International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

An Arbitrary Model to Explore Data Center Performance in IaaS Cloud Computing Systems

Mohammad Neelofar [1]; Ch Lavanya Susanna [2]& Sayeed Yasin [3]

[1] M.Tech (CSE), Nimra College of Engineering and Technology, A.P., India.
[2] Asst Professor, Dept. of Computer Science & Engineering, Nimra College of Engineering and Technology, A.P., India.
[3] Head of the Department, Dept. of Computer Science & Engineering, Nimra College of Engineering and Technology, A.P., India.

Abstract *—*
Stochastic model to evaluate the performance of an IaaS cloud system. 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 — Cloud computing; stochastic reward nets; cloud-oriented performance metrics; resiliency; responsiveness

I. Introduction

In a market-oriented area, such as the cloud computing, an accurate evaluation of these parameters is required to quantify the offered QoS and opportunely manage SLAs.Cloud computing is a promising technology able to strongly modify the way computing and storage resources will be accessed in the near future [1].

Through the provision of on-demand access to virtual resources available on the Internet, cloud systems offer services at three different levels: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). In particular, IaaS clouds provide users with computational resources in the form of virtual machine (VM) instances deployed in the provider data center, while PaaS and SaaS clouds offer services in terms of specific solution stacks and application software suites, respectively. 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 us to implement particular resource management techniques such as M multiplexing [2] or VM live migration [3] that, even if transparent to final users, have to be considered in the design of performance models to accurately understand the system behaviour. 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 [4], allows us to provide

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

and release resources on demand, thus providing elastic capabilities to the whole infrastructure.

For these reasons, typical performance evaluation approaches such as simulation or on-the-field measurements cannot be easily adopted. Simulation [5], [6] does not allow us to conduct comprehensive analyses of the system performance due to the great number of parameters that have to be investigated. On-the-field experiments [7], [8] 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. On the contrary, analytical techniques [9], [10] represent a good candidate, thanks to the limited solution cost of their associated models. However, to accurately represent a cloud system, an analytical model has to be Scalable. To deal with very large systems composed of hundreds or thousands of resources Flexible. Allowing us to easily implement different strategies and policies and to represent different working conditions. 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 us to investigate different mixed strategies.

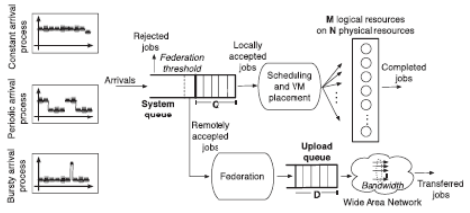


Fig. 1. An IaaS cloud system with federation.

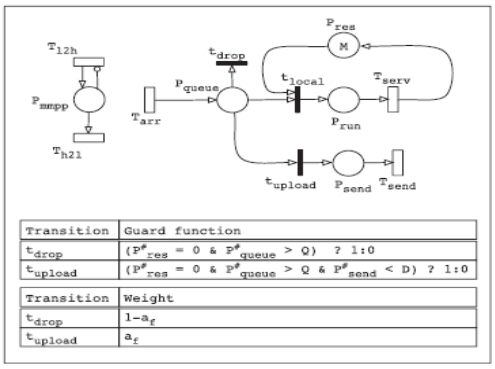
An exhaustive set of performance metrics is defined regarding both the system provider (e.g., utilization) and the final users (e.g., responsiveness). Moreover, different working conditions are investigated and a resiliency analysis is provided to take into account the effects of load bursts. Finally, 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.

II .PROBLEM STATEMENT

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. In this paper, we have presented a stochastic model to evaluate the performance of an IaaS cloud system.

III .RELATED WORK

Finally, with 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. 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 [7]. For these reasons, to take into account the dependences 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),



Transition	Guard function
t_{drop}	$(P^{\#}_{res} = 0 \ \& \ P^{\#}_{queue} > Q) \ ? \ 1:0$
t_{upload}	$(P^{\#}_{res} = 0 \ \& \ P^{\#}_{queue} > Q \ \& \ P^{\#}_{send} < D) \ ? \ 1:0$

Transition	Weight
t_{drop}	$1-a_f$
t_{upload}	a_f

Fig. 2. The proposed SRN cloud performance model.

