IMPACT OF RESOURCE MANAGEMENT AND SCALABILITY ON PERFORMANCE OF CLOUD APPLICATIONS – A SURVEY

P.Ganesh[1], D Evangelin Geetha[2], TV Suresh Kumar[3]
[1, 2, 3] Department of MCA, BMSIT, Bangalore, India

ABSTRACT

Cloud computing facilitates service providers to rent their computing capabilities for deploying applications depending on user requirements. Applications of cloud have diverse composition, configuration and deployment requirements. Quantifying the performance of applications in Cloud computing environments is a challenging task. In this paper, we try to identify various parameters associated with performance of cloud applications and analyse the impact of resource management and scalability among them.

KEYWORDS

Service Level Agreement (SLA), Virtual Machine (VM), Quality of Service (QoS), Software Performance Engineering (SPE), Software Development Life Cycle (SDLC)

1. INTRODUCTION

Cloud computing has been envisaged as a novel paradigm for providing required services on demand [1]. Through Virtualization, physical resources are converted into a pool of resources to provide services on demand. In this context, a service provider possesses and accomplishes various physical resources and users/end users access them through the Internet. Considered as a paradigm shift to make software more convenient and attractive as a service and shape the way IT hardware is provided [2], cloud computing moves from concept to reality with overwhelming speed as seen currently.

Cloud computing services differ from traditional ones in 3 features. Initially, they are massively scalable; second, the services can be dynamically configured and delivered on demand; third, multiple VMs to run on the same physical server. As the complexity and diversity of user requests of the cloud computing system varies dynamically, its performance is difficult to model and analyse. Though enough attention is being paid by the cloud providers, performance un-predictability is a major concern for many cloud users and it is considered as one of the major obstacles for cloud computing acceptance. Hence it is imperative that cloud providers make quality of service guarantees by offering SLAs based on performance features [2]

Performance is a runtime attribute of a software system. It is required to describe the software runtime behaviour called dynamics of software, suitably, in order to analyze performance. Once good performance goes bad, the problem can originate within any one of a number of interacting application and infrastructure components. However, as the complexity of the cloud computing system varies dynamically, modelling its performance and analysing the same becomes difficult [3]. However, Model based approach is still preferred for performance analysis, as it can be

DOI : 10.5121/ijcsea.2016.6401

applied to any phase of the software life cycle. Most widely used performance analysis models include Queuing Networks, Stochastic Petri nets, simulation etc. To evaluate performance models to get proper performance indices, analytical methods and simulation techniques can be used. These indices include, throughput, utilization, response time, power consumption etc. in general. These indices in particular to cloud applications include resource sharing, responsiveness, scalability, latency, service availability, user satisfaction, reliability, accessibility etc.

The Section II provides a review on resource management; Section III provides review on scalability. Section IV summarizes the literature available on resource management and scalability.

2. Review On Resource Management

Dan Marinescu [4], defines resource management as a core function required in any man-made system. According to him, it affects the three basic criteria for system evaluation viz. performance, functionality and cost. Performance and cost are directly and heavily influenced by inefficient resource management. In addition, it can also indirectly affect system functionality. As an effect, due to poor performance, some functions of the system may become excessively expensive or futile.

Like for any other system, the above aspects hold good for a cloud computing environment as well. After all, the infrastructure of a cloud computing system is complex with a huge number of resources being shared. Cloud computing system is subject to random requests and hence it may be affected by external events beyond our control. Resources can be categorised as Computing resources, network resources, virtual resources etc. Typically cloud resources include Network, CPU and Disk among others. Cloud resource management thus requires complex policies and decisions for optimal utilization. It is also subject to continuous and unpredictable interactions with the environment.

The strategies for cloud resource management associated with the three cloud delivery models, namely, Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS), differ from each other, essentially. In all the three cases, the cloud service providers are faced with large and changing loads that challenge the privilege of cloud elasticity. As changes are recurrent and unpredictable in the cloud, it is unlikely that centralized control will provide continuous service and performance guarantees. In fact, adequate solutions cannot be provided by the centralized control to the pool of cloud management policies that we need to impose.

