ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Dynamic Resource Allocation Using VM for Cloud Computing Environment

V.Indhumathi[*1], N.Nagalakshmi[2]

[*1, 2] Student, Department of CSE, Jayam College of Engineering & Technology, Dharmapuri-636813, Tamilnadu, India

indu063@gmail.com

Abstract

As cloud computing becomes more and more general, kind the economics of cloud computing becomes censoriously important. To maximize the profit, a cloud service provider should understand both service charges and business costs, and how they are gritty by the characteristics of the applications and the configuration of a multi-server system. The problem of optimal multi-server arrangement for profit maximization in a cloud computing environment is studied. The pricing model takes such factors into considerations as the amount of a service, the workload of an application location, the formation of a multi-server system, the service-level agreement, the satisfaction of a consumer, the quality of a service, the disadvantage of a low-quality service, the cost of renting, and the cost of energy depletion, and a service provider's margin and profit. The approach is to treat a multi-server system as an M/M/m queuing model, such that the optimization problem can be formulated and solved critically.Two server speed and power consumption models are measured, namely, the idle-speed model and the constant-speed model. The probability mass function of the waiting time of a newly arrived service request is derived. The expected service charge to a cloud service request is considered. The expected remaining business gain in one unit of time is gained. Numerical calculations of the optimal server size and the optimal server speed are established.

Keywords: Cloud computing multi-server system, pricing model, profit and queuing model, response time, server formation, service charge, service-level covenant, waiting time..

Introduction

As cloud computing becomes more and more general, kind the economics of cloud computing becomes censoriously important. To maximize the profit, a cloud service provider should understand both service charges and business costs, and how they are gritty by the characteristics of the applications and the configuration of a multi-server system..

The problem of optimal multi-server arrangement for profit maximization in a cloud computing environment is studied. The pricing model takes such factors into considerations as the amount of a service, the workload of an application location, the formation of a multi-server system, the service-level agreement, the satisfaction of a consumer, the quality of a service, the disadvantage of a low-quality service, the cost of renting, and the cost of energy depletion, and a service provider's margin and profit.

The approach is to treat a multi-server system as an M/M/m queuing model, such that the optimization problem can be formulated and solved critically. Two server speed and power consumption models are measured, namely, the idle-speed model and the constant-speed model. The probability mass function of the waiting time of a newly arrived service request is derived. The expected service charge to a cloud service request is considered. The expected remaining business gain in one unit of time is gained.

Numerical calculations of the optimal server size and the optimal server speed are established. Cloud computing is a large-scale distributed computing paradigm in which a pool of computing resources is available to users via the Internet. Computing resources, e.g., processing control, storage, software, and network bandwidth, are represent to cloud consumers as the accessible public utility services. Infrastructure- as-a-Service is a computational service model widely applied in the cloud computing theory. In the model, virtualization technology can be used to provide resources to cloud consumers.

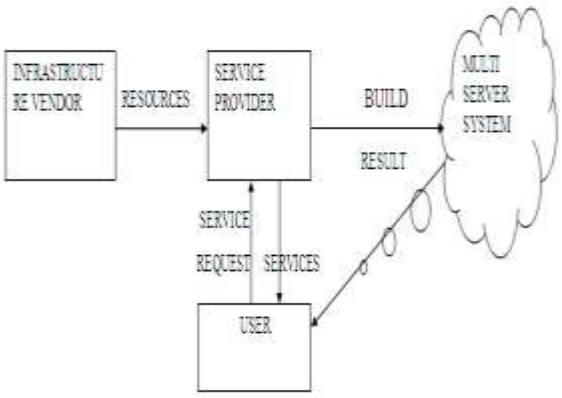
The consumers can identify the required software load, e.g., operating systems and applications; then package them all together into virtual machines. The hardware requirement of VMs can also be adjusted by the consumers. at last, those VMs will be outsourced to host in computing environments operated by third-party sites owned by

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

cloud providers. A cloud provider is responsible for guarantee the Quality of Services for running the VMs. Since the computing resources are maintain by the donor, the total cost of ownership to the consumers can be bargain. In cloud computing, a resource provisioning mechanism is required to supply cloud consumers a set of computing resources for processing the jobs and storing the data. Cloud provider can offer cloud consumers two resource provisioning plans, namely short-range on-demand and long-term reservation plans.

