A Profit Maximization Scheme with Guaranteed Quality of Service in Cloud Computing

Kanchan Naithani,
Asst. Professor, Department of Comp. Sc. & Info. Tech., Graphic Era Hill University, Dehradun, Uttarakhand India 248002,

Abstract

Cloud computing has grown in popularity being a practical and efficient method to offer clients computing resources and services on demand. One of the most important elements from the standpoint of cloud service providers is revenue, which is mostly influenced by how a cloud service platform is built in light of market demand. To design a cloud platform, nevertheless, a single long-term rental plan is often used. This cannot ensure the quality of the service and results in significant resource waste. In order to address the difficulties at hand, this study first designs a twofold resource rental model that combines short-term and long-term rentals. This system of duplicate renting may significantly cut down on resource waste and efficiently ensure the caliber of service for every request.In second step, a service system is modeled as an M/M/m+D queuing model, and performance metrics that influence profitability of our twofold renting scheme are examined, such as average fee, the percentage of requests that require temporary servers, and other factors. Third, a profit maximization challenge for the dual renting scheme is created, and the optimal design of a cloud platform is produced by resolving issue. In order to evaluate profit of our suggested plan with that of only one rental scheme, a number of computations are finally performed. The findings demonstrate that our plan not only ensures that every request will receive high-quality service, but also generates more profit than the alternative.

DOINumber: 10.48047/nq.2021.19.7.NQ21121 **NeuroQuantology2021;19(7):345-350**

1. INTRODUCTION

Consolidating computer resources and services through cloud computing is a practical and successful method. There are three layers in a cloud computing environment: services providers, infrastructure providers, and consumer. When establishing a cloud service platform, a service provider frequently chooses a single renting scheme in which the servers are all rented for long time. The optimal service provider configuration problem for profit maximization is covered in this article. The waiting period for every incoming service request must be confined within a specified limit, as established by a Service-Level Agreement (SLA), with the goal to meet criteria for service quality. For service providers, a brand-new double-renting plan that mixes long-term and short-term renting is put out. An M/M/m+D queuing model is used to describe a multi-server system, and performance indicators such as average service fee, proportion of requests that require temporary servers, and other metrics are examined. In assumption that the service quality is totally assured, the suggested Double-Quality-Guaranteed (DQG) rental plan can generate greater profit than the Single-Quality-Unguaranteed (SQU) renting scheme. This paper presents the models that were used, suggests the DQG renting scheme, introduces profit optimization problem, illustrates effectiveness of proposed scheme by contrasting it with the conventional SQU renting scheme, and reviews recent works pertinent to profitability of cloud service providers. Dynamic pricing is a compelling alternative to the prevalent method of static pricing for managing erratic consumer demand.Customer happiness, which is influenced by both the service quality and the price, is the second

element that has an impact on service providers' profits. A SLA is made among the service provider and the clients with the goal to increase customer satisfaction. This study examines how to set up a cloud service platform to ensure the quality of all requests and minimize wastage of resourcesby using double renting system. While the SLA is a discussion among service providers and consumers on the service quality and price, the usage-based pricing model is used.

2. LITERATURE SURVEY

By utilizing technologies like Virtual Machines (VMs) the framework for building Clouds with market-focused allocation of resource is presented in this article. In order to support resource allocation that is focused on Service Level Agreements (SLA), it also offers insights on market-based resource management solutions that include both computational risk management and customer-driven service management. It also emphasizes the distinction between workload generated by Internet-based services and workload generated by High Performance Computing (HPC). In order to realize the 21st century vision, it is necessary for opposing IT paradigms to converge. Computing is evolving towards a paradigm that consists of commodities that are distributed similarly to conventional amenities like gas, water, telecommunications, andelectricity.This utility computing vision has been offered by a number of computing paradigms. An extension of this paradigm is cloud computing, in which commercial programs are made available as advanced services that can be utilized through a network. Consumers are drawn to cloud services by the possibility of lowering or completely removing expenses related to "in-house" provision of these services, while cloud service providers are motivated by the potential revenues that may be earned by billing users to use similar services. In order to offer redundancy and guarantee dependability in the event of site failures, providers including Google, Microsoft, Amazon, IBM, Sun Microsystems and Salesforcehave started to create additional data centers for hosting Cloud computing applications in numerous places across globe. Applications may be isolated from the hardware that underlies them and other VMs using virtual machines, and the platform can be tailored to the user's needs.However, the usage of VMs creates new difficulties, such as balancing the conflicting resource needs of the users through the intelligent distribution of physical resources. Enterprise service customers with worldwide operations need quicker response times, therefore they may cut down on time by simultaneously sending workload requests to many Clouds in different places. The development of such clouds and cloud connections is fraught with difficulties [1].

