ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

A Stochastic Model to Investigate Data Center Performance and QOS in IAAS Cloud Computing Systems

Mohammed Abdul Gaffer
M.Tech Student,
Software Engineering,
Lords College of Engineering and Technology.

C Siva Jyothi
Associate Professor,
M.S.(Software Systems),
Lords College of Engineering and Technology.

ABSTRACT:

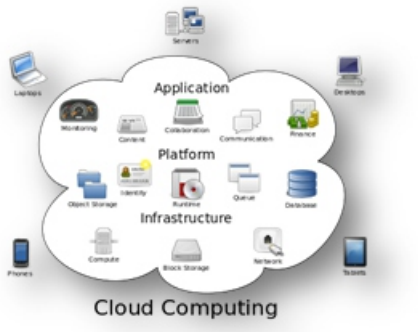
Cloud data center management is a key problem due to the numerous and heterogeneous strategies that can be applied, ranging from the VM placement to the federation with other clouds. Performance evaluation of Cloud Computing infrastructures is required to predict and quantify the cost-benefit of a strategy portfolio and the corresponding Quality of Service (QoS) experienced by users. Such analyses are not feasible by simulation or on-the-field experimentation, due to the great number of parameters that have to be investigated. In this paper, we present an analytical model, based on Stochastic Reward Nets (SRNs), that is both scalable to model systems composed of thousands of resources and flexible to represent different policies and cloud-specific strategies. Several performance metrics are defined and evaluated to analyze the behavior of a Cloud data center: utilization, availability, waiting time, and responsiveness. A resiliency analysis is also provided to take into account load bursts. Finally, a general approach is presented that, starting from the concept of system capacity, can help system managers to opportunely set the data center parameters under different working conditions.

KEYWORDS: Performance evaluation, Cost-benefit, Quality of service, Cloud-specific, Resiliency analysis.

INTRODUCTION

What is cloud computing?

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. These services typically provide access to advanced software applications and high-end networks of server computers.



Structure of cloud computing
How Cloud Computing Works?

The goal of cloud computing is to apply traditional supercomputing, or high-performance computing power, normally used by military and research facilities, to perform tens of trillions of computations per second, in consumer-oriented applications such as financial portfolios, to deliver personalized information, to provide data storage or to power large, immersive computer games.The cloud computing uses networks of large groups of servers typically running low-cost consumer PC technology with specialized connections to spread data-processing chores across them. This shared IT infrastructure contains large pools of systems that are linked together. Often, virtualization techniques are used to maximize the power of cloud computing.

Characteristics and Services Models:

The salient characteristics of cloud computing based on the definitions provided by the National Institute of Standards and Terminology (NIST) are outlined below:

•On-demand self-service: A consumer can unilaterally provision computing capabilities, such as server time and network storage, as needed automatically without requiring human interaction with each service's provider.

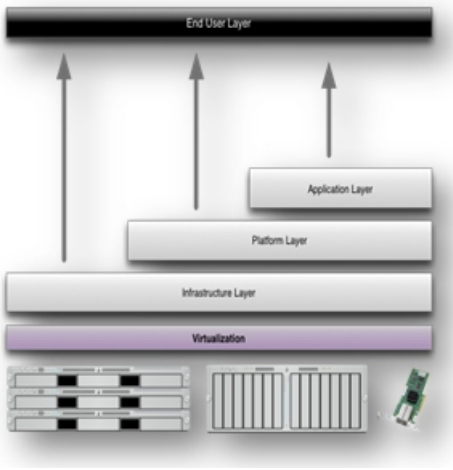
ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

•Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
•Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. There is a sense of location-independence in that the customer generally has no control or knowledge over the exact location of the provided resources but may be able to specify location at a higher level of abstraction (e.g., country, state, or data center). Examples of resources include storage, processing, memory, network bandwidth, and virtual machines.
•Rapid elasticity: Capabilities can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.
•Measured service: Cloud systems automatically control and optimize resource use by leveraging a metering capability at some level of abstraction appropriate to the type of service (e.g., storage, processing, bandwidth, and active user accounts). Resource usage can be managed, controlled, and reported providing transparency for both the provider and consumer of the utilized service.

