Prediction of Cloud Application's Performance using SMTQA Tool

P.Ganesh[#1], D EvangelinGeetha[#2], TV Suresh Kumar[#3]

[#1]*Associate Professor, Dept. of MCA, BMSIT, India*
[#2]*Associate Professor, Dept. of Computer Applications, MSRIT, India*
[#3]*Professor&HOD, Dept. of Computer Applications, MSRIT, India*

***Abstract**—Performance of cloud applications is critical for its user acceptance. Resource management and scalability play important role in cloud performance. As a result, cloud environments fundamentally aim for resource consolidation and management. Also, it is challenging for the cloud service providers to allocate the cloud resources dynamically and efficiently. Through proper management of cloud resources, the scalability issue can be mitigated significantly. Given the significance of resource management in assessing the cloud application performance, we focus on evaluatingthe cloud performance considering resource utilization aspects of a cloud application. In this paper, we attempt to identifythe key actors in cloud environment with respect to resource management and design UML model for it. Also, we predict the performance of sample cloud application through SMTQA simulation.*

Keywords—*Service Level Agreement, Service Level Objectives, Unified Modelling Language, Software Performance Engineering*

I. INTRODUCTION

Cloud environment is blend of parallel and distributed systems that comprises of a collection of interconnected and virtualized resources. Cloud resources are dynamically provisioned on-demand and provided as one or more integrated computing resources as per service-level agreements, among the service providers and consumers. As, the consumers of cloud can access its services/ applications and associated data from anywhere at any time [5],it is difficult for the cloud service providers to distribute the cloud resources dynamically and efficiently [6].

Resources can be either physical or logical. Physical resources include Computer, Processor, Disk, Database, Network, Bandwidth and the logical resources include Execution, Monitoring and Application etc. Resource Management can be defined as the process of allocating resources of cloud such as storage, computing and networking to a bundle of applications. Cloud environments primarilytarget resource consolidation and management[2].

Resource management in cloud environment is a challenging task for the reasons of data centre scaling, diversified resource types, inter dependencies of resources, unpredictable load and objectives of cloud system actors [1]. Resource managementis expected to meet the performance objectives of applications, infrastructure providers and users of cloud services.It also influences scalability of application.Hence, resource management and scalability are two important issues identified as critical for the performance assessment of a cloud application [9].

Performance generally refers to system responsiveness, either the time required to respond to specific events, or number of events processed in a given time interval. The performance objectives are specified as: response time, throughput, and constraints on resource usage. These objectives are used for the performance assessment.Performance responsiveness and scalability are make-or-break qualities for software [15].

Predicting the performance of software systems enables developing more responsive and performance oriented systems. Software performance engineering (SPE) is used to predict the performance of the software system.SPE is a methodology to predict performance of software systems early in the life cycle [11].SPE continues through the other stages of the life cycle as well to predict and manage the performance of the system in consideration. Also, SPE monitors and reports actual performance against predictions. Thus SPEis considered important for software quality assessment. In this paper, using SPEapproach we predict the performance of a cloud application considering its resource utilization aspects.

In section II, we, examine the available literature on SPE and cloud performance; NIST reference architecture on cloud in section III, modelling using UML in section IV, conduct simulation using SMTQA tool in section V and present the simulation results in section VI.

II. LITERATURE SURVEY

Survey of the literature available on SPE, performance of cloud applications especially with respect toresource management is carried out to understand the status of performance aspects of cloud applications and related concerns/challenges. As reported in [16],cloud is not performing as satisfactorily as expected.The descriptions of software dynamics, referred as software runtime behavior, are required to analyze performance. Performance related problems can originate from any of a number of interacting application and infrastructure components.

