ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 64 - 67 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

OMC USED IN CLOUDS FOR MAXIMUM PROFIT

[1]J Sravanthi, [2]Devavarapu Sreenivasarao,[3]Bandaru A Chakravarthi

[1]Student of M.Tech, Dept CSE/IT, MallaReddy Institute of Eng. & Tech., Hyd. j.sravanthi526@gmail.com
[2]Associate Professor, Dept CSE/IT MallaReddy Institute of Eng & Tech, Hyd, sreenivasaraodevavarapu@gmail.com
[3]Assistant Professor, Dept CSE/IT MallaReddy Institute of Eng & Tech, Hyd, amar.chakry@gmail.com

ABSTRACT: To maximize the profit, a service provider should understand both service charges and business costs, and how they are determined by the characteristics of the applications and the configuration of a multi server system. The problem of optimal multi server configuration for profit maximization in a cloud computing environment is studied. Our pricing model takes such factors into considerations as the amount of a service, the workload of an application environment, the configuration of a multi server system, the service-level agreement, the satisfaction of a consumer, the quality of a service, the penalty of a low-quality service, the cost of renting, the cost of energy consumption, and a service provider's margin and profit. Our approach is to treat a multi server system as an M/M/m queuing model, such that our optimization problem can be formulated and solved analytically. Two server speed and power consumption models are considered, namely, the idle-speed model and the constant-speed model. The probability density function of the waiting time of a newly arrived service request is derived. The expected service charge to a service request is calculated. The expected net business gain in one unit of time is obtained. Numerical calculations of the optimal server size and the optimal server speed are demonstrated

KEYWORDS: Cloud computing, multi server system, pricing model, profit, queuing model, response time.

I. INTRODUCTION

CLOUD computing is quickly becoming an effective and efficient way of computing resources and computing services consolidation. By centralized management of resources and services, cloud computing delivers hosted services over the Internet, such that accesses to shared hardware, software, databases, information, and all resources computing is able to provide the most cost-effective and energy-efficient way of computing resources management and computing services provision. Cloud computing turns information technology into ordinary commodities and utilities by using the pay-per-use pricing model . However, cloud computing will never be free, and understanding the economics of cloud computing becomes critically important. One attractive cloud computing environment is a three tier=structure, which consists of infrastructure vendors, service providers, and consumers. The three parties are also called cluster nodes, cluster managers, and consumers in cluster computing systems, and resource providers, service providers, and clients in grid computing systems. An infrastructure vendor maintains basic hardware and software facilities. A service provider can build different multi server systems for different application domains, such that service requests of different nature are sent to different multi server systems. Each multi server system contains multiple servers, and such a multi server system can be devoted to serve one type of service requests and applications. An application domain is characterized by two basic features, i.e., the workload of an application environment and the expected amount of a service. The configuration of a multi server system is characterized by two basic features, i.e., the size of the multi server system (the number of servers) and the speed of the multi server system (execution speed of the servers).Like all business, the pricing model of a service provider in cloud computing is based on two components, namely, the income and the cost. For a service provider, the income (i.e., the revenue) is the service charge to users, and the cost is the renting cost plus the utility cost paid to infrastructure vendors. A pricing model in cloud computing includes many considerations, such as the amount of a service (the requirement of a service), the workload of an application environment, the configuration (the size and the speed) of a multi server system, the service-level agreement, the satisfaction of a consumer (the expected service time), the quality of a service (the task waiting time and the task response time), the penalty of a low-quality service, the cost of renting, the cost of energy consumption, and a service provider's margin and profit. The profit (i.e., the net business gain) is the income minus the cost. To maximize the profit, a service provider should understand both service charges and business costs, and in particular, how they are determined by the characteristics of the applications and the configuration of a multiserver system. The service