Characteristics of cloud computing Services Models:

Cloud Computing comprises three different service models, namely Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). The three service models or layer are completed by an end user layer that encapsulates the end user perspective on cloud services.

The model is shown in figure below. If a cloud user accesses services on the infrastructure layer, for instance, she can run her own applications on the resources of a cloud infrastructure and remain responsible for the support, maintenance, and security of these applications herself. If she accesses a service on the application layer, these tasks are normally taken care of by the cloud service provider.



Structure of service models

Benefits of cloud computing:

1.Achieve economies of scale – increase volume output or productivity with fewer people. Your cost per unit, project or product plummets.
2.Reduce spending on technology infrastructure. Maintain easy access to your information with minimal upfront spending. Pay as you go (weekly, quarterly or yearly), based on demand.
3.Globalize your workforce on the cheap. People worldwide can access the cloud, provided they have an Internet connection.
4.Streamline processes. Get more work done in less time with less people.
5.Reduce capital costs. There's no need to spend big money on hardware, software or licensing fees.
6.Improve accessibility. You have access anytime, anywhere, making your life so much easier!
7.Monitor projects more effectively. Stay within budget and ahead of completion cycle times.
8.Less personnel training is needed. It takes fewer people to do more work on a cloud, with a minimal learning curve on hardware and software issues.
9.Minimize licensing new software. Stretch and grow without the need to buy expensive software licenses or programs.
10.Improve flexibility. You can change direction without serious "people" or "financial" issues at stake.

Advantages:

1.Price: Pay for only the resources used.
2.Security: Cloud instances are isolated in the network from other instances for improved security.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

3.Performance: Instances can be added instantly for improved performance. Clients have access to the total resources of the Cloud's core hardware.
4.Scalability: Auto-deploy cloud instances when needed.
5.Uptime: Uses multiple servers for maximum redundancies. In case of server failure, instances can be automatically created on another server.
6.Control: Able to login from any location. Server snapshot and a software library lets you deploy custom instances.
7.Traffic: Deals with spike in traffic with quick deployment of additional instances to handle the load.

PROBLEM STATEMENT:

In order to integrate business requirements and application level needs, in terms of Quality of Service (QoS), cloud service provisioning is regulated by Service Level Agreements (SLAs): contracts between clients and providers that express the price for a service, the QoS levels required during the service provisioning, and the penalties associated with the SLA violations. In such a context, performance evaluation plays a key role allowing system managers to evaluate the effects of different resource management strategies on the data center functioning and to predict the corresponding costs/benefits.Cloud systems differ from traditional distributed systems. First of all, they are characterized by a very large number of resources that can span different administrative domains. Moreover, the high level of resource abstraction allows to implement particular resource management techniques such as VM multiplexing or VM live migration that, even if transparent to final users, have to be considered in the design of performance models in order to accurately understand the system behavior. Finally, different clouds, belonging to the same or to different organizations, can dynamically join each other to achieve a common goal, usually represented by the optimization of resources utilization. This mechanism, referred to as cloud federation, allows to provide and release resources on demand thus providing elastic capabilities to the whole infrastructure.

DRAW BACKS:

•On-the-field experiments are mainly focused on the offered QoS, they are based on a black box approach that makes difficult to correlate obtained data to the internal resource management strategies implemented by the system provider.

•Simulation does not allow to conduct comprehensive analyses of the system performance due to the great number of parameters that have to be investigated.