It is difficult to model performance of cloud system and analysethe same due to the complexity of cloud computing system[18]. Jin Shao et al., in their work presented a performance model based approach to guarantee the performance of cloud applications [17]. Queuing Networks, Stochastic Petri nets, Simulation etc. are most widely used performance analysis models.Using analytical methods and simulation techniques performance models can be evaluated in order to get performance indices. In general, these indices include response time, utilization, power consumption, throughput etc.In particular to cloud applications, these indices include resource sharing, scalability, responsiveness, user satisfaction, reliability, latency, service availability, accessibility etc.

Among these performance indices of cloud applications, resource management and scalability are identified as critical for the performance assessment of a cloud applications [9].Resource management is defined as a core function required in any man-made system by Dan Marinescu [20]. According to him, resource management affects the three basic criteria of system evaluation viz. performance, functionality and cost.

SPE supports the construction of software and systems by meeting performance objectives early in the SDLC. SPE approach, initially proposed by Connie U Smith, was to integrate software performance analysis into the software engineering process [11].Recently, the performance analysis of distributed systems during feasibility study and during the early stages of software development was presented by EvangelinGeetha D et al. [21]. Also, a process model, called Hybrid Performance Prediction (HP^3) model, was proposed by these authors to model and evaluate distributed systems with the goal of assessing performance of software system during feasibility study. In addition, an execution environment based on dynamic workload to achieve the defined performance goal was determined by them [22].

Through the literature available, it is observed that performance assessment for cloud applications during the early stages of development is not carried out [15]. Also most of the software performance prediction methodologies are concentrated on modeling of transformation techniques and solving the performance models.

III. CLOUD REFERENCE ARCHITECTURE

The National Institute of Standards and Technology(NIST) view of Cloud computing definition and architecture are widely accepted towards understanding and developing cloud technologies and cloud services. As per NIST, thereare three cloud services namely IaaS, PaaS, SaaS; four cloud deployment models namely Public Cloud, Private Cloud, Hybrid Cloud and Community Cloud. As public cloud is more general and widely used, we consider it as our deployment model and all our further discussions are based on the public cloud, unless otherwise specified. The exhaustive cloud reference architecture presented in Fig.1, is as defined by NIST. There exist five major actors in this architecture.

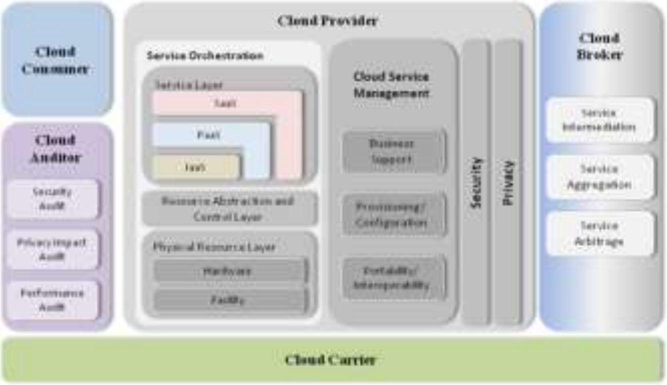

Fig 1: NIST Cloud Reference Architecture

The five major actors as described in the architecture can be further generalized for our UML modeling purposes. Brendon Jennings et al. and Armbrust M et al, proposes, for simplicity, such three important and general actors [1][10] and we too prefer consider the same for our discussion in this paper. These identified actors include: *Cloud Provider, Cloud User and End-User*. These actors are not new but are part of the NIST architecture. Based on the usage of IaaS or PaaS or SaaS services, these actors' roles can be suitably assumed as presented in Fig.2.

Thus the roles of these actors, from cloud resource management stand point, can be defined as follows:

Cloud Provider: A set of data center related hardware and software resources, are managed by Cloud Provider. In public

cloud context, it means,either IaaS or PaaS services of these resources to Cloud Users.It is accountable for distributing the resources to its users. The general SLAs include the aspects such as availability, service performance, data access, problem resolution, security, privacy etc. Also, the typical SLOswrt Cloud Provider include load balancing, fault tolerance, energy efficiency, among others.