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 64 - 67 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

charge to a service request is determined by two factors, i.e., the expected length of the service and the actual length of the service. The expected length of a service (i.e., the expected service time) is the execution time of an application on a standard server with a baseline or reference speed. Once the baseline speed is set, the expected length of a service is determined by a service request itself, i.e., the service requirement (amount of service) measured by the number of instructions to be executed. The longer (shorter, respectively) the expected length of a service is, the more (less, respectively) the service charge is. The actual length of a service (i.e., the actual service time) is the actual execution time of an application. The actual length of a service depends on the size of a multi server system, the speed of the servers (which may be faster or slower than the baseline speed), and the workload of the multi server system. Notice that the actual service time is a random variable, which is determined by the task waiting time once a multi server system is established. There are many different service performance metrics in service-level agreements. Our performance metric in this paper is the task response time (or the turn around time), i.e., the time taken to complete a task, which includes task waiting time and task execution time. The service-level agreement is the promised time to complete a service, which is a constant times the expected length of a service.

If the actual length of a service is (or, a service request is completed) within the service-level agreement, the service will be fully charged. However, if the actual length of a service exceeds the service-level agreement, the service charge will be reduced. The longer (shorter, respectively) the actual length of a service is, the more (less, respectively) the reduction of the service charge is. In other words, there is penalty for a service provider to break a service-level agreement. If the actual service time exceeds certain limit (which is service request dependent), a service will be entirely free with no charge. Notice that the service charge of a service request is a random variable, and we are interested in its expectation. The cost of a service provider includes two components, i.e., the renting cost and the utility cost. The renting cost is proportional to the size of a multi server system, i.e., the number of servers. The utility cost is essentially the cost of energy consumption and is determined by both the size and the speed of a multi server system. The faster (slower, respectively) the speed is, the more (less, respectively) the utility cost is. To calculate the cost of energy consumption, we need to establish certain server speed and power consumption models. Hence, a powerful multi server system reduces the penalty of breaking a service-level agreement and increases the revenue. However, more servers (i.e., a larger multi server system) increase the cost of facility renting from the infrastructure vendors and the cost of base power consumption. Furthermore, faster servers increase the cost of energy consumption. Such increased cost may counterweight the gain from penalty reduction.

II.PROPOSED WORK

Therefore, for an application environment with specific workload which includes the task arrival rate and the average task execution requirement, a service provider needs to decide an optimal multi server configuration (i.e., the size and the speed of a multi server system), such that the expected profit is maximized. In this paper, we study the problem of optimal multi server configuration for profit maximization in a cloud computing environment. Our approach is to treat a multi server system as an M/M/m queuing model, such that our optimization problem can be formulated and solved analytically. We consider two server speed and power consumption models, namely, the idle-speed model and the constant-speed model. Our main contributions are as follows.

To the best of our knowledge, there has been no similar investigation in the literature, although the method of optimal multi core server processor configuration has been employed for other purposes, such as managing the power and performance trade off. One related research is user-centric and market-based and utility-driven resource management and task scheduling, which have been considered for cluster computing systems and grid computing systems. To compete and bid for shared computing resources through the use of economic mechanisms such as auctions, a user can specify the value (utility, yield) of a task, i.e., the reward (price, profit) of completing the task.

III. MODELS

MULTISERVER MODEL

A cloud computing service provider serves users' service requests by using a multi server system, which is constructed and maintained by an infrastructure vendor and rented by the service provider. The architecture detail of the multi- server system can be quite flexible.

WAITING TIME DISTRIBUTION

Let W denote the waiting time of a new service request that arrives to a multi server system. To this end, we consider W in different situations, depending on the number of tasks in the queuing system when a new service request arrives. Let Wk denote the waiting time of a new task that arrives to an M/M/m queuing system under the condition that there are k tasks in the queuing system when the task arrives.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 64 - 67 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

We define a *unit impulse function* $u_z(t)$ as follows:

$$u_z(t) = \begin{cases} z, & 0 \le t \le \frac{1}{z}; \\ 0, & t > \frac{1}{z}. \end{cases}$$

The function $u_z(t)$ has the following property:

$$\int_0^\infty u_z(t)dt = 1,$$

namely, $u_z(t)$ can be treated as a pdf of a random variable with expectation

$$\int_0^\infty t u_z(t)dt = z\int_0^{1/z} tdt = \frac{1}{2z}.$$

Notice that a multi server system with multiple identical servers has been configured to serve requests from certain application domain. Therefore, we will only focus on task waiting time in a waiting queue and do not consider other sources of delay, such as resource allocation and provision, virtual machine instantiation and deployment, and other overhead in a complex cloud computing environment