In order to fulfill its mandates under the Federal Information Security Management Act, the National Institute of Standards and Technology (NIST) created this paper. It complies with the guidelines in Section 8b(3), "Securing Agency Information Systems," of Office of Management and Budget Circular A-130, as described in A-130, Appendix IV: Analysis of Key Sections. Federal agencies are allowed to use this policy, but nongovernmental groups are free to utilize it as well. It is meant to give a foundation for debate on anything from what cloud computing is to the best ways to use it, as well as a way to make general comparisons between cloud services and deployment approaches. The service and deployment models described create a straightforward taxonomy that does not aim to impose or restrict any specific deployment, service delivery, or business operating technique.System planners, program managers, technologists, and other people embracing cloud computing as users or producers of cloud services are the target audience for this publication. A concept called "cloud computing" makes it possible to access a pool of shared computer resources that may be reconfigured (like storage, servers, applications, networks, and services) from anywhere at any time. It is made up of four deployment types, three service models, and five key features. On-demand self-service is the capacity to provision computing resources unilaterally, broad network access is the capacity to access resources over the network, resource pooling is the capacity to pool resources to serve multiple consumers, rapid elasticity is the capacity to provision and release capabilities elastically, and measured service is the capacity to track, manage, and report resource usage[2].

This study introduces a novel dynamic reuse-based optimization technique for profit-driven service request scheduling that takes into consideration the unique SLA features of client requests and the recent demand on the system. The suggested method creates a dynamic virtual machine resource pool as needed, accomplishes efficient cloud service request scheduling in a fair length of time, drastically lowers operating costs for cloud service providers, and boosts their profitability. When compared to many baseline algorithms, simulation studies demonstrate that the suggested algorithm enhances the usage of virtual resources and boosts the revenues of cloud service providers. The cloud service provider receives a service request from the client, accepts it, and sends it to the vendors of the underlying cloud infrastructure for virtual resources on demand.In

response to the request for resource leasing, the cloud infrastructure vendor assigns VM instances to the appropriate cloud service provider for executing cloud user request. In order to obtain the best request scheduling, this study introduces a novel optimization approach called PSRSDR (Profit-driven Service Request Scheduling based on Dynamic Reuse).In order to reduce VM rental costs while still guaranteeing service performance satisfies SLAs, it utilizes both of the segmentability of user requests and the flexibility of SLA attributes. The suggested method beats conventional revenue-aware algorithms with regards to virtual resource consumption and operating profit, according to simulation trials. The following is how the paper is set up: The cloud service request model with SLA restrictions and cloud service provider's revenue function are presented in Section 2, our cloud service request scheduling strategy is described in Section 3, the outcomes of the simulation experiment are provided in Section 4, and Section 5 serves as the article's conclusion. [3].