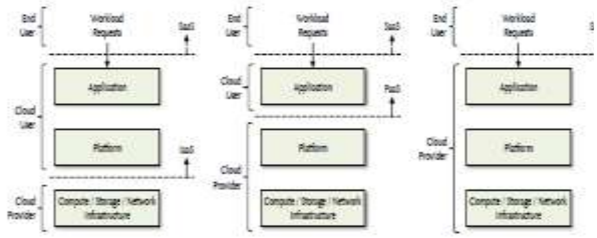

Fig 2: Roles of the actors

Cloud User: Cloud User offers its services (applications) to End Users by hosting them onpublic clouds. It is responsible for fulfilling SLAs agreed with its customers (i.e., End Users) and is more concerned in doing so to reduce its costs and optimise its gains. For achieving this, Cloud User need to ensure that the set of resources rented from the Cloud Provider scales in line with demands from its End Users.

End User: End User generates the workloads that are handled using leased cloud resources. Itusually does not play a direct role in resource management, but its application usage behaviour can influence the resource management decisions of the Cloud User and Cloud Provider.

As our concentration in this paper is on resource management and associated scalability, let us consider various resource types from cloud computing perspective. These resources can be basically of three categories: Compute resources, Network resources, Storage resources, Power resources.

Compute Resources**:** Collection of Physical Machines (PMs), each involving processors, memory, network and local I/O. PMs ironically deploy on them the virtualization layer that enables PMs host Virtual Machines (VMs) [13][14].

Network Resources: PMs must be solidly connected with a high-bandwidth network. Cloud-hosted applications are mainly influenced by communication overhead involved in data center networking technologies and networking protocols. Predicting latency and bandwidth in a data center network under the varying traffic patterns is crucial part of these resources.

Storage Resources: Storage services, range from virtual disks and database services to object stores. Each of these services has varying levels of data consistency assurances and reliability.Achieving elasticity in order for a service to dynamically scale up with an increasing number of workload is a difficult issue.

Power Resources: Data centers use significant portion of energy resources and energy costs, and hence it accounts for a substantial portion of the total cost of a data center.

IV. MODELING USING UML

We consider the cloud reference architecture as approved by NIST to design UML model. Based on the classification of the actors as highlightedin this literature and their functionalities, Fig.3 depicts the UML use case diagram. The description of the use cases is as below:

a) ***Global Scheduling of Virtual Resources****:* Deals with system-wide monitoring and control of virtualised resources, meeting cloud providers' management objectives, admission control of requests from IaaS cloud users to use virtual infrastructure deployment and placement of virtual infrastructure on physical infrastructure.

*b)****Resource demand profiling*** *– Demand Profiler:* Engages to find balance between reactivity vs proactivity in balancing resources and implement(proactive) demand profiling based ondemand patterns for virtual infrastructure and individual applications.

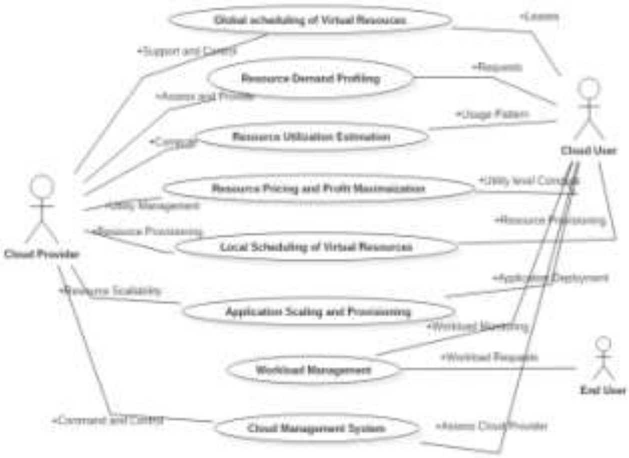

Fig 3: UML diagram

*c)****Resource utilisation estimation****:* It predicts the state of physical and virtual resources required, provides resource utilisation estimation for compute, network, storage and power resources and supports cloud monitoring and resource scheduling process.

d.) ***Resource pricing and profit maximisation****:* It concentrates on resource pricing based on the usage through dynamic pricing of cloud resources used, encouraging users to use more virtual infrastructure through dynamic pricing to increase(maximise) profit and on accurate demand profile input in order to maximise profit.