PROBLEM DEFINITION:

In this paper, we present a stochastic model, based on Stochastic Reward Nets (SRNs), that exhibits the above mentioned features allowing to capture the key concepts of an IaaS cloud system. The proposed model is scalable enough to represent systems composed of thousands of resources and it makes possible to represent both physical and virtual resources exploiting cloud specific concepts such as the infrastructure elasticity. With respect to the existing literature, the innovative aspect of the present work is that a generic and comprehensive view of a cloud system is presented. Low level details, such as VM multiplexing, are easily integrated with cloud based actions such as federation, allowing to investigate different mixed strategies. An exhaustive set of performance metrics are defined regarding both the system provider (e.g., utilization) and the final users (e.g., responsiveness).

ADVANTAGES:

To provide a fair comparison among different resource management strategies, also taking into account the system elasticity, a performance evaluation approach is described. Such an approach, based on the concept of system capacity, presents a holistic view of a cloud system and it allows system managers to study the better solution with respect to an established goal and to opportunely set the system parameters.

IMPLEMENTATION

1.System Queuing:

Job requests (in terms of VM instantiation requests) are en-queued in the system queue. Such a queue has a finite size Q, once its limit is reached further requests are rejected. The system queue is managed according to a FIFO scheduling policy.

2.Scheduling Module:

When a resource is available a job is accepted and the corresponding VM is instantiated. We assume that the instantiation time is negligible and that the service time (i.e., the time needed to execute a job) is exponentially distributed with mean$1/\mu$.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

3.VM Placement:

According to the VM multiplexing technique the cloud system can provide a number M of logical resources greater than N. In this case, multiple VMs can be allocated in the same physical machine (PM), e.g., a core in a multicore architecture. Multiple VMs sharing the same PM can incur in a reduction of the performance mainly due to I/O interference between VMs.

4.Federation Module:

Cloud federation 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.

5.Arrival Process:

Finally, we respect to the arrival process we 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, large scale distributed systems with thousands of users, such as cloud systems, could exhibit self-similarity/long-range dependence with respect to the arrival process. 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.

CONCLUSION:

In this paper, we have presented a stochastic model to evaluate the performance of an IaaS cloud system. Several performance metrics have been defined, such as availability, utilization, and responsiveness, allowing to investigate the impact of different strategies on both provider and user point-of-views. In a market-oriented area, such as the Cloud Computing, an accurate evaluation of these parameters is required in order to quantify the offered QoS and opportunely manage SLAs. Future works will include the analysis of autonomic techniques able to change on-thefly the system configuration in order to react to a change on the working conditions. We will also extend the model in order to represent PaaS and SaaS Cloud systems and to integrate the mechanisms needed to capture VM migration and data center consolidation aspects that cover a crucial role in energy saving policies.

REFERENCES:

[1] R. Buyya et al., "Cloud computing and emerging it platforms: Vision, hype, and reality for delivering computing as the 5th utility," Future Gener. Comput. Syst., vol. 25, pp. 599–616, June 2009.

[2] X. Meng et al., "Efficient resource provisioning in compute clouds via vm multiplexing," in Proceedings of the 7th international conference on Autonomic computing, ser. ICAC '10. New York, NY, USA: ACM, 2010, pp. 11–20.

[3] H. Liu et al., "Live virtual machine migration via asynchronous replication and state synchronization," Parallel and Distributed Systems, IEEE Transactions on, vol. 22, no. 12, pp. 1986 –1999, dec. 2011.

[4] B. Rochwerger et al., "Reservoir - when one cloud is not enough," Computer, vol. 44, no. 3, pp. 44 –51, march 2011.

[5] R. Buyya, R. Ranjan, and R. Calheiros, "Modeling and simulation of scalable cloud computing environments and the cloudsim toolkit: Challenges and opportunities," in High Performance Computing Simulation, 2009. HPCS '09. International Conference on, june 2009, pp. 1 –11.

[6] A. Iosup, N. Yigitbasi, and D. Epema, "On the performance variability of production cloud services," in Cluster, Cloud and Grid Computing (CCGrid), 2011 11th IEEE/ACM International Symposium on, may 2011, pp. 104 –113.