The principal guiding decisions are normally referred by a policy while, its associated mechanisms represent the means to implement such policies. As a guiding principle of computer science, we need to separate policies from their mechanisms. Cloud resource management policies can be roughly grouped into five classes namely; Admission Control, Capacity Allocation, Load Balancing, Cost Minimisation and Quality of Service.

The system can be prevented from accepting workloads in violation of high-level system policies using Admission control policy. As an example, an additional workload that might stop it from completing work already in progress or contracted may not be accepted by the system. Capacity Allocation takes care of allocating resources for individual instances. At the time when the state of individual systems is changing so rapidly, it requires to search a large space to locate resources that are subject to multiple global optimization constraints. The cost of providing services will be affected by the correlated effect of load balancing and energy optimization. Load balancing refers to evenly distributing the load to a set of servers. In cloud computing, another critical goal is minimizing the cost of providing the service. In precise, it refers to minimizing energy consumption.

This points to a different sense of the word, load balancing. Thus, we may need to concentrate the load and use the smallest number of servers switching the others to standby mode, where in a server uses less energy, in lieu of having the load evenly distributed among all servers. QoS is the key feature of resource management which is perhaps the most hard to address and, at the same time, probably the most critical and challenging to the prospects of cloud computing. Performance and power consumption are often together targeted by resource management policies.

All the optimal or near-optimal mechanisms to address the five policy classes do not scale up, virtually. They normally target either of the aspect of resource management, say, admission control, but overlook energy conservation. Many require complex computations that can't be done effectively in the time available to respond. Too complex performance models, intractable analytical solutions and the monitoring systems used to collect state information for these models are too invasive and unable to provide precise data.

As a result, most of the techniques of system performance are focussed on throughput and time in system. Energy trade-offs or QoS guarantees are seldom included by these techniques. Some techniques are based on unrealistic assumptions. As an example, under the assumption that servers are protected from overload, the capacity allocation is viewed as an optimization problem. Hence, based on a disciplined approach, the allocation techniques in computer clouds are identified, rather than ad-hoc methods. Following is the list of four basic mechanisms to implement resource management policies:

Control theory: To guarantee system stability and predict transient behaviour, It uses feedback, but can only predict local behaviour.

Machine learning: Machine-learning techniques don't need a performance model of the system.

Utility-based: It requires a performance model and a mechanism to correlate user-level performance with cost.

Market-oriented/economic mechanisms: It does not require a system model, such as combining auctions for bundles of resources.

3. REVIEW ON SCALABILITY

The prime goal of cloud environment is to provide equal or near-equal access to every user of cloud to its services. In this direction, one of the most related issues facing cloud computing is the ability of various services, let it be apps or web objects, to effectively scale up or down in order to match the system that they're running on. The ability of providing applications, processes etc. to ever growing numbers of users can be viewed as scalability. We shall consider a cloud's scalability as its performance report card. Most cloud users believe that all scalability issues are automatically dealt with by the cloud itself, due to virtualization of resources [5]. This is certainly not the case at all. It may be true that many cloud networks are set up to grant certain individuals (users) access to greater amounts of system resources than other cloud users. For example, Gmail (SaaS) access to subscribed users and normal users. Such kinds of provision of services selectively can cause scalability issues as well. Yet, many a times scalability problems are just a part of too many individuals accessing the similar data at the same time. This signals that the data delivery capabilities need to be significantly upgraded.

Two methods exist that are used to enhance the potentiality of scalability, namely: *Vertical Scalability and Horizontal Scalability.* By adding additional hardware such as hard drives, servers, CPU's, etc., we can achieve *Vertical scalability.* Though this is not a total solution in itself, it might be necessary if we need a cloud network to be able to accommodate ever increasing numbers of virtual machines.

*Horizontal scalability emphasises c*reating more access points within the current system. Horizontal scalability can be just a quick fix for a growing or expanding cloud. Horizontal scalability also serves as a kind of supportive for vertical scalability. Like the manner with which points of access are distributed across a network, horizontal scalability is closely associated.

Alternatively, integrating multiple load balancers into a cloud system is probably the viable solution for dealing with scalability issues. Currently, there exist many diverse forms of load balancers to choose from, such as, server farms, software to even hardware which have been intended to handle and distribute increased traffic.

Following list highlights the items that hamper the scalability.