*e.) **Local scheduling of virtual resources**:* Decides on how to share access between virtual resources placed on physical resources, enable the control of VM sharing in Hypervisor / VMM and inline tuning of local scheduling with cloud management objectives.

*f.) **Application scaling and provisioning**:* Mainly comprises of placing applications on virtual infrastructure – either by cloud provider or cloud user. This is important as application demand changes, the placement and configuration of application modules tend to dynamically change – thus its success influences demand estimation for future.

*g.) **Work load management**:* Involves inmanaging cloud user controls the end user work load requests. Once accepted, it assigns the workload to one of themany instantiations of software module that manages its execution.

*h.) **Cloud management systems**:* It is used by cloud providers and cloud users to provide feedback forresource management systems enabling the critical component, SLA, being met satisfactorily.

V.SMTQA SIMULATION

SMTQA (Simulating Multi Tier Queueing Application) is a process oriented simulation tool developed for the performance evaluation of software that follows multi tier architecture [12]. It provides complete visualization of model, parameters and output. It addresses the following under distributed environment.

1. Simulation of multi tier architecture with open work load and multi classes.
2. Use Case perfomance Engineering approach
3. Simulation of server behaviours with replicas
4. dynamic load balancing
5. Generation of performance metrics and associated graphs.

As described, the cloud environment has layered architecture and hence multi-tier architecture oriented simulation is best suited for it. The SMTQA simulation tool [12], which is appropriate for such environments, is considered in our study. For simplicity, the cloud provider and cloud user are assumed to be one and the same, so as to clearly differentiate from service provider and service end user. Thus, we consider only two actors namely cloud provider and end user. Accordingly, the use cases reduce to 6 from 8.

TABLE I
Overhead Matrix for Cloud Application Performance

DEVICE	CPU	DISK	INET	DELAY
SERVICE UNITS	SEC	PHY I/O	KB	SEC
INPUT	0.00006	-	1	-
DB ACCESS	0.00005	-	-	0.025
LOCAL DB	0.0001	2	1	-
DATA SIZE	0.00005	-	1	-
SERVICE TIME	1	0.0003	0.002	1

Also, the system is further sliced as end user system, internet1, application server, internet2, data center. Suitable application usage sizes, in KB, with respect to resource scheduling, demand profiling, resource estimation, resource pricing, scaling, work load management are considered.The overhead matrix for the sample cloud application is presented in Table 1. The resources like CPU, Disk, Internet are considered.

VI. RESULTS AND DISCUSSION

We have simulated the environment using SMTQA tool. In our analysis, the simulation is carried out for 1000 requests. Performance metrics like average response time, average waiting time and average service time are obtained and presented in Table 2. Through the tool, graphs are also generated for clarity and better understanding.

Through the simulation, it is observed that, the waiting time for the given work load is high with data center and moderate with application server. Thus response time is very high in data center, moderate in application server and low in other resources.Also, service time of data center is high compared to application server. On the other hand, dropping of sessions is more with application server compared to data center.

Hence, data center and application server are identified as bottleneck resources either due to more waiting time or high probability of dropping of sessions.The graphs, as given in Fig.4 through Fig.7, provide some interesting details on the relationship among these resourcesand their impact onapplication’s performance. The averageresponse time and waiting timeof application server and data center,

TABLE II
Performance Metrics Obtained Through SMTQA for Cloud Application

	AVG. RESPONSE TIME	AVG. SERVICE TIME	AVG. WAITING TIME	PROB. OF IDLE SERVER	PROB. OF DROPPING OF SESSIONS
EU	0.001	0.001	0.000	0.989	0.000
INTERNET	0.043	0.025	0.017	0.526	0.047
ASCPU	2.532	0.648	1.884	0.010	0.918
ASDB	0.011	0.010	0.001	0.987	0.000
INT2	0.129	0.072	0.057	0.853	0.009
DCCPU	1.299	0.702	0.507	0.101	0.130
DCDB	16.357	5.312	11.044	0.107	0.429

initially increase gradually. Later, after reaching certain arrival rates, say 10 or20, they decrease steadily. This pattern can be attributed to droppingof sessions in previous resources. From the Fig.8, it is evident thathigher thearrival rate more will be the application server utilization.