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

We also choose to model the job arrival process as a Markov Modulated Poisson Process (MMPP). To capture the main features of a typical IaaS cloud, we make use of SRNs [11]. SRNs are an extension of generalized stochastic Petri Nets (GSPNs) [14] that allow us to associate reward rates with the marking (i.e., the distribution of tokens in the various places) [9]. In the reminder of the paper, we will use the notation P# to refer to the number of token in place P. Moreover, in the function definitions, we adopt a C-like syntax using the ternary operator (? :) instead of the if _ else construct. We give a formal overview of the SRN notation in the Appendix A of the supplementary file available online. The proposed SRN cloud performance model is depicted in Fig. 2. Transition Tarr models the arrival process. If the constant arrival process is taken into account, we can characterize transition Tarr with an exponentially distributed firing time with mean.

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. Moreover, to take into account the federated cloud availability, concurrent enabled transitions tupload and tdrop are managed by setting their weights with the values af and $1\text{-}a_f$, respectively. To respect the scalability requirement of the proposed model, we need to analyze its complexity. In particular, we are interested in the analysis of the state-space cardinality that is the parameter that mainly influences the performance of the numerical solution techniques. The state space S of the model is given by the set of all its tangible markings [8]. Considering the SRN of Fig. 2 and the corresponding guard functions, we can make some considerations on the token distributions. SRNs allow us to define reward functions that can be associated to a particular state of the model to evaluate the performance level reached by the system during the sojourn in that state [11]. The expected data center utilization U can be computed as the ratio between the number of physical resources used at steady state and the total number N of physical resources. It is the steady-state probability R that the system is able to accept a request within a given time deadline _. The computation of such a parameter requires the knowledge of the waiting time cumulative distribution function (CDF). Through a transient solution of the cloud performance model of Fig. 2, it is possible to investigate the trend over time of some performance metrics. Such an analysis is straightforward to assess the resiliency of the cloud infrastructure, in particular when the load is characterized by bursts. In fact, even if the infrastructure is optimally sized with respect to the expected load, during a load burst, users can experience a degradation of the perceived QoS with corresponding violations of SLAs. For this reason, it is needed to predict the effects of a particular load condition to study the ability of the system to react to an overload situation.

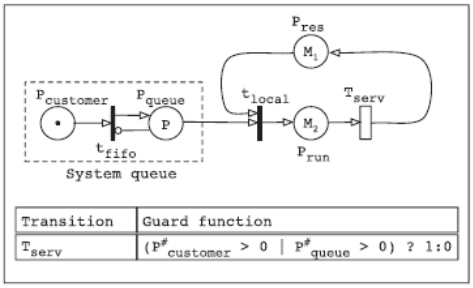


Fig. 3. The SRN tagged customer model used in the responsiveness evaluation.

To this end, it is possible to apply the tagged customer technique [10], [11] by modifying the SRN model to isolate the behavior of a single user request u and to observe its movements through the system. In the tagged customer model shown in Fig. 3, the system queue is modelled through two places. Place Pcustomer contains a single token that represents the arrival of request u. The P tokens initially present in place Pqueue represent the number of requests still waiting in the queue when u arrives, while the M_1 and M_2 tokens initially present in places Pres and Prun represent the corresponding system status. The FIFO policy is modeled through transition tfifo that is enabled only when place Pqueue is empty. Transitions t_{local} and T_{serv} behave as in the model of Fig. 2.

The bursty arrival process is modeled by opportunely changing the exponentially distributed firing time of the transition T_{arr} in the cloud performance model through the adoption of the technique described in [12], [13]. First of all, we can identify three temporal phases: In each phase, the model is solved in transitory by setting the firing rate of Tarr with the corresponding mean value. During the resiliency analysis, we are interested in the quantitative evaluation of the performance degradation experienced by the system during a load burst. To this end, we propose some

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 11, November 2015
Available at http://internationaljournalofresearch.org

temporal indices (see Fig. 4) able to capture the performance degradation trend. Such indices can be applied to both the Availability and Instant service probability metrics.

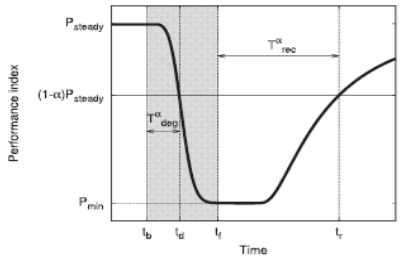


Fig. 4. Graphical representation of the quantitative resiliency metrics: the dashed area corresponds to the burst duration.

As can be observed, the Federation strategy increases the 0.05-degradation time of 25 minutes with respect to the Base configuration and of a value ranging from 3 to 5 minutes with respect to the other strategies. Then, using such a strategy the system is able to maintain its performance for a longer time, thus reducing the QoS degradation. Similar consideration can be made with respect to the 0.05-recovery time; in this case, the Federation strategy produces the shortest time interval. Such a trend is also confirmed by the value of MPL (%) that, in the case of the Federation strategy, is reduced by half with respect to the Base configuration. On the other hand, observing the data referred to the instant service probability, we can quantify the improvements obtained using the Multiplexing strategy. From a user perspective, we will investigate the impact of the system capacity on the total service time (i.e., the sum of the waiting time and the service time) and on the system responsiveness. In fact, taking as reference the Base configuration, we can observe that the 0.05-degradation time is increased by 20 minutes, while the 0.05-recovery time is reduced by more than 100 minutes. Moreover, we can argue that the Federation strategy does not influence such a system performance index and that the Queuing strategy gives rise to the worst results, in particular with respect to the 0.05-recovery time that reaches a value near to 7 hours (very high if compared with burst duration of about 1 hour). Finally, the high values of MPL (%) confirm that such a performance metric is strongly influenced by a burst load.

IV Conclusion

In a market-oriented area, such as the cloud computing, an accurate evaluation of these parameters is required to quantify the offered QoS and opportunely manage SLAs 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 us to investigate the impact of different strategies on both provider and user point of views. Future works will include the analysis of autonomic techniques able to change on-the-fly the system configuration to react to a change on the working conditions. We will also extend the model 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 Fifth Utility," Future Generation Computer System, vol. 25, pp. 599-616, June 2009.

[2] X. Meng et al., "Efficient Resource Provisioning in Compute Clouds via VM Multiplexing," Proc. Seventh Int'l Conf. Autonomic Computing (ICAC '10), pp. 11-20, 2010.

[3] H. Liu et al., "Live Virtual Machine Migration via Asynchronous Replication and State Synchronization," IEEE Trans. Parallel and Distributed Systems, 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, Mar. 2011. [5] R. Buyya, R. Ranjan, and R. Calheiros, "Modeling and Simulation of Scalable Cloud Computing Environments and the Cloudsim Toolkit: Challenges and Opportunities," Proc. Int'l Conf. High Performance Computing Simulation (HPCS '09), pp. 1-11, June 2009.

[5] M. Mishra and A. Sahoo, "On Theory of VM Placement: Anomalies in Existing Methodologies and Their Mitigation Using a Novel Vector Based Approach," Proc. IEEE Fourth Int'l Conf. Cloud Computing (CLOUD '11), pp. 275-282, July 2011. [16] A.V. Do et al., "Profiling Applications for Virtual Machine Placement in Clouds," Proc. IEEE Int'l Conf. Cloud Computing (CLOUD '11), pp. 660-667, July 2011.

[6] A. Iosup, N. Yigitbasi, and D. Epema, "On the Performance Variability of Production Cloud Services," Proc. IEEE/ACM 11th Int'l Symp. Cluster, Cloud and Grid Computing (CCGrid), pp. 104- 113, May 2011.

[7] V. Stantchev, "Performance Evaluation of Cloud Computing Offerings," Proc. Third Int'l Conf. Advanced Eng. Computing and Applications in Sciences (ADVCOMP '09), pp. 187-192, Oct. 2009.

[8] S. Ostermann et al., "A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing," Proc. Int'l Conf. Cloud Computing, LNCS vol. 34, pp. 115-131, Springer, Heidelberg, 2010.

[9] H. Khazaei, J. Misic, and V. 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.

[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," Proc. IEEE 16th Pacific Rim Int'l Symp. Dependable Computing (PRDC), pp. 125-132, Dec. 2010.

[11] G. Ciardo et al., "Automated Generation and Analysis of Markov Reward Models Using Stochastic Reward Nets," Linear Algebra, Markov Chains, and Queuing Models, vol. 48, pp. 145-191, Springer, 1993.

[12] D. Gupta, L. Cherkasova, R. Gardner, and A. Vahdat, "Enforcing Performance Isolation across Virtual Machines in Xen," Proc. ACM/IFIP/USENIX Int'l Conf. Middleware, pp. 342-362, 2006.

[13] M. Armbrust et al., "A View of Cloud Computing," Comm. ACM, vol. 53, pp. 50-58, Apr. 2010.

[14] J.N. Matthews et al., "Quantifying the Performance Isolation Properties of Virtualization Systems," Proc. Workshop Experimental Computer Science (ExpCS '07), 2007.