SERVICE CHARGE

If all the servers have a fixed speed s, the execution time of a service request with execution requirement. The response time T is related to the service charge to a customer of a service provider in cloud computing. To study the expected service charge to a customer, we need a complete specification of a service charge based on the amount of a service, the service-level agreement, the satisfaction of a consumer, the quality of a service, the penalty of a low-quality service, and a service provider's margin and profit. Let s0 be the baseline speed of a server. We define the service charge function for a service request with execution requirement r and response time T to be

$$C(r,T) = \begin{cases} ar, & \text{if } 0 \le T \le \frac{c}{s_0}r; \\ ar - d\left(T - \frac{c}{s_0}r\right), & \\ & \text{if } \frac{c}{s_0}r < T \le \left(\frac{a}{d} + \frac{c}{s_0}\right)r; \\ 0, & \text{if } T > \left(\frac{a}{d} + \frac{c}{s_0}\right)r. \end{cases}$$

If the response time T to process a service request is no longer than a constant c times the task execution time with speed s0), where the constant c is a parameter indicating the service level agreement, and the constant s0 is a parameter indicating the expectation and satisfaction of a consumer, then a service provider considers that the service request is processed successfully with high quality of service and charges a customer ar, which is linearly proportional to the task execution requirement r(i.e., the amount of service), where a is the service charge per unit amount of service (i.e., a service provider's margin and profit). . If the response time T to process a service request is longer than but no longer than, then a service provider considers that the service request is processed with low quality of service and the charge to a customer should decrease linearly as T increases. The parameter d indicates the degree of penalty of breaking the service-level agreement then a service provider considers that the service request has been waiting too long, so there is no charge and the service is free.

NET BUSINESS GAIN

Since the number of service requests processed in one unit of time is _ in a stable M/M/m queuing system, the expected service charge in one unit of time is which is actually the expected revenue of a service provider. Assume that the rental cost of one server for unit of time is Also, assume that the cost of energy is per Watt. The cost of a service provider is the sum of the cost of infrastructure renting and the cost of energy consumption, i.e., _m þ P Then, the expected net business gain (i.e., the net profit) of a service provider in one unit of

In the first case, there is no enough business (i.e., service requests). In this case, a service provider should consider reducing the number of servers m and/or server speed s, so that the cost of infrastructure renting and the cost of energy consumption can be reduced. In the second case, there is too much business (i.e., service requests). In this case, a service provider should consider increasing the number of servers and/or server speed, so that the waiting time can be reduced and the revenue can be increased.

IV. RESULTS

To formulate and solve our optimization problems analytically, we need a closed-form expression of C. To this end, let us use the following closed-form approximation.

ISSN 2278-3091

International Journal of Advanced Trends in Computer Science and Engineering, Vol. 3 , No.1, Pages : 64 - 67 (2014)
Special Issue of ICETETS 2014 - Held on 24-25 February, 2014 in Malla Reddy Institute of Engineering and Technology, Secunderabad– 14, AP, India

7.2 Optimal Speed

Given λ, $\bar{r}$, m, P^*, α, β, γ, a, c, and d, our second problem is to find s such that G is maximized. To maximize G, we need to find s such that

$$\frac{\partial G}{\partial s} = \lambda \frac{\partial C}{\partial s} - \gamma \lambda \bar{r} \xi (\alpha - 1) s^{\alpha - 2} = 0,$$

for the idle-speed model, and

$$\frac{\partial G}{\partial s} = \lambda \frac{\partial C}{\partial s} - \gamma m \xi \alpha s^{\alpha - 1} = 0,$$

for the constant-speed model, where

$$\frac{\partial C}{\partial s} = \frac{a\bar{r}}{(D_1 D_2 D_3)^2} \times \left(D_2 D_3 \frac{\partial D_1}{\partial s} + D_1 D_3 \frac{\partial D_2}{\partial s} + D_1 D_2 \frac{\partial D_3}{\partial s} \right).$$

Similar to the calculation in the last section, we have

$$\frac{\partial \rho}{\partial s} = -\frac{\lambda \bar{r}}{m s^2} = -\frac{\rho}{s},$$

and

$$\frac{1}{R}\frac{\partial R}{\partial s} = m\left(1 - \frac{1}{\rho}\right)\frac{\partial \rho}{\partial s} = \frac{m}{s}(1 - \rho),$$

and

$$\frac{\partial R}{\partial s} = \frac{m}{s}(1 - \rho)R.$$

V. CONCLUSION

We have proposed a pricing model for cloud computing which takes many factors into considerations, such as the requirement r of a service, the workload _ of an application environment, the configuration (m and s) of a multiserver system, the service level agreement c, the satisfaction (r and s0) of a consumer, the quality (W and T) of a service, the penalty d of a low-quality service, the cost (_ and m) of renting, the cost (P_, and P) of energy consumption, and a service provider's margin and profit a. By using an M/M/m queuing model, we formulated and solved the problem of optimal multiserver configuration for profit maximization in a cloud computing environment. Our discussion can be easily extended to other service charge functions. Our methodology can be applied to other pricing models

VI. REFERENCES

[1] http://en.wikipedia.org/wiki/CMOS, 2012.
[2]http://en.wikipedia.org/wiki/Service_level_agreement, 2012.
[3] M. Armbrust et al., "Above the Clouds: A Berkeley View of Cloud Computing," Technical Report No. UCB/EECS-2009-28, Feb. 2009.
[4] R. Buyya, D. Abramson, J. Giddy, and H. Stockinger, "Economic Models for Resource Management and Scheduling in Grid Computing," Concurrency and Computation: Practice and Experience, vol. 14, pp. 1507-1542, 2007.
[5] 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 Fifth Utility," Future
Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 2009.
[6] A.P. Chandrakasan, S. Sheng, and R.W. Brodersen, "Low-Power CMOS Digital Design," IEEE J. Solid-State Circuits, vol. 27, no. 4, pp. 473-484, Apr. 1992.
[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 Performability 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] D.E. Irwin, L.E. Grit, and J.S. Chase, "Balancing Risk and Reward
in a Market-Based Task Service," Proc. 13th IEEE Int'l Symp. HighPerformance Distributed Computing, pp. 160-169, 2004.
[13] H. Khazaei, J. Misic, and V.B. Misic, "Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems,"IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 5,pp. 936-943, May 2012.
[14] L. Kleinrock, Queueing Systems: Theory, vol. 1. John Wiley and Sons, 1975.
[15] Y.C. Lee, C. Wang, A.Y. Zomaya, and B.B. Zhou, "Profit-Driven Service Request Scheduling in Clouds," Proc. 10th IEEE/ACM Int'l Conf. Cluster, Cloud and Grid Computing, pp. 15-24, 2010.
[16] K. Li, "Optimal Load Distribution for Multiple Heterogeneous Blade Servers in a Cloud Computing Environment," Proc. 25th IEEE Int'l Parallel and Distributed Processing Symp. Workshops, pp. 943-952, May 2011.
[17] K. Li, "Optimal Configuration of a Multicore Server Processor for Managing the Power and Performance Tradeoff," J. Supercomputing,
vol. 61, no. 1, pp. 189-214, 2012.
[18] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," Nat'l Inst. of Standards and Technology, http://csrc.nist.gov/groups/SNS/cloud-computing/, 2009.
[19] F.I. Popovici and J. Wilkes, "Profitable Services in an Uncertain World," Proc. ACM/IEEE Conf. Supercomputing, 2005.
[20] J. Sherwani, N. Ali, N. Lotia, Z. Hayat, and R. Buyya, "Libra: A Computational Economy-Based Job Scheduling System for Clusters," Software - Practice and Experience, vol. 34, pp. 573-590, 2004.
[21] C.S. Yeo and R. Buyya, "A Taxonomy of Market-Based Resource Management Systems for Utility-Driven Cluster Computing," Software - Practice and Experience, vol. 36, pp. 1381-1419, 2006.