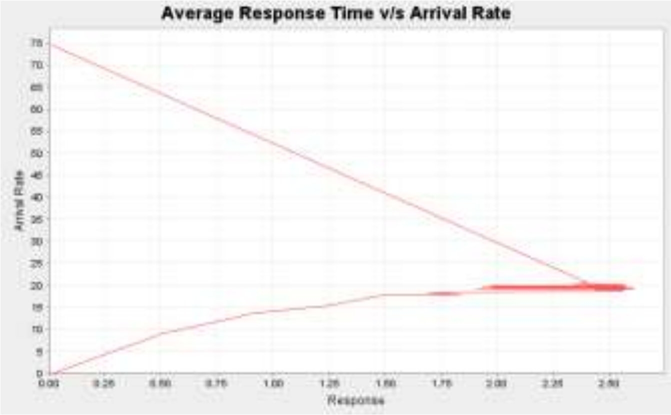

Fig 4: Average response time of Application Server

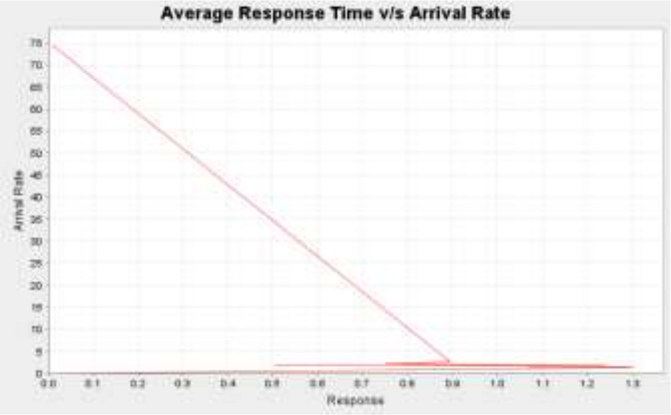

Fig 5: Average response time of Data center

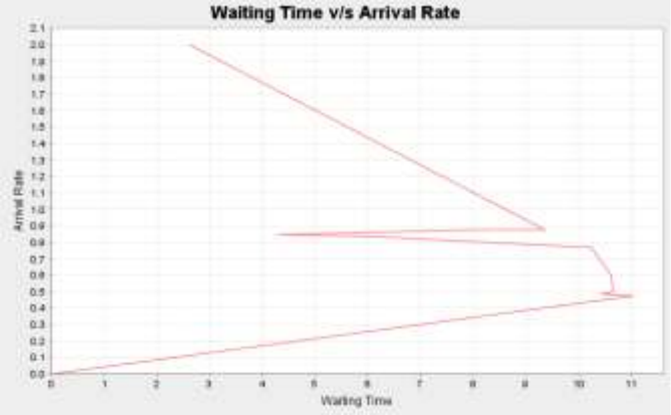

Fig 6: Average waiting time of Data center DB

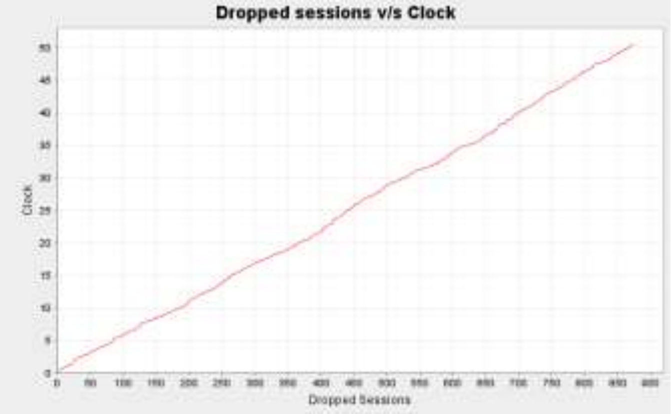

Fig 7: Sessions drop in Application Server CPU

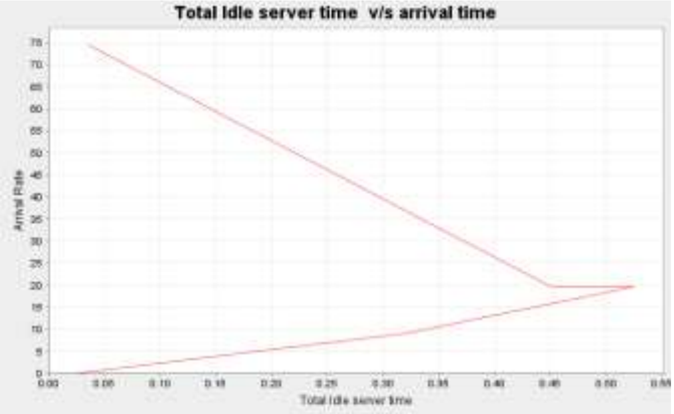

Fig 8: Idle server time of Application Server

Also, as depicted in Fig.9, the average response time of end user system, is almost uniform throughout thesimulation for all the arrival ratesand it is not affected by the performance of previous resources. This is due to the fact that end user system is the first resource being used followed by other resources.

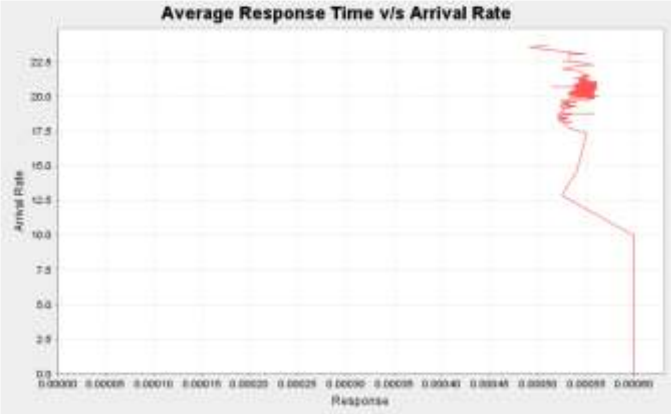

Fig 9: Average response time of End User System

VII. CONCLUSION

Resource management in cloud environment is critical and crucial aspect for its performance. Also it influences another cloud parameter, scalability.SPE methodology predicts the performance of software system early in the SDLC thereby providing scope to improve it further. We have designed UML model for a cloud application considering resource management aspect of it. Also using SMTQA tool, which is developed in line with SPE principles, we have simulated a cloud application to gather statistics to assess performance. Various