1. Too much software clutter within the hardware stack(s)
2. Overuse of third-party scaling
3. Reliance on the use of synchronous calls
4. Not enough caching
5. Databases not being used properly

Finally, creation of a cloud network that offers the maximum level of scalability potential can be feasible if we apply a more "diagonal" solution. The benefits of horizontal and vertical scaling can be realised by incorporating best solutions present in both of them. When cloud servers reach the threshold (no growth state), we should simply start imitating them. By doing so, we shall be able to keep a consistent architecture when adding new components, software, apps and users.

4. LITERATURE SURVEY

Cloud computing, viewed as major IT revolution of the recent years, offering computing resources in a pay-as-you-go mode, is anticipated to minimize service operators' cost without compromising the performance of services. Yet, it is considerably harder to guarantee the performance of cloud applications given the architectural complexities of cloud, the types of interaction between co-deployed applications and the unpredictability of workload. There are numerous research resources available on various aspects of performance in cloud setup. The main motive behind the development and deployment of cloud computing is the on-demand dynamic and scalable resource allocation [6]. Hence, here we attempt to concentrate on two important aspects of cloud paradigm namely resource and scalability, in order to understand their relevance with performance of cloud applications.

Survey on the literature available on performance of Cloud applications and its related issues (Resource Management and Scalability) is carried out to understand the status of performance aspects of cloud applications and related concerns/challenges.

The capability to construct software and systems to provide performance objectives early in the SDLC is enabled by SPE. Conie U Smith, was first to propose SPE approach to integrate software performance analysis into the software engineering process. [7] Evangelin Geetha D et al., addresses the problem of performance analysis of distributed systems during feasibility study and also during the early stages of software development.[8] In this, a process model, called Hybrid Performance Prediction (HP3) model, was proposed to model and evaluate distributed systems with the goal of assessing performance of software system during feasibility study. Also it was determined that an execution environment can achieve the defined performance goal based on dynamic workload. [9]

Network resources is one of the central resources in cloud computing. Chrysa Papagianni et al., discusses allocation of virtual resources and describes that network performance, is key to the cloud performance.[10]

According to Sunil Kumar S. Manvi et al., resource management in IaaS can offer the benefits like Scalability, QoS, optimal utility, reduced latency, reduced overheads, improved throughput etc. provided the performance metrics such as delay, bandwidth overhead, security, reliability etc. are taken into consideration while addressing a resource management scheme.[11]

BahmanJavadi et al., explains that resource failure due to increasing functionality and complexity of hybrid cloud are inevitable. Such failures result in frequent performance degradation, premature SLAs, loss of revenue. A failure aware resource provisioning method, that is capable of addressing the end-users' QoS requirements was suggested.[12]

SaifU.R.Maliket. et al., work on virtualization emphasizes that any increase in the number of VMs does not disturb the working of the models in a highly scalable environment.[13] As resources available in a given single data center are restricted and hence if a large demand for an elastic application arises in a given time, then that cloud provider will not be able to deliver uniform QoS. In such a scenario, Rodrigo N Calheiros et al., proposes to enable further growing of application across multiple or independent cloud centers through cloud coordinators. In InterCloud project, this approach was realised using agents as cloud coordinators where increase in performance, reliability and scalability was observed.[14]

Initialising a new virtual instance is not instantaneous in Cloud. Sadeka Islam et al., propose a prediction based resource measurement and provisioning strategies using neural networks and linear progression.[15]

Disadvantage of cloud-based resource-sharing approaches, in particular to private clouds, is the existence of administrative boundaries due to the organizational principles. It is proposed to re-apply proxy-based resource sharing methods that supported solving similar problems in grid computing to overcome such limitations. D. CenkErdil strongly proposed that information proxies (proxies which distribute information on behalf of a resource or a collection of resources) can be viewed as main assisting mechanism for merging of grids and clouds. As per his findings, dissemination overhead can be reduced significantly under different scenarios, using information proxies.[16]

A self-organizing cloud can be modeled as scalable model that provide powerful computing ability with distributed computers. However, the resource allocation on such cloud will be very challenging as it involves various types of divisible resources and needs to manage with social competitions. Sheng Di et al., proposed ex-post efficient resource allocation for self-organizing cloud. [17]

