International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on QoS in IaaS

Shital Makwana

Computer Engineering Department, Silver oak college of Engineering & Technology, Gujarat Technological University, India

Abstract: ***This paper aims at surveying various elasticity algorithms to improve Quality of Service (QoS) in Infrastructure as a Service (IaaS) model in cloud computing. The paper briefly discusses cloud computing basics, its service and deplyment models, Qulaity of Service in IaaS model and its parameters. Various elasticity methods available in literature and commercial are discussed.***

Keywords: Cloud Computing, Quality of Service (QoS), Infrastructure as a Service (IaaS), Elasticity.

1. Introduction

A *cloud* is a distinct IT environment which is designed for the aim of remotely provisioning scalable and measured IT resources. We can compare cloud environment with the Internet which is, a network of networks which provides remote access to decentralized IT resources. The symbol of a cloud is commonly used to represent the Internet in many architectures and specifications. The symbol of cloud is now used to represent the boundary of a cloud environment.

Cloud computing can change a large part of the IT industry by making software more attractive as a service and altering the way IT hardware is designed and purchased. Cloud computing includes both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services, thus including software & hardware both. Developers with innovative ideas for new Internet services no longer require the large capital outlays in hardware to deploy their service or the human expense to operate it. They need not be concerned about over provisioning for a service whose popularity does not meet their predictions, thus wasting costly resources, or under provisioning for one that becomes wildly popular, thus missing potential customers and revenue [1].

NIST defines cloud computing as "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" [2]. Cloud provides ubiquitous access as services are available online so you get service wherever internet is available .different physical and virtual resources are dynamically assigned to multiple consumers on demand. As per application need resources are provisioned and released manually or automatically. This elasticity feature distinguishes cloud computing with other technology. Users are charged in pay-per-use model.

2. Cloud Service Models

There are basically three Service Models of Cloud computing as shown in Fig.1.

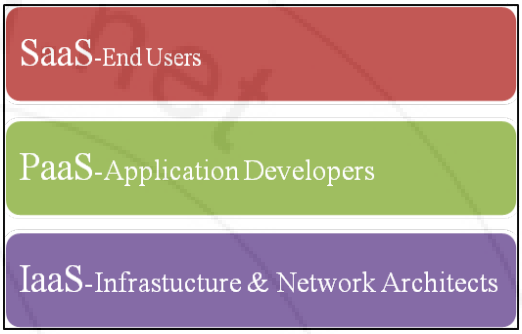

Figure 1: Cloud Service Models

A. Software as a Service (SaaS)

SaaS is software that is owned, delivered and managed remotely. Cloud provider maintains and manages the software services which are used by the cloud consumer. The software is deployed on the internet which is accessed by user via a web browser. It requires zero installation of the software and is accessible from anywhere with an internet connection. Cloud provider updates software as new version comes. Google Apps such as Google Mail and Google Docs and Spreadsheets are examples for SaaS.

B. Platform as a Service (PaaS)

PaaS provides ready to use environment for users to support entire lifecycle of custom application. Developers can write their applications according to the specifications of a particular platform. PaaS provides platform that basically includes resources like operating system, database, web server, programming language, libraries which automatically scales to meet the demands. Consumer do not need to manage resources but can configure settings for application hosting environment. Google App Engine, Red Hat OpenShift, Engine Yard are PaaS providers.

C. Infrastructure as a Service (IaaS)

IaaS offers computing resources such as storage or processing which can be obtained as a service without requiring any physical hardware on their own site. It is the base layer for cloud computing which basically deals with virtual machines, storage, servers, networks, load balancers, and the IaaS cloud providers supply these resources on-demand [3]. Hardware cost for organization can be greatly reduced here. Amazon Web Service (AWS), Rackspace, Windows Azure etc provide Infrastructure as a Service.

3. Cloud Deployment Models

Cloud is generally classified by the owner of data center, whether it is owned by some third party, organization, community or combination of any of this. There are four deployment models as shown in Fig.2.