The issue of energy reduction for repetitive preemptive hard real-time jobs planned on a similar multiprocessor architecture with dynamic voltage scaling capabilities is covered in this study. It suggests a three-part integrated strategy made up of the speed assignment method, the partitioning heuristic, and the RMS admission control test. Experimental results demonstrate that Worst-Fit outperforms other popular heuristics in off-line conditions and an approach that ensures constant performance by allocating a fraction of processors for light workloads. Partitioning heuristics like First-Fit and Best-Fit have acceptable average-case performance, despite the fact that partitioned multiprocessor real-time scheduling is NP-Hard. The highest priority jobs on mprocessors are chosen for execution by a global scheduler from a single ready queue in a different way than local scheduling.However, because of Rate Monotonic Analysis &constrained several priority levels in the majority of commercial operating systems, real-time and embedded systems continue to adopt static-priority rules. This results in lower CPU usage. Energy management on multiprocessor platforms has been studied in research articles, including dynamic speed adjustment and slack reclamation. While Baruah and Anderson use global EDF to handle the system synthesis issue of periodic RT workloads on similar multiprocessors, Aydin et al. suggest two shared slack reclamation strategies [4].

The difficulties of end-to-end quality of service management for common Internet applications and related price incentive schemes are covered in this note. The first study of service differentiation uses a twoclass model traffic that is sensitive to delays and throughput. Under overload situations, flat-rate pricing is preferable to usage-based pricing from the standpoints of both the system and the individual users. Typically, core networks only offer single-tier bulk transit. The issue of network neutrality has recently gained attention since core providers should be able to charge for packets in part based on where they came from or the sort of application they are used for.Because to "terms of use" limitations and unequal access bandwidth for subscriber traffic aggregates, the access network are currently not neutral. The Internet core in the US is made up of a collection of private networks whose owners are trying to protect the income streams from their managed phone and video services. This essay makes the case for a variety of service classes, such as access to service classes that are both differently priced and subject to SLA regulation. Use-priced premium classes-of-service (CoSs) or tier-based flat charges for premium CoSs might lower the amount of premium traffic in use. For some high-volume real-time data applications, usage-based charging is compatible with the temporary requirement for reserved end-to-end bandwidth, but it necessitates the installation of an authenticated billing system, which might be expensive.Other security advantages of detailed authentication and network monitoring systems include flow attribution and financial deterrent for attack and bothersome/spam behavior in premium serviceclasses. If linked revenues have the potential to be high and "scalable" accounting techniques are applied, mounting a billing system could be possible [5].

3. PROPOSED SYSTEM

Infrastructure designs, software delivery, and development methodologies have all been changed by cloud computing. The introduction of a Trusted Third Party, entrusted with guaranteeing particular security features inside a cloud environment, is suggested in this study. To guarantee the authenticity, integrity, and secrecy of relevant data and communications, the suggested solution uses cryptography, more especially Public Key Infrastructure, working in collaborationwith SSO and LDAP. This article describes a horizontal level of service, available to all involved parties, that builds a security mesh and upholds fundamental trust. This essay suggests a security solution that relies on clients trusting a Third Party to relieve them of the security responsibility.The Third Party is entrusted with generating a trust mesh between the relevant entities, creating cloud federations, while

guaranteeing specified security features inside a distributed information system. The research strategy used to accomplish this aim is based on information systems design and software engineering methodologies.

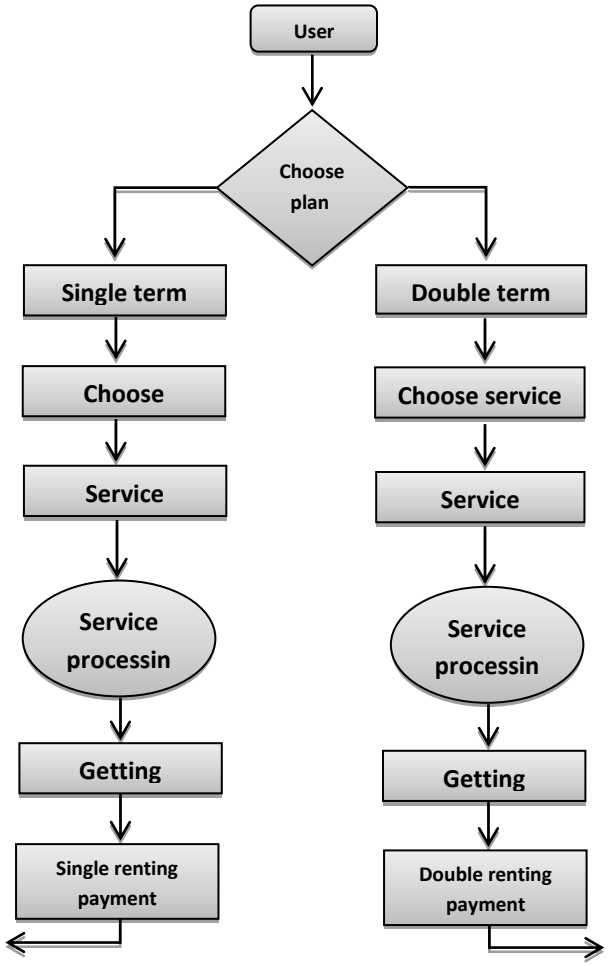

Fig 1: Flow Diagram

The gathering of requirements and the examination of functionally abstracted specifications are the fundamental phases in creating the system architecture. The goal of this study is to demystify security concerns specific to cloud environments and to make security-related issues more understandable. The 2 models for systems with a waiting time-dependent service discipline are discussed in this study.The 1st model employs a single server that continually adjusts its service rate depending on how long first customer in line has been waiting, as opposed to the second model, which involves a single queue served by a primary server that is supplemented by a backup server when the 1stconsumer in line is waiting for more time than a threshold. Finding the steady-state waiting time distribution for queueing systems whose service qualities rely on how long customers wait in line before being served is the major objective of this work. Despite being widely used in the industry, this sort of service control is new in queuing literature.

Following section explains several stages that are involved in putting the suggested technique into practice:

User Login:

The user's ID (email-id) is required to log in. Virtual machines (VMs) are resources offered by the clouds for use in tasks. A task queuing system like SGE, PBS, or Condor is employed in the cloud, where the users also upload their jobs. The job scheduler centrally schedules every work and distributes it to several VMs. Consequently, we may think of it as a queue for service requests.

Plan choosing:

The user chooses whether they want a short-term or long-term plan. A specialized workload management system called Condor offers a task queuing mechanism, priority scheme, scheduling strategy, resource management, and resource monitoring for compute-intensive workloads. Condor receives user-submitted jobs, adds them to a queue, and decides where and when to execute them depending on a set of rules.

Service accessing:

The user's utilization of various services and their access to them are measured in the background process. A service provider creates a virtual machine (VM) by renting resources from infrastructure providers to put together a collection of services. There are 2kinds of resource rental programs presented by infrastructure suppliers.

Viewing payment:

Payment information is provided to the user for services accessed and the frequency of each service's use. A consumer submits a service request to a service provider who provides services on demand. a specific SLA, depending on the service's scope and level of quality, the client receives pays for the service after the expected result from provider. Customers pay service providers for handling their service requests, which creates cost, while infrastructure providers are paid for renting out their physical resources, which generates income.

Viewing customer details:

The admin may access client information such as payment information, each customer's plan, etc.

Profit analysis:

Profit analysis of the higher quality services offered to the client. We compare the maximum profit made by our DQG renting scheme with that of SQU renting system with the aim to further demonstrate the

superior performance of our suggested scheme with regards to profit. The settings for this series of comparisons are the same as those in Section 5, with set to 6.99, D to 5, r ranging from 0.75 to 2.00 in steps of 0.25, and the other variables being the same. Two rental schemes are set up in accordance with the optimal profit and are displayed.

4. RESULTS

The use of the cloud to provide clients with computer resources and services on demand has increased in popularity. From the standpointof cloud service providers, revenue is among the most crucial variables, and it is mostly influenced by how a cloud service platform is structured considering the market requirement. This study initially creates a two-part resource leasing model that mixes short-term and long-term rents with the aim to address the current issues. A service system is modeled as an M/M/m+D queuing model, and performance metrics that influence the twofold renting scheme's profitability are looked at. The ideal architecture of a cloud platform is developed by finding a solution for the profit maximization issue for the twofold renting scheme.Several calculations are eventually carried out to evaluate the profitability of our recommended strategy as compared to the single rental plan. To ensure the caliber of service requests and increase service providers' profits, the study has created a special Double-Quality-Guaranteed (DQG) leasing system. With this strategy, short-term and long-term rentals are merged, potentially reducing resource waste and allowing for flexibility in response to changing demand for computing power. An ideal configuration issue for profit maximization is developed using an M/M/m+D queueing model. A variety of calculations are also done to compare the Single-Quality-Unguaranteed (SQU) and DQG rental schemes' profits. The outcomes show that the plan performs better than the SQU system with regards to profit and service quality.

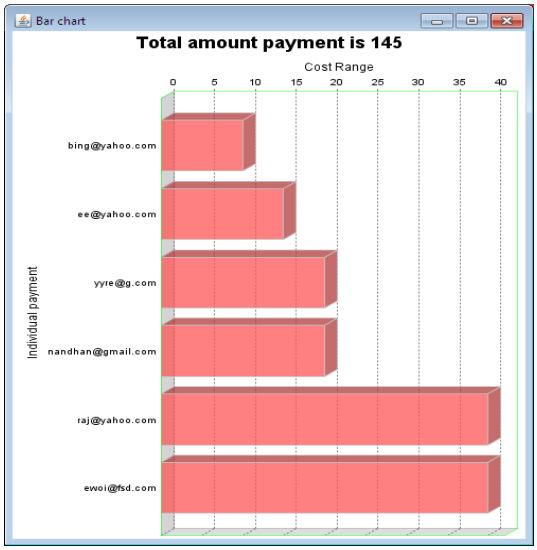

Fig 2: Performance Analysis

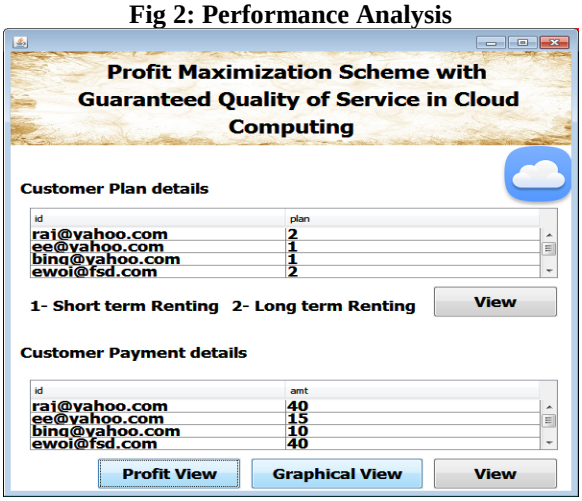

Fig 3: Profit View

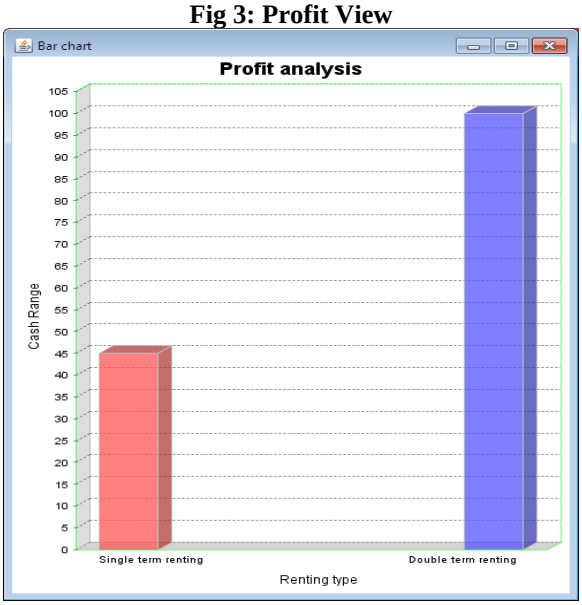

Fig 4: Comparative Analysis

5. CONCLUSION

This study has developed a unique Double-Quality-Guaranteed (DQG) rental system for service providers with the aim of ensuring the service quality requests and optimizes the revenue of service providers. This plan mixes short-term and long-term renting, which may significantly decrease wastage of resource and adjust to the fluctuating need for computing power. For our multiserver system with variable system size, an M/M/m+D queueing model is developed. Then, an ideal configuration problem for profit maximization is created, taking into account a number of variables like request workload, SLA, cost of energy consumption, server rental cost, market demand, and other. The ideal optimum solutions and the real optimal solutions are two separate conditions for which the optimal solutions are solved. A number of computations are made to compare the profit made by the Single-Quality-Unguaranteed(SQU) and DQG rental schemes. According to the findings, our plan beats the SQU system with regards to both profit and service quality.