Amazon EC2 and Go Grid are, for instance, cloud providers which offer IaaS services with both plans. In general, pricing in on-demand plan is charged by pay-per-use basis. Then, purchasing this on-demand plan, the consumers can dynamically provision resources at the moment when the resources are needed to fit the fluctuated and changeable demands. For reservation plan, pricing is charged by a onetime fee typically before the computing resource will be utilized by cloud consumer. With the reservation plan, the cost to utilize resources is cheaper than that of the on-demand plan.

In this way, the consumer can decrease the cost of computing resource provisioning by using the reservation plan. For example, the reservation plan accessible by Amazon EC2 can reduce the total provisioning cost up to 49 percent when the reserved resource is fully utilized.



The cost of a service provider includes two components, i.e., the renting cost and the service cost. The renting cost is proportional to the size of a multi server system, i.e., the number of servers. The service cost is essentially the cost of energy consumption and is determined by both the size and the speed of a multiserver system. The earlier (slower, respectively) the speed is, the more (less, respectively) the utility cost is. To calculate the cost of energy utilization, the need to establish certain server speed and power consumption models. To increase the profits of business, a service provider can construct and configure a multi server system with many servers of high speed. Since the actual service time (i.e., the task response time) contains job waiting time and task execution time, more servers decrease the waiting time and faster servers reduce both waiting time and execution time. Hence, a powerful multi server system reduces the penalty of breaking a service-level agreement and increases the profits. However, more servers (i.e., a larger multiserver system) increase the cost of facility renting from the infrastructure vendors and the cost of base power consumption. Moreover, faster servers raise the cost of energy exploit. Such increased cost may counterweight the gain from penalty decrease. Therefore, for an application environment with specific workload which includes the task arrival rate and the average task execution obligation, a service provider needs to decide an optimal multiserver configuration (i.e., the size and the speed of a multiserver system), such that the expected profit is maximize.

In this paper, study the problem of optimal multiserver configuration for profit maximization in a cloud computing environment. The approach is to treat a multi server system as an M/M/m queuing model, such that the optimization problem can be formulated and solved analytically. They consider two server speed and power utilization models, namely, the idle-speed model and the constant-speed model. The main contributions are as follows. We derive the probability mass function of the waiting time of a newly arrived service request. This result is important in its own right and is the base of the discussion.

To calculate the expected service charge to a service request. Based on these results, to get the expected net business gain in one unit of time, and find the optimal server size and the optimal server speed numerically. To the best of the 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 purpose, such as managing the power and performance tradeoff.

A multiserver model

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 and0) of a consumer, the quality (W and T) of a service, the price of a low-quality service, the cost and mof

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

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 e-search 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.

A cloud computing service provider serves users' service requests by using a multiserver system, which is construct and maintained by an infrastructure vendor and rented by the service source. The architecture detail of the multiserver system can be quite flexible. Assume that a multiserver system S has m identical servers. In this paper, a multi server system is treated as an M/M/m queuing system which is elaborate as follows. There is a Poisson stream of service requests with arrival rate i.e., the inter arrival times are independent and identically distributed (i.i.d.) exponential random variables with mean 1=_. A multiserver system S maintains a queue with infinite capacity for waiting tasks when all the m servers are busy.

The first-come-first-served (FCFS) queuing discipline is adopted. The task execution requirements (measured by the number of instructions to be executed) are i.i.d. exponential random variables r with mean _r.The m servers (i.e., blades/processors/cores) of S have identical execution speed s (measured by the number of instructions that can be executed in one unit of time). thus, the task execution times on the servers of S are i.i.d. exponential random variables x ¼ r=s with mean _x ¼ _r=s. Notice that although an M/G/m queuing system has been considered (see, e.g., [13]), the M/M/m queuing model is the only model that accommodates an analytical and closed form expression of the probability density function of the waiting time of a newly arrived service request.

Power consumption model

In the existing literature, the power consumption of a server has been modelled in two different ways: offline and online. In the former case, Simple Power, Software Watt and Mambo estimate the power consumption of an entire server. though these models use analytical methods based on some low-level information such as number of used CPU cycles. The main advantage is that they provide high accuracy. however, the offline nature of such models requires extensive simulation, which results in a important amount of time for estimating the power consumption .accordingly, these models are infeasible for predicting the power consumption of highly dynamic environments like cloud computing data centres. To conquer this problem, an online (run-time) methodology is proposed. Such models are based on the information monitored through performance counters. These counters keep track of activities performed by applications such as amount of accesses (e.g. to caches) and switching activities within processors.

