ijecs open access www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 03 Issue 07 July, 2014 Page No. 7032-7034

A Generosity Model to Study Data Center Performance and QoS in IaaS Cloud Computing Systems

D.S.Selvi, A.Balasubramani, P.Nirupama
M.Tech,Department Of Cse
Sietk,Puttur,India
Professor
Department Of Cse
Sietk,Puttur,India
Head Of Department
Department Of Cse
Sietk,Puttur,India

ABSTRACT

Cloud computing is a general term for system architectures that involves delivering hosted services over the Internet, made possible by significant innovations in virtualization and distributed computing, as well as improved access to high-speed Internet. A cloud service differs from traditional hosting in three principal aspects. First, it is provided on demand, typically by the minute or the hour; second, it is elastic since the user can have as much or as little of a service as they want at any given time; and third, the service is fully managed by the provider – user needs little more than computer and Internet access. Typically a contract is negotiated and agreed between a customer and a service provider; the service provider is required to execute service requests from a customer within negotiated quality of service (QoS) requirements for a given price.

Due to dynamic nature of cloud environments, diversity of user's requests, resource virtualization, and time dependency of load, providing expected quality of service while avoiding over-provisioning is not a simple task. To this end, cloud provider must have efficient and accurate techniques for performance evaluation of cloud computing centers. The development of such techniques is the focus of this thesis. This thesis has two parts. In first part, monolithic performance models are developed for cloud computing performance analysis. Poisson task arrivals, generally distributed service times, and a large number of physical servers. Later on, to extend the model to include finite buffer capacity, batch task arrivals, and virtualized servers with a large number of virtual machines in each physical machine. However, a monolithic model may suffer from intractability and poor scalability due to large number of parameters. Therefore, in the second part of the thesis we develop and evaluate tractable functional performance sub-models for different servicing steps in a complex cloud center and the overall solution obtain by iteration over individual sub-model solutions. Also extend the proposed interacting analytical sub-models to capture other important aspects including pool management, power consumption, resource assigning process and virtual machine deployment of nowadays cloud centers. Finally, a performance model suitable for cloud computing centers with heterogeneous requests and resources using interacting stochastic models is proposed and evaluated.

Index Terms—Cloud computing, Federation, stochastic reward nets, cloud-oriented performance metrics, resiliency, responsiveness.

1.INTRODUCTION

Cloud Computing as a new paradigm of computing. The main challenge and obstacle of cloud computing, Quality of Service (QoS), is defined and the general aspects of QoS are explored. Later the performance evaluation in cloud computing. With the advancement of human society, basic essential services are commonly provided such that everyone can easily obtain access to them. Today, utility services such as water, electricity, gas, and telephony are deemed necessary for fulfilling daily life routines. These utility services are accessed so frequently that they need to be available at any time. Consumers are then able to pay service providers based on their usage of these utility services.

In 1969, Leonard Kleinrock said: "As of now, computer networks are still in their infancy, but as they grow up and become sophisticated, we will probably see the

spread of 'computer utilities' which, like present electric and telephone utilities, will service individual homes and offices across the country." This vision of the computing utility based on a service provisioning model anticipates the massive transformation of the entire computing industry in the 21st century, whereby computing services will be readily available on demand, like other utility services. As a result, there will be no need for the consumers to invest heavily in building and maintaining complex IT infrastructure ; instead, they will pay only when they access computing services. Nowadays, significant innovations in virtualization and distributed computing, as well as improved access to high-speed Internet, have accelerated interest in cloud computing . It is quickly gaining acceptance: According to IDC, 17 billion dollars was spent on cloud-related technologies, hardware and software in 2009, and spending is expected to grow to 45 billion by 2013 .