REFERENCES

[1] R. Buyya, C. S. Yeo, S. Venugopal, J. Broberg, andI. Brandic, "Cloud computing and emerging it platforms:Vision, hype, and reality for delivering computing as the5th utility," Future Gener. Comp. Sy., vol. 25, no. 6, pp. 599–616, 2009.

[2] 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.

[3] J. Chen, C. Wang, B. B. Zhou, L. Sun, Y. C. Lee, andA. Y. Zomaya, "Tradeoffs between profit and customersatisfaction for service provisioning in the cloud," in Proc.20th Int'l Symp. High Performance Distributed Computing.ACM, 2011, pp. 229–238.

[4] J. Mei, K. Li, J. Hu, S. Yin, and E. H.-M. Sha, "Energyaware preemptive scheduling algorithm for sporadic tasksondvs platform," MICROPROCESS MICROSY., vol. 37,no. 1, pp. 99–112, 2013.

[5] G. Kesidis, A. Das, and G. de Veciana, "On flat-rateand usage-based pricing for tiered commodity internetservices," in 42nd Annual Conf. Information Sciences andSystems. IEEE, 2008, pp. 304–308.

[6] J. Cao, K. Hwang, K. Li, and A. Y. Zomaya, "Optimalmultiserver configuration for profit maximization in cloudcomputing," IEEE Trans. Parallel Distrib. Syst., vol. 24, no. 6,pp. 1087–1096, 2013.

[7] A. Fox, R. Griffith, A. Joseph, R. Katz, A. Konwinski,G. Lee, D. Patterson, A. Rabkin, and I. Stoica, "Abovethe clouds: A berkeley view of cloud computing," Dept.Electrical Eng. and Comput. Sciences, vol. 28, 2009.

[8] P. de Langen and B. Juurlink, "Leakage-aware multiprocessor scheduling," J. Signal Process. Sys., vol. 57, no. 1, pp.73–88, 2009.

[9] G. P. Cachon and P. Feldman, "Dynamic versus staticpricing in the presence of strategic consumers," Tech. Rep.,2010.

[10] Y. C. Lee, C. Wang, A. Y. Zomaya, and B. B. Zhou, "Profitdriven scheduling for cloud services with data accessawareness," J. Parallel Distr. Com., vol. 72, no. 4, pp. 591–602, 2012.

[11] M. Ghamkhari and H. Mohsenian-Rad, "Energy and performance management of green data centers: a profit maximization approach," IEEE Trans. Smart Grid, vol. 4, no. 2,pp. 1017–1025, 2013.

[12] A. Odlyzko, "Should flat-rate internet pricing continue,"IT Professional, vol. 2, no. 5, pp. 48–51, 2000.

[13] K. Hwang, J. Dongarra, and G. C. Fox, Distributed andCloud Computing. Elsevier/Morgan Kaufmann, 2012.

[14] S. Shakkottai, R. Srikant, A. Ozdaglar, and D. Acemoglu,"The price of simplicity," IEEE J. Selected Areas in Communications, vol. 26, no. 7, pp. 1269–1276, 2008.

[15] H. Xu and B. Li, "Dynamic cloud pricing for revenuemaximization," IEEE Trans. Cloud Computing, vol. 1, no. 2,pp. 158–171, July 2013.