Scalability heavily depends on the support of efficient network communications in virtualized environs. Roberto R. E. et al., analyzed the main bottlenecks of performance in HPC application scalability on Amazon EC2 and as a result it was echoed that network communication is the key performance logjam for scalability.[18]

Over provisioning and under provisioning of resources are yet unresolved issues in cloud computing. Over utilization and underutilization of cloud resources are considered as hitches because elasticity in pay-per-use cloud models has not been achieved yet. [19] Javier Espadas et al., proposed the practicality of a cost effective scalable environment via resource allocation mechanisms built on tenant isolation and VM instance allocation in a way to deploy SaaS applications over cloud platforms.[20]

5. OBSERVATIONS

An exhaustive survey on available literature relating to cloud resource management and scalability was carried out. This includes public cloud to hybrid cloud; IaaS to SaaS. Based on the above available literature, the following observations are derived.

1. Performance of cloud applications or services is dynamic in nature.

2. Performance issues can originate within any one of a number of interacting application and infrastructure components.

3. Performance analysis requires a great deal of understanding on the related aspects like resource sharing, responsiveness, scalability, latency, service availability, user satisfaction, reliability, accessibility etc.

4. Performance analysis can be effectively carried through model based approach.

5. Resource management and scalability are major areas of concern to performance of cloud services.

6. Over provisioning and under provisioning of resources are yet unresolved issues in cloud.

7. Network resources play crucial role in improving performance of cloud applications as it also affects scalability.

8. Information proxies can be used to reduce dissemination overheads.

9.Mere increase in the number of VMs does not affect the performance in a highly scalable environment.

10. Virtualization alone cannot solve the issues of scalability.

The below tables summarize the literature available on resource management and scalability related to performance of cloud computing.

Table 1. Summary of papers wrt Resource Management

Sl. No.	Author	Title	Observation
1	Chrysa Papagianni et al.	On the Otimal Allocation of Virtual resources in cloud computing networks	Network performance is key to the cloud performance.
2	Sunil Kumar S. Manvi et al.	Resource management for IaaS cloud : A survey	Resource management in IaaS can offer key benefits provided the performance metrics are taken into consideration while addressing a resource management scheme.
3	BahmanJavadi et al.	Failure-aware resource provisioning for hybrid cloud infrastructure	Resource failure due to increasing functionality and complexity of hybrid cloud are inevitable
4	Saif U.R.Malik et. al.	Modeling and Analysis of VM based cloud management platforms	Any increase in the number of VMs does not disturb the working of the models in a greatly scalable environment.
5	Rodrigo N Calheiros et al.	A coordinator for scaling elastic applications across multiple clouds	Cloud coordinators enable further growing of application across multiple or independent cloud centers.
6	Sadeka Islam et al.	Empirical prediction models for adaptive resource provisioning in the cloud	Prediction based resource measurement and provisioning strategies using neural networks and linear progression to initialise a new instance of VM

Table 2. Summary of papers wrt Scalability

Sl. No.	Author	Title	Observation
1	Sheng Di et al.	Ex-Post efficient resource allocation for self-organizing cloud	Ex-post efficient resource allocation for self-organizing cloud, is possible.
2	Roberto R. Exposito et al.	Performance analysis of High Performance Computing applications in the Cloud	Network communication is the key bottleneck of performance for scalability.
3	M. Stillwella et al.	Resource allocation methods for virtualised service platforms	Over provisioning and under provisioning of resources are still unresolved issues in cloud.
4	Javier Espadas et al.	Tenant-based resource allocation model for scaling SaaS over cloud infrastructure	Feasibility of cost effective scalable environment via resource allocation methods built on tenant isolation and VM instance allocation.

6. CONCLUSION

It is required to describe the software runtime behaviour called dynamics of software, suitably, in order to analyze performance. . Once good performance goes bad, the problem can originate within any one of a number of interacting application and infrastructure components. However, as the complexity of the cloud computing system varies dynamically, modelling its performance and analysing the same becomes difficult. Performance is affected by various parameters like resource sharing, responsiveness, scalability, latency, service availability, user satisfaction, reliability, accessibility etc. and is measured in terms of throughput, utilization, response time, power consumption etc. in general. Though there are multiple parameters that influence cloud services, resource management and scalability are the important two parameters that play crucial role in determining the QoS. In this paper we tried to analyze the available literature on cloud performance in general and on these two aspects in particular. It was observed that network resources are the one critical resource that impact scalability. Numerous methods as pointed in the literature can be used to mitigate the issues like over utilization and underutilization of resources, reduction of dissemination overheads etc. In our future work, we would like to extend the survey to analyze other critical aspects of performance.

REFERENCES

[1] 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

[2] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R.H. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, “Above the Clouds: A Berkeley view of cloud computing”,EECS Department, UCB, Tech Rep. UCB/EECS-2009-28,2009.

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

[4] Dan Marinescu, “Cloud Computing: Theory and Practice”, Elsevier Science & Technology

[5] http://artofservice.com.au/cloud-computing-and-scalability-issues, 2 May 2011

[6] Ashraf Zia and Muhammad Naeem Ahmad Khan, “Identifying Key Challenges in Performance Issues in Cloud Computing”, Int. Journal of Modern Education and Computer Science, 2012, 10, 59-68

[7] C U smith, Performance Engineering of Software Systems, Addison Wesley, 1990.

[8] Evangelin Geetha 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

[9] Evangelin Geetha 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

[10] ChrysaPapagianni et al., “On the Otimal Allocation of Virtual resources in cloud computing networks, IEEE Transactions on Computers”, Vol 62, No.6, June 2012

[11] Sunil Kumar S Manvi and Gopal Krishna Shyam, “Resource management for infrastructure as a service(IaaS) in cloud computing : A survey”, Journal of Network and Computer Applications, Elsevier, 2013

[12] BahmanJavadi, JemalAbawajy, RajkumarBuyya, “Failure-aware resource provisioning for hybrid cloud infrastructure”, Journal of parallel distributed comuting, Elsevier, June 2012

[13] Saif U.R. Malik et al., “Modeling and Analysis of state of the art VM based cloud management platforms”, IEEE Transactions on Cloud Computing, Vol.1, No.1, January-June 2013

[14] Rodrigo N Calheiros, Adel NadjaranToosi, Christian Vecchiola, RajkumarBuyya, “A coordinator for scaling elastic applications across multiple clouds”, Future Generation Computer Systems, Elsevier, 2012

[15] Sadeka Islam, Jacky Keung, Kevin Lee, Anna Liu, “Empirical prediction models for adaptive resource provisioning in the cloud”, Future Generation Computer Systems, Elsevier, 2011.

[16] D.CenkErdil, “Autonomic cloud resource sharing for intercloud federations”, Future Generation Computer Systems, Elsevier, 2012.

[17] Sheng Di, Cho-Li Wang, Ling Chen, “Ex-Post efficient resource allocation for self-organizing cloud”, Computers and Electrical Engineering, Elsevier, 2013.

[18] Roberto R. Exposito, Guillermo L. Taboada, Sabela Ramos, Juan Tourino, Ramon Doallo, “Performance analysis of HPC applications in the Cloud”, Future Generation Computer Systems, Elsevier, 2013

[19] M. Stillwella, D. Schanzenbacha , F. Vivienb, H. Casanova, “Resource allocation algorithms for virtualised service hosting platforms”, Journal of Parallel and Distributed Computing, 70(9)(2010) 962-974.

[20] Javier Espadas, Arturo Molina, Guillermo Jimenez, Martin Molina, Raul Ramirez, David Concha, “A tenant-based resource allocation model for scaling software-as-a-service applications over cloud computing infrastructure”, Future Generation Computer Systems, Elsevier, 2011.

AUTHORS

P. Ganesh, working as Associate Professor, Department of MCA at BMS Institute of Technology and Management, Bangalore, India. He has 14 years of teaching experience. His areas of interest are Software Engineering, SPE, Cloud Computing.

Dr. Evangelin Geetha, working as Associate Professor, Department of MCA at MS Ramaiah Institute of Technology, Bangalore, India. She has 20 years of teaching and 4 years of research experience. Her areas of interest are Software Engineering, SPE, Distributed Computing, Cloud Computing.

Dr. T V Suresh Kumar, working as Professor and HoD, Department of MCA, at MS Ramaiah Institute of Technology, Bangalore, India. He has 25 years of teaching and 10 years of research experience. His areas of interest are Software Engineering, SPE, Distributed Computing, Cloud Computing, Business Analytics.