[7] V. Stantchev, "Performance evaluation of cloud computing offerings," in Advanced Engineering Computing and Applications in Sciences, 2009. ADVCOMP '09. Third International Conference on, oct. 2009, pp. 187 –192.

[8] S. Ostermann et al., "A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing," in Cloud Computing, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer Berlin Heidelberg, 2010, vol. 34, ch. 9, pp. 115–131.

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

[9] H. Khazaei, J. Misic, and V. Misic, "Performance analysis of cloud computing centers using m/g/m/m+r queuing systems," Parallel and Distributed Systems, IEEE Transactions on, vol. 23, no. 5, pp. 936 –943, may 2012.

[10] R. Ghosh, K. Trivedi, V. Naik, and D. S. Kim, "End-to-end performability analysis for infrastructure-as-a-service cloud: An interacting stochastic models approach," in Dependable Computing (PRDC), 2010 IEEE 16th Pacific Rim International Symposium on, dec. 2010, pp. 125 –132.

[11] G. Ciardo et al., "Automated generation and analysis of Markov reward models using stochastic reward nets." IMA Volumes in Mathematics and its Applications: Linear Algebra, Markov Chains, and Queueing Models, vol. 48, pp. 145–191, 1993.

[12] D. Gupta, L. Cherkasova, R. Gardner, and A. Vahdat, "Enforcing performance isolation across virtual machines in xen," in Proceedings of the ACM/IFIP/USENIX International Conference on Middleware, New York, NY, USA: Springer-Verlag New York, Inc., 2006, pp. 342–362.

[13] M. Armbrust et al., "A view of cloud computing," Commun. ACM, vol. 53, pp. 50–58, Apr. 2010.

[14] J. N. Matthews et al., "Quantifying the performance isolation properties of virtualization systems," in Proceedings of the 2007 workshop on Experimental computer science, ser. ExpCS '07. New York, NY, USA: ACM, 2007.

[15] M. Mishra and A. Sahoo, "On theory of vm placement: Anomalies in existing methodologies and their mitigation using a novel vector based approach," in Cloud Computing (CLOUD), 2011 IEEE International Conference on, july 2011, pp. 275 –282.

[16] A. V. Do et al., "Profiling applications for virtual machine placement in clouds," in Cloud Computing (CLOUD), 2011 IEEE International Conference on, july 2011, pp. 660 –667.

[17] A. Verma et al., "Server workload analysis for power minimization using consolidation," in Proceedings of the USENIX Annual technical conference, Berkeley, CA, USA, 2009, pp. 28–28.

[18] G. Balbo et al., Modelling with Generalized Stochastic Petri Nets. John Wiley and Sons, 1995.

[19] R. Sahner, K. S. Trivedi, and A. Puliafito, Performance and reliability analysis of computer systems: an example based approach using the SHARPE software package. Kluwer Academic Publishers, 1995.

[20] J. K. Muppala, K. S. Trivedi, and S. P. Woolet, "On modeling performance of real-time systems in the presence of failures." Readings in Real-Time Systems, pp. 219–239, 1993.

[21] A. Puliafito, S. Riccobene, and M. Scarpa, "Evaluation of performability parameters in client-server environments." The Computer Journal, vol. 39, no. 8, pp. 647–662, 1996.

[22] R. Ghosh, F. Longo, V. Naik, and K. Trivedi, "Quantifying resiliency of iaas cloud," in Reliable Distributed Systems, 29th IEEE Symposium on, 2010, pp. 343 –347.

[23] "Spnp manual," www.ee.duke.edu/chirel/MANUAL/SPNPv6-manual.pdf.

[24] G. Ciardo, J. Muppala, and K. S. Trivedi, "SPNP: Stochastic Petri Net Package," in 3rd International Workshop on Petri-nets and Performance Models Los Alamitos, California, 1989, pp. 142–151.