopted in our paper is modeled as an *M/M/m+D* queuing model and the performance indicators are analyzed such as the average service charge, the ratio of requests that need short term servers, and so forth.
- The optimal configuration problem of service providers for profit maximization is formulated and two kinds of optimal solutions, i.e., the ideal solutions and the actual solutions, are obtained respectively.
- A series of comparisons are given to verify the performance of our scheme. The results show that the proposed Double-Quality-Guaranteed (DQG) renting scheme can achieve more profit than the compared

Single-Quality-Unguaranteed (SQU) renting scheme in the premise of guaranteeing the service quality completely.

The rest of the paper is organized as follows. Section 2 reviews the related work on profit aware problem in cloud computing. Section 3 presents the used models, including the three-tier cloud computing model, the multi server system model and the revenue and cost models. Section 4 proposes our DQG renting scheme and formulates the profit optimization problem. Section 5 introduces the methods of finding the optimal solutions for the profit optimization problem in two scenarios. Section 6 demonstrates the performance of the proposed scheme through comparison with the traditional SQU renting scheme. Finally, Section 7 concludes the work.

II. Related Work

In this section, we review recent works relevant to the profit of cloud service providers. Profit of service providers is related with many factors such as the price, the market demand, the system configuration, the customer satisfaction and so forth. Service providers naturally wish to set a higher price to get a higher profit margin; but doing so would decrease the customer satisfaction, which leads to a risk of discouraging demand in the future. Hence, selecting a reasonable pricing strategy is important for service providers. To configure a cloud service platform, a service provider usually adopts a single renting scheme. That's to say, the servers in the service system are all long-term rented. Because of the limited number of servers, some of the incoming service requests cannot be processed immediately. So they are first inserted into a queue until they can handled by any available server. However, the waiting time of the service requests cannot be too long. In order to satisfy quality-of-service requirements, the waiting time of each incoming service request should be limited within a certain range, which is determined by a service-level agreement (SLA). If the quality of service is guaranteed, the service is fully charged, otherwise, the service provider serves the request for free as a penalty of low quality. To obtain higher revenue, a service provider should rent more servers from the infrastructure providers or scale up the server execution speed to ensure that more service requests are processed with high service quality. Thus the majority of the lessons already learnt within the research topic are highly relevant to Cloud computing. The motivation behind research into Cloud Computing was initially the need to manage large scale resource intensive scientific applications across multiple administrative domains that require many more resources than that can b e provided by a single computer. Cloud computing shares this motivation but within a new context oriented towards business rather than academic resource management, for the stipulation of reliable services rather than batch oriented scientific applications. This difference in application domain and requirements being pushed by industry does not mean that the scientific community cannot leverage Cloud Computing, far from it, as illustrated by Cloud Batch[2]. There is much crossover between the two paradigms and many goals are shared. Cloud Computing will b e enabled through the next generation of data centre technology. The current generation of data centers are already leaning heavily towards the virtualization of compute and storage resources, the technological foundation of a Cloud, enabling the consolidation of proprietary servers running legacy software. This is being achieved through the creation of virtual machines which run on large physical servers utilizing the latest technology. This provides the benefits of being able to both reduce maintenance cost and minimize lost revenue due to downtime and also takes advantage of the improvements in computer efficiency facilitated by hardware vendors such as Intel and AMD. Monitoring tools are essential in ascertaining the availability of resources and providing feedback to schedulers within Cloud. Monitoring tools enable guarantees to be made on the performance of any given resource by making sure that the computational resource in question is not over utilized and is on- line. Performance is characterized by the amount of useful work accomplished by a computer system in comparison to the time and resources used. Monitoring tools are also essential in providing fault tolerance and the migration of tasks in the event of a resource failure in the Cloud. Fault tolerance involves the identification of a resource failure via monitoring tools, the rescheduling of the task to an alternative available resource and migration of the state of the task to the newly allotted resource, at which point the task continues execution.

III. Models in Cloud Computing

In this section, we first describe the three-tier cloud computing structure. Then, we introduce the related models used in this paper, including a multi server system model, a revenue model, and a cost model.

3.1 A Cloud System Model

The cloud structure (see Fig. 1) consists of three typical parties, i.e., infrastructure providers, service providers and customers. This three-tier structure is used commonly in existing literatures [2, 6, 10].

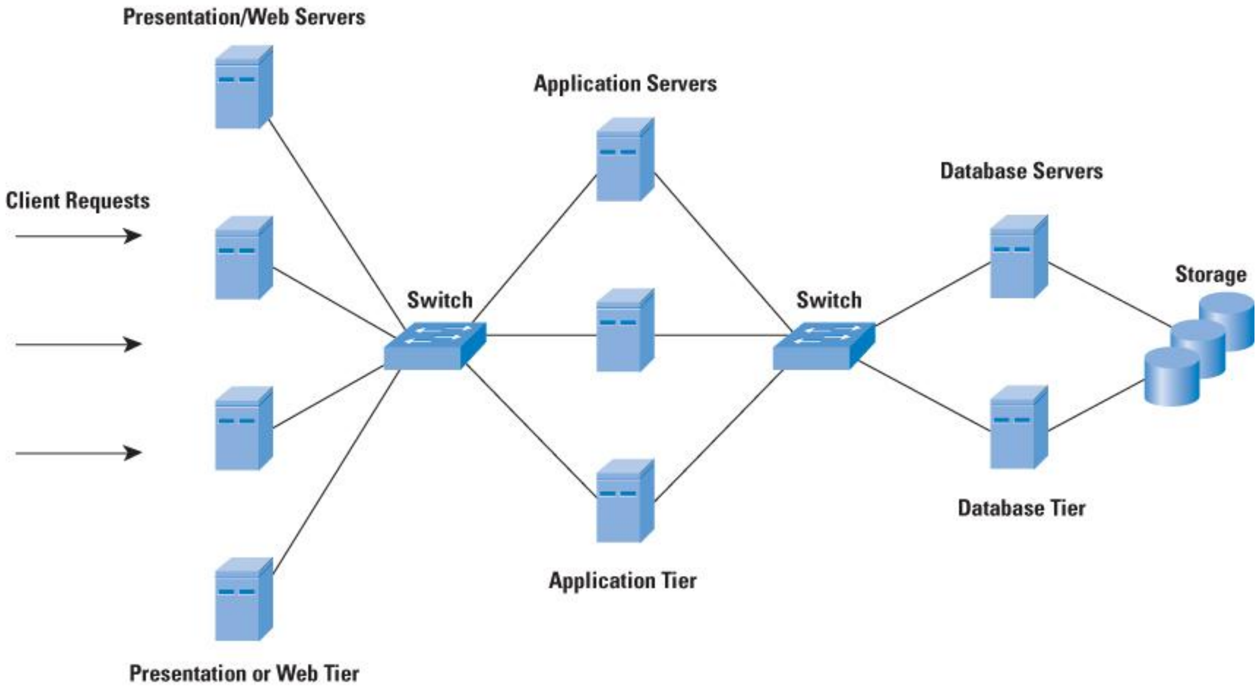
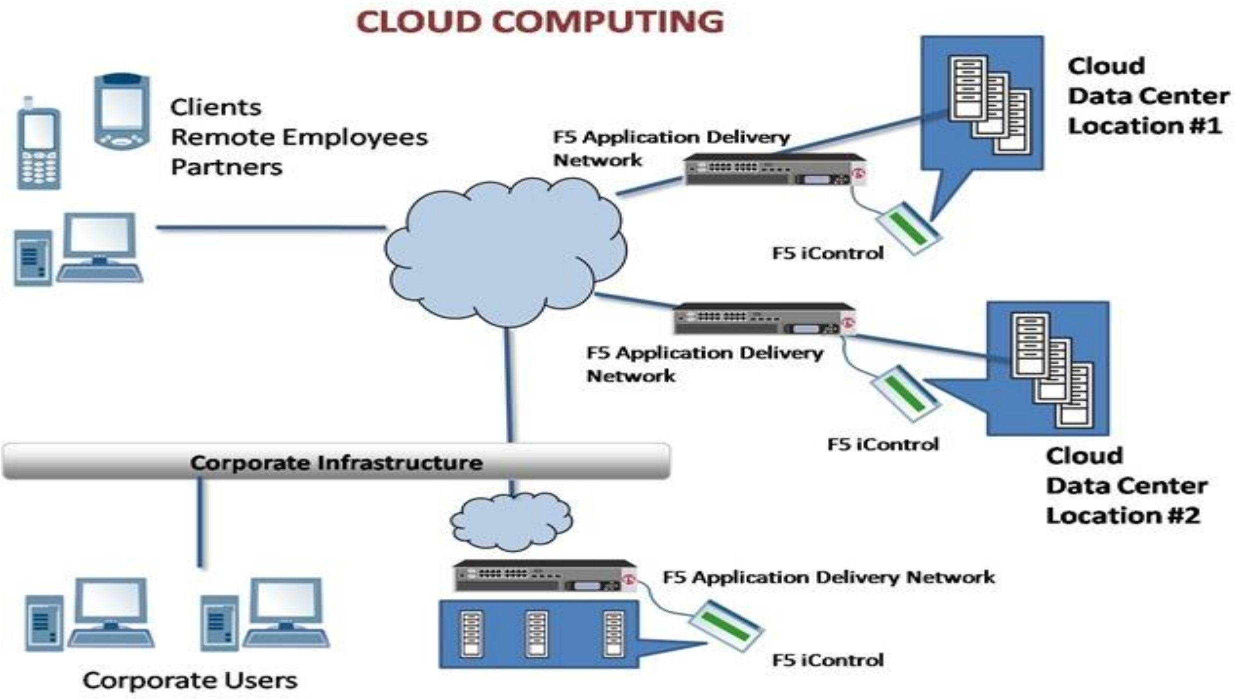


Fig 3.1: The Three-Tier Cloud Structure.

The presentation, business logic, and data handling is realized as separate tiers to scale stateless presentation and compute-intensive processing independently of the data tier, which is harder to scale and often handled by the cloud provider. A Distributed Application is decomposed into application components to scale individual application functions independently. There can be many differentiating factors of application tiers. For example, if Processing Components are more computation intensive or are used less frequently than User Interface Components, aligning the elastic scaling of these two components by summarizing their implementation in one tier can be inefficient. This issue arises every time components experiences different Application Workloads. The number of provisioned component instances cannot be aligned well to the different workloads if they are summarized to coarse grained tiers. Applications component instances and data handled by them are globally distributed to meet the access performance required by a global user group.

3.2 A Multi Server Model


CLOUD COMPUTING
Clients
Remote Employees
Partners
F5 Application Delivery Network
Cloud Data Center Location #1
F5 iControl
F5 Application Delivery Network
F5 iControl
Cloud Data Center Location #2
Corporate Infrastructure
F5 Application Delivery Network
F5 iControl
Corporate Users

They have proposed a pricing model for cloud computing which takes many factors into considerations, such as the requirement of a check, the workload of an application environment, the configuration (m and s) of a multi server system, the service stage agreement c, the satisfaction (r and 0) of a consumer, the quality (W and T) of a service, the price of a low quality service, the cost and mof renting, the cost (P and P) of energy utilization, and a service provider's margin and profit. The cloud caching service can maximize its profit using an optimal pricing scheme. Optimal pricing necessitate an appropriately simplified price demand model that incorporates the correlations of structures in the cache services. Provides a multi cloud service for an research application that achieves optimal pricing for the products available in different cloud services (like Amazon, Azure, eBay, etc)in a clustered environment. This work propose a novel pricing scheme designed for a cloud cluster that offers inter querying services and aims at the maximization of the cloud profit. An appropriate price demand and formulate the optimal pricing problem.

IV. A Quality Guaranteed Scheme

The traditional single resource renting scheme cannot guarantee the quality of all requests but wastes a great amount of resources due to the uncertainty of system workload. To overcome the weakness, we propose a double renting scheme as follows, which not only can guarantee the quality of service completely but also can reduce the resource waste greatly.

The Proposed Scheme:

In this section, we first propose the Double-Quality- Guaranteed (DQG) resource renting scheme which combines long-term renting with short-term renting. The main computing capacity is provided by the long-term rented servers due to their low price. The short-term rented servers provide the extra capacity in peak period. Since the requests with waiting time *D* are all assigned to temporary servers, it is apparent that all service requests can guarantee their deadline and are charged based on the workload according to the SLA. Hence, the revenue of the service provider increases. However, the cost increases as well due to the temporarily rented servers. Moreover, the amount of cost spent in renting temporary servers is determined by the computing capacity of the long-term rented multi server system. Since the revenue has been maximized using our scheme, minimizing the cost is the key issue for profit maximization. Next, the tradeoff between the long term rental cost and the short-term rental cost is considered, and an optimal problem is formulated in the following to get the optimal long-term configuration such that the profit is maximized.

V. Conclusion

In order to guarantee the quality of service requests and maximize the profit of service providers, this paper has proposed a novel Double-Quality-Guaranteed (DQG) renting scheme for service providers. This scheme combines short-term renting with long-term renting, which can reduce the resource waste greatly and adapt to the dynamical demand of computing capacity. An *M/M/m+D* queuing model is build for our multi server system with varying system size. And then, an optimal configuration problem of profit maximization is formulated in which many factors are taken into considerations, such as the market demand, the workload of requests, the server-level agreement, the rental cost of servers, the cost of energy consumption, and so forth. The optimal solutions are solved for two different situations, which are the ideal optimal solutions and the actual optimal solutions. In addition, a series of calculations are conducted to compare the profit obtained by the DQG renting scheme with the Single-Quality-Unguaranteed (SQU) renting scheme. The results show that our scheme outperforms the SQU scheme in terms of both of service quality and profit. They have proposed a pricing model for cloud computing which takes many factors into consideration, such as the requirement of a check, the workload _ of an application Environment, the configuration (m and s) of a multi server system, the service level concurrence c, the satisfaction (r ands0) of a consumer, the quality (W and T) of a service, the price d of a low quality service, the cost (_ and m) of renting, the cost (_, , P_, and P) of energy consumption, and a cloud service provider's margin and earnings a. By using an M/M/m queuing model, the formulated and solved the problem of optimal multi server configuration for profit maximization in a cloud computing environment.

References

1. K. Hwang, J. Dongarra, and G. C. Fox, Distributed and Cloud Computing. Elsevier/Morgan Kaufmann, 2012.
2. J. Cao, K. Hwang, K. Li, and A. Y. Zomaya, “Optimal multiserver configuration for profit maximization in cloud computing,” IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6, pp. 1087–1096, 2013.
3. A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, and I. Stoica, “Above the clouds: A berkeley view of cloud computing,” Dept. Electrical Eng. and Comput. Sciences, vol. 28, 2009.
4. R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, “Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility,” Future Gener. Comp. Sy., vol. 25, no. 6, pp. 599– 616, 2009.
5. P. Mell and T. Grance, “The NIST definition of cloud computing. national institute of standards and technology,” Information Technology Laboratory, vol. 15, p. 2009, 2009.
6. J. Chen, C. Wang, B. B. Zhou, L. Sun, Y. C. Lee, and A. Y. Zomaya, “Tradeoffs between profit and customer satisfaction for service provisioning in the cloud,” in Proc. 20th Int’l Symp. High Performance Distributed Computing. ACM, 2011, pp. 229–238.
7. B.N. Chun and D.E. Culler, “User-Centric Performance Analysis of Market Based Cluster Batch Schedulers,” Proc. Second IEEE/ ACM Int’l Symp. Cluster Computing and the Grid, 2002.
8. D. Durkee, “Why Cloud Computing Will Never be Free,” Comm. ACM, vol. 53, no. 5, pp.62 -69, 2010.
9. R. Ghosh, K.S. Trivedi, V.K. Naik, and D.S. Kim, “End to End Perform ability Analysis for Infrastructure-as-a-Service Cloud: An Interacting Stochastic Models Approach,” Proc. 16th IEEE Pacific Rim Int’l Symp.Dependable Computing, pp. 125-132, 2010.
10. K. Hwang, G.C. Fox, and J.J. Dongarra, Distributed and Cloud Computing. Morgan Kaufmann, 2012.
11. “Enhanced Intel Speed Step Technology for the Intel Pentium M Processor, “White Paper, Intel, Mar. 2004.
12. L. Klein rock, Queuing Systems: Theory, vol. 1. John Wiley and Sons, 1975.

© 2016, IJARCSMS All Rights Reserved