Cloud computing has a service oriented architecture in which services are broadly divided into three categories: Infrastructure-as-a-Service (IaaS), where equipment such as hardware, storage, servers, and networking components are made accessible over the Internet); Platform-asa- Service (PaaS), which includes computing platforms—hardware with operating systems, virtualized servers, and the like; and Software-as-a-Service (SaaS), which includes software applications and other hosted services . A cloud service differs from traditional hosting in three principal aspects. First, it is provided on demand, typically by the minute or the hour; second, it is elastic since the user can have as much or as little of a service as they want at any given time; and third, the service is fully managed by the provider. There is no unique definition for cloud computing. The definition of cloud computing as follows: "Cloud computing is a new computing paradigm, whereby shared resources such as infrastructure, hardware platform, and software applications are provided to users on-demand over the Internet (Cloud) as services." the computing paradigm shift of the last half century .

Authors et al. have listed 10 top obstacles and opportunities for growth of Cloud Computing .the first three concern adoption, the next five affect growth, and the last two are policy and business obstacles. Each obstacle is paired with an opportunity, ranging from product development to research projects, which can overcome that obstacle Although benefits and opportunities that cloud computing has been bringing about are tremendous, its challenges and problems that require huge amount of effort to be addressed by researchers. Since 2007 that cloud computing has gotten high attention until now, the trend of search volume index has been increasing sharply. The trend of search volume index for Cloud Computing versus Grid Computing using Google trend service during past years.

2.The Analytical Model

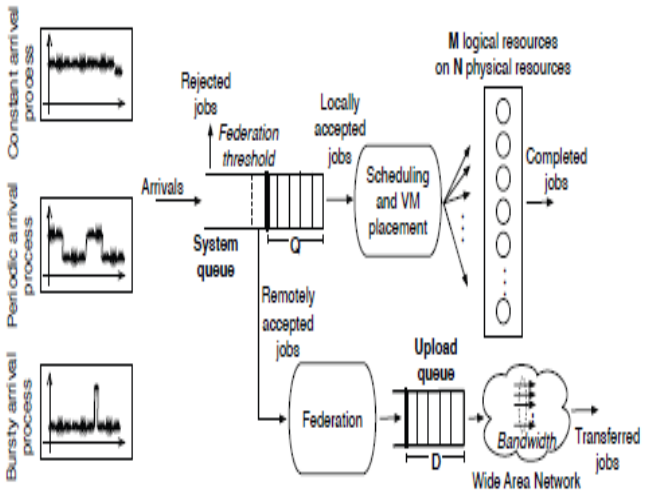

Fig: An Iaas Cloud System with Federation

An IaaS cloud system composed of *N* physical resources (see Fig. 1). Job requests (in terms of VM instantiation requests) are enqueued in the *system queue*. Such a queue has a finite size *Q*, once its further requests are rejected. The system queue is managed according to a FIFO scheduling policy. When a resource is available a job is accepted and the corresponding VM is instantiated.

Cloud federation [4] allows the system to use, in particular situations, the resources offered by other public cloud systems through a sharing and paying model. In this way, elastic capabilities can be exploited in order to respond to particular load conditions. Job requests can be redirected to other clouds by transferring the corresponding VM disk images through the network. . With respect to the federation technique the following assumptions:

• A job is redirected only if it arrives when the system queue is full.
• Federate clouds are characterized by an availability *af* .
• Federate clouds are also characterized by a quality level *qf* ($0 < qf \leq 1$) that determines the QoS reached by a request, in terms of expected service time (i.e., a VM that needs a time $T = 1/\mu$ to accomplish it works will experience an execution time $Tf = 1/(qf \cdot \mu) \geq T$).
• A redirected job is inserted in the *upload queue* waiting for the VM transfer completion.
• There is a maximum number of concurrent redirected jobs (elasticity level) i.e., the upload queue has a finite size equal to *D*.
• The network bandwidth allows to transmit up to *k* VMs in parallel.
• The time needed to transfer a VM disk image is exponentially distributed with mean $1/\eta$.

Finally, with respect to the arrival process will investigate three different scenarios. In the first one (*Constant arrival process*) we assume the arrival process be a homogeneous Poisson process with rate λ. However, largescale distributed systems with thousands of users, such as cloud systems, could exhibit self-similarity/long-range dependence with respect to the arrival process. For these reasons, in order to take into account the dependencies of the job arrival rate on both the days of a week and the hours of a day, in the second scenario (*Periodic arrival process*) also choose to model the job arrival process as a Markov

Modulated Poisson Process (MMPP). In particular, refer to an *MMPP*(λh, λl, $\lambda h2l$, $\lambda l2h$), where λh and λl represent the expected arrival rate in high and low load conditions while $1/\lambda h2l$ and $1/\lambda l2h$ represent the expected duration of the two load conditions. The last scenario (*Bursty arrival process*) takes into account the presence of a burst whit fixed and short duration and it will be used in order to investigate the system resiliency. To capture the main features of a typical IaaS cloud to use of SRNs.

2.1 Modeling cloud federation

Federation with other clouds is modeled allowing tokens in place *Pqueue* to be moved, through transition *tupload*, in the upload queue represented by place *Psend*. In accordance with the assumptions made before, transition *tupload* is enabled only if the number of tokens in place *Pqueue* is greater than *Q* and the number of tokens in place *Psend* is less than *D.*

2.2 Modeling VM multiplexing

When VM multiplexing is allowed, the number of running VMs can be greater than *N*, i.e., $0 \leq P\#run \leq M$ and each PM can be loaded with more than one VM.

2.3 Understanding the model complexity

In order to respect the scalability requirement of the proposed model, to analyze its complexity. In particular, the analysis of the state space cardinality that is the parameter that mainly influences the performance of the numerical solution techniques.

3.CONCLUSION

We have extended our proposed interacting analytical model to capture important aspects including pool management, power consumption, resource assigning process and virtual machine deployment of nowadays cloud centers. The performance model can assist cloud providers to predict the expected servicing delay, task rejection probability, steadystate arrangement of server pools and power consumption. We carried out extensive numerical experiments to study the effects of various parameters such as arrival rate of supertasks, task service time, virtualization degree, super-task size and pool check rate on the task rejection probability, response time and normalized power consumption. The behavior of cloud center for given configurations has been characterized in order to facilitate the capacity planning, SLA analysis, cloud economic analysis and trade offs by cloud service providers. Using the proposed pool management model, the most appropriate arrangement of server pools and the amount of required electricity power can be identified in advance for anticipated arrival process and super-task characteristics.

Finally, we have also presented a performance model suitable for cloud computing centers with heterogeneous requests and resources using interacting stochastic models. More specifically, a user task may request different types of VMs; VMs can be varied in terms of CPU cores per virtual CPU. Also, unlike previous performance models that have been presented our final performance model can support heterogeneous PMs.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. A view of cloud computing. Commun. ACM, 53:50–58, Apr. 2010.

[2] F. Bonomi and A. Kumar. Adaptive optimal load balancing in a nonhomogeneous multiserver system with a central job scheduler. IEEE Trans. on Comput., 39:1232–1250, Oct. 1990.

[3] G. P. Cosmetatos. Some practical considerations on multi-server queues with multiple poisson arrivals. Omega, 6(5):443–448, 1978.

[4] D. Efrosinin and V. Rykov. On performance characteristics for queueing systems with heterogeneous servers. Automation and Remote Control, 69:61–75, 2008.

[5] J. Fu, W. Hao, M. Tu, B. Ma, J. Baldwin, and F.B. Bastani. Virtual services in cloud computing. In IEEE 2010 6th World Congress on Services, pages 467–472, 2010.

[6] B. Furht. Cloud Computing Fundamentals. In Handbook of Cloud Computing, pages 3–19. Springer US, 2010.

[7] D. F. Garc´ıa, J. Garc´ıa, J. Entrialgo, M. Garc´ıa, P. Valledor, R. Garc´ıa, and A. M Campos. A QoS control mechanism to provide service differentiation and overload protection to internet scalable servers. IEEE Trans. Serv. Comput., 2:3–16, Jan. 2009.