performance metrics obtained through the simulation ascertain that application server and data center, which are under the control of a cloud provider, are bottle neck resources and takes considerably more response time. We further propose to enhance our work to develop a framework to predict performance of cloud applications.

REFERENCES

[1] Resource Management in Clouds: Survey and Research Challenges by Brendan Jennings and Rolf Stadler, February 2013, Springer, http://dx.doi.org/10.1007/s10922-014-9307-7

[2] Efficient Resource Management for Cloud Computing Environments, byAndrew J. Younge, Gregor von Laszewski, Lizhe Wang, Sonia Lopez-Alarcon, Warren Carithers

[3] On Resource management for cloud users: A generalised Kelly Mechanism approach, by Richard T.B. Ma, dah Ming Chiu, John C.S. Lui, Vishal Misra and Dan Rubenstein

[4] Cloud Computing: State-of-the-art and research challenges byQi Zhang, Lu Cheng, RaoufBoutaba, Springer Research Gate May 2010, DOI: 10.1007/s13174-010-0007-6

[5]Resource Management and Scheduling in Cloud Environment by Vignesh V, Sendhil Kumar KS, Jaisankar N, International Journal of Scientific and Research Publications, Volume 3, Issue 6, June 2013 1 ISSN 2250-3153

[6] Hitoshi Matsumoto, Yutaka Ezaki," Dynamic Resource Management in Cloud Environment", July 2011, FUJITSU science & Tech journal, Volume 47, No: 3, page no: 270-276.