The total power dissipation of a server is computed as the power consumption of each activity.Although, these counters in certain processors (e.g. AMD Opteron) can report only four out of 84 events. Therefore, such models are unable to predict accurate power consumption in real-life cases. Another run-time methodology is to use high-level information as the one proposed by . These authors assumed that processors are the main contributors to the total server's power utilization. Thus, a linear model based on the processor's utilisation is proposed. How-ever, such a model suffers from a significant inaccuracy as the server's power consumption is not exactly linear.

The key reason is that the impact of other components (e.g. multiple level caches, RAM, I/O activities) and their interactions are not considered. To prevent this problem designed a constituent level model. Through this approach, a calibration phase is performed before predicting the power consumption of a server.

For the duration of this phase, this model analyses the system parameters (e.g. CPU utilisation, hard disk I/O rate) influencing on its power consumption. Though, implementing such a model within a data centre (having homogenous and/or heterogenous resources) is very difficult since it needs calibration whenever a new hardware is installed within the existing servers. Because the component level is flexible for modelling a generic server, they have also adopted the same approach.

In contrast to this model, which provides one linear model for the whole servers, the approach designs different models for different components based on their behaviours. Another characteristic property is that the approach does not need calibration phase. It is worthwhile to note that above mentioned approach, which depend on low-level information in order to predict the power consumption, are not suitable in real-life case simply because the

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

underlying monitoring systems of the data centres are not able to provide the low-level information that these approaches require. Consequently,

in this paper the identified the most-relevant energy-related attributes of ICT resources to which the monitoring systems of data centres typically provide the necessary information.

Server configuration

Plan reservation

In reservation plan, the cloud uses reserve the cloud in advance for their requirements. In this way, the pay the payment of the reservation in on the spot. That is, when we will reserve the cloud space mean, at the time we pay the payment also.

Space utilization

The space timing calculates by the reference of cloud usage. That is, the cost also calculates based on cloud space utilization and cloud usage

Analysis of performance

The analyze and compare the performance offered by different configurations of the computing collect, focused in the execution of loosely coupled applications. In particular, we have chosen nine different cluster configurations with different number of worker nodes from the three clouds**.** Providers and different number of Jobs (depending on the cluster size), as shown in the definition of the different cluster configurations,

we use the following acronyms infrastructure; Amazon EC2 Europe cloud AmazonEC2 US cloud and flexible Hosts cloud. The number preceding the site acronym represents the number of worker nodes. For example, is a cluster with four worker nodes deployed in the local infrastructure; and is an eight-node cluster, four deploy in the local infrastructure and four in Amazon. To represent the execution profile of loosely coupled applications, the will use the Embarrassingly Distributed benchmark from the Numerical Aero dynamic Simulation Benchmarks.

Analysing the priority

The want to enable the use of large-scale distributed systems for task-parallel applications, which are linked into useful workflows through the looser task coupling model of passing data via files between dependent tasks .This potentially larger class of task-parallel quality Extraction. The need to expand the computational resources in a massive surveillance network is clear but traditional means of purchasing new equipment for short-term tasks every year is inefficient.

In this work I will provide evidence in support of utilizing a cloud computing infrastructure to perform computationally intensive feature extraction tasks on data streams. Competent off-loading of computational tasks to cloud resources will involve a minimization of the Time needed to expand the cloud resources, an professional model of communication and a study of the interaction between the in-network computational resources and remote resources in the cloud.

Task scheduling

Each and every user assigns the task to cloud, so that task will assign to the cloud in priority scheduling basis or if anyone cloud is free mean, user job assign to that cloud

Conclusion

They have proposed a pricing model for cloud computing which takes many factors into consideration, such as the requirement r 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. The discussion can be easily extended to other service charge functions.

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*

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

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 SpeedStep 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. *High Performance 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. *B. Zhai, D. Blaauw, D. Sylvester, and K. Flautner, "Theoretical and Practical Limits of Dynamic Voltage Scaling," Proc. 41st Design Automation Conf., pp. 868-873, 2004.*