A Stochastic Model To Investigate Data Center Performance and Qos in Iaas Cloud Computing Systems

Dario Bruneo, Ph.D. in Advanced
Technologies for Information Engineering at the
University of Messina (Italy) in 2005
IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS 2013.

1. *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.

2. INTRODUCTION

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.For these reasons, typical performance evaluation approaches such as simulation or on-the-field measurements can not be easily adopted. Simulation does not allow conduct comprehensive analyses of the system performance due to the great number of parameters that have to be investigated. 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. On the contrary, analytical techniques represent a good candidate thanks to the limited solution cost of their associated models. However, accurately represent a cloud system an analytical model has to be:

• Scalable. In order to deal with very large systems composed of hundreds or thousands of resources.
• Flexible. Allowing to easily implement different strategies and policies and to represent different working conditions.

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 investigating 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). 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.

3 RELATED WORKS

3.1. Markovian Workload Characterization for QoS Prediction in the Cloud

Resource allocation in the cloud is usually driven by performance predictions, such as estimates of the future incoming load to the servers or of the quality-of-service(QoS) offered by applications to end users. In this context, characterizing web workload fluctuations in an accurate way is fundamental to understand how to provision cloud resources under time-varying traffic intensities. we investigate the Markovian Arrival Processes (MAP) and the related MAP/MAP/1queuing model as a tool for performance prediction of servers deployed in the cloud. MAPs are a special class of Marko models used as a compact description of the time-varying characteristics of workloads. In addition, MAPs can fit heavy-tail distributions that are common in HTTP traffic, and can be easily integrated within analytical queuing models to efficiently predict system performance without simulating. By comparison with traced riven simulation, we observe that existing techniques for MAP parameterization from HTTP log files often lead to inaccurate performance predictions. We then define a maximum likelihood method for fitting MAP parameters based on data commonly available in Apache log files, anda new technique to cope with batch arrivals, which are notoriously difficult to model accurately. Numerical experiments demonstrate the accuracy of our approach for performance prediction of web systems.

3.2Performance Analysis of Cloud Computing Center Using M/G/m/m+r Queuing Systems:

Successfuldevelopment of cloud computing paradigm necessitates accurate Performance evaluation of cloud data centers. As exact modeling of cloud centers is not feasible due to the nature of cloudcenters and diversity of user requests, we describe a novel approximate analytical model forperformance evaluation of cloud server farms and solve it to obtain accurate estimation of the complete probability distribution of the request response time and other important performance indicators. The model allows cloud operators to determine the relationship between the number of servers and input buffer size, on one side, and the performance indicators such as mean number of tasks in the system, blocking probability, and probability that a task will obtain immediate service, on the other.

3.3 Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing:

Cloud computing is an emerging commercial infrastructure paradigm that promises to eliminate the need for maintaining expensive computing facilities by companies and institutes alike. Through the use of virtualization and resource time sharing, clouds serve with a single set of physical resources a large user base with different needs. Thus, clouds have the potential to provide to their owners the benefits of an economy of scale and, at the same time, become an alternative for scientists to clusters, grids, and parallel production environments. However, the current commercial clouds have been built to support web and small database workloads, which are very different from typical scientific computing workloads. Moreover, the use of virtualization and resource time sharing may introduce significant performance penalties for the demanding scientific computing workloads. In this work, we analyze the performance of cloud computing services for scientific computing workloads. We quantify the presence in real scientific computing workloads of Many-Task Computing (MTC) users, that is, of users who employ loosely coupled applications comprising many tasks to achieve their scientific goals. Then, we perform an empirical evaluation of the performance of four commercial cloud computing services including Amazon EC2, which is currently the largest commercial cloud. Last, we compare through trace-based simulation the performance characteristics and cost models of clouds and other scientific computing platforms, for general and MTC-based scientific computing workloads. Our results indicate that the current clouds need an order of magnitude in performance improvement to be useful to the scientific community, and show which improvements should be considered first to address this discrepancy between offer and demand

4 EXISTING SYSTEM

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.

Disadvantages of Existing System

- 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.

5 PROPOSED SYSTEM

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). 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.

Advantages of Proposed System

- 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.

Architecture Diagram

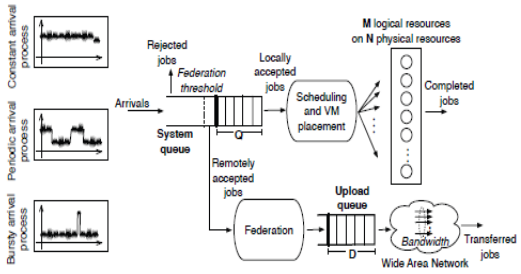

Fig. 1. An IaaS cloud system with federation.

6 DESIGN GOALS

The aim of the project is to 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 system 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.

7.CONCLUSION

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-the-fly 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.

8.REFERENCE:

[1] Dario Bruneo, *Member,* IEEE-" A Stochastic Model to Investigate Data Center Performance and QoS in IaaS Cloud Computing Systems"- IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS 2013.

[2]. 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.

[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]. 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.

[5]. 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.

[6]. Berlin Heidelberg, 2010, vol. 34, ch. 9, pp. 115–131. 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.

[7]. 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.

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

[9] .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.

[10]. 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.