[7]Mell P, Grance T. The NIST definition of cloud computing (draft).NIST special publication. 2011; 800(145):1–7.

[8]http://cloudpatterns.org/mechanisms/resource_
management_system

[9]P.Ganesh, D EvangelinGeetha, T V Suresh Kumar, "Impact of resource management and scalability on performance of cloud applications – A survey", International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.6, No.4, August 2016.

[10]Armbrust M., Fox A., Griffith R., Joseph A.D., Katz R., Konwinski A., Lee G., Patterson D., Rabkin A., Stoica I., Zaharia M.: A view of Cloud Computing. Communications of the ACM 53(4), 50-58 (2010).
DOI 10.1145/1721654.1721672

[11] Connie U . Smith, Performance Engineering of Software Systems, Reading, Addison-Wesley, 1990.

[12] D. EvangelinGeetha, T. V. Suresh Kumar, P. Mayank, K. Rajanikanth, 2010, "A tool for simulating multitier queuing applications", Technical Report, Department of MCA, MSRIT, TRMCA 04.

[13] Ahn, J., Kim, C., Choi, Y.r., Huh, J.: Dynamic virtual machine scheduling in clouds for architectural shared resources. In: Proc. 4th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud 2012) (2012)

[14] Govindan, S., Liu, J., Kansal, A., Sivasubramaniam, A.: Cuanta: quantifying e_ects of shared on-chip resource interference for consolidated virtual machines. In: Proc. 2nd ACM Symposium on Cloud Computing (SoCC 2011), pp. 22:1-22:14. DOI 10.1145/2038916.2038938

[15] P.Ganesh, D EvangelinGeetha, T V Suresh Kumar, "Software Performance Engineering for Cloud Applications – A Survey", International Journal on Recent and Innovation Trends in Computing and Communication(IJRITCC), ISSN: 2321-8169 Vol.4, No.2, February 2016.

[16] J. Schad, J. Dittrich, and J.A. Quiane-Ruiz, "Runtime Measurements in the Cloud: Observing, Analyzing, and Reducing Variance," Proceedings of the VLDB Endowment, vol. 3, 2010, pp. 460-471.

[17] Jin Shao and Qianxiang Wang, "A Performance Guarantee Approach for Cloud Applications Based on Monitoring", Proceedings of the 35th IEEE Annual Computer Software and Applications Conference Workshops, 2011.

[18] KhazaeiH(2012), "Performance analysis of cloud computing centers using M/G/m/m+r queuing system", IEEE Trans Parallel Distributed Systems, 23:936-943

[19] Xiaodong Liu, Weiqin Tong, XiaoliZhi, Fu ZhiRen, Liao WenZhao, "Performance analysis of cloud computing services considering resources sharing among virtual machines",Springer, Online, 20th March 2014

[20] Dan Marinescu, "Cloud Computing: Theory and Practice", Elsevier Science & Technology

[21] EvangelinGeetha D., Suresh Kumar T V , Rajanikanth K, Predicting the software performance during feasibility study, IET Software, April 2011, Vol.5, Issue 2, pp 201-215

[22] EvangelinGeetha D., Suresh Kumar T V , Rajanikanth K, Determining suitable execution environment based on dynamic workload during early stages of software development life cycle: a simulation approach, Int. Journal of Computational Science and Engg., Inderscience, Vol X., No.Y, 200X