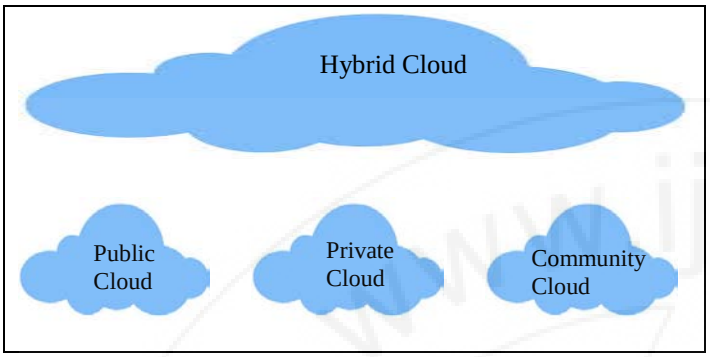

Figure 2: Cloud Deployment Models

D. Public cloud

A Public cloud is owned and managed by third parties. It is generally on premises of cloud provider. Public Cloud is provisioned for open use by the general public on pay per use basis. Public cloud may raise number of issues like management, security, performance, level of control, privacy etc.

E. Private cloud

A Private Cloud is fully owned by a single company who has total control over the applications run on the infrastructure, the place where they run, and the people or organizations using it [4].Private Cloud Infrastructure is used by organization staff only. It may exist on premises or may be managed by third party but is not shared among general public.

F. Community cloud

Different communities with same objectives come together to form a cloud so it is available to specific group of people. Services are shared by all members of community. Community cloud may exist on or off premise.

G. Hybrid cloud

It is made by composition of two or more infrastructures like public, private or community cloud. User can use private or community cloud and can expand to public cloud as needed.

4. Quality of Service in Cloud

Quality of Service (QoS) for clouds is very necessary topic in research community. Certain level of QoS is required in order to deliver and give a smooth user experience. QoS is expressed in terms of Service Level Agreements (SLA) that is contract between cloud user and provider for users to get predefined quality in service.SLA includes obligations, price & penalties in case agreement violates. QoS criteria are dependent on the application. QoS parameters in the IaaS are described below.

A. Availability

Availability is expressed as the percentage of time a service that is hosted on a cloud is running and active. It is denoted as up-time of a service. Most of public cloud provider guarantees more than 99.9% availability. User usually receives service credits if the provider fails to meet the agreed SLA.

B. Performance

IaaS provides services through Virtualization. Cloud provides an isolated and secure environment to the users in the form of Virtual Machine but resources such as processing cores, I/O devices and memory are mostly shared. Resource congestion raised due to Variability in the workload of VMs can cause performance degradation. Dynamic workload requires dynamic resources to keep performance intact without leaving resources idle.

C. Scalability

If a system can accommodate future growth or can continue functioning as better as before enlargement then it is said to be as scalable. Here we talk about scalable cloud where it is possible add resources whenever the demand rises, in order to keep applications performing as SLA do not violates. People use elasticity and scalability interchangeably but both terms have major difference. Elasticity is the ability for users to quickly request, receive, and release as many resources as they require. The scalability metrics involve storage scalability, server scalability horizontal and vertical both.

5. Elasticity: A Means to Improve QoS

Applications requests for resources as their requirement varies with time. Elasticity in clouds provides the facility to the users not to purchase the resources for peak workload, rather request for more resources when demand increases and release resources when not required, hence decreasing total cost of ownership. Open Data Center Alliance (ODCA) [6] defines elasticity as the configurability and expandability of the solution. Centrally, it is the ability to scale up and scale down capacity based on subscriber workload". This is because all of the factors like delay in allocation of resources, scheduling points etc. finally cause a mismatch in demand and supply curve, leading to over-allocation or under-allocation which overall affects on Quality of Service.

A. Horizontal Scaling

Horizontal scaling also called as replication is a method used in which resource scaling is achieved by adding or removing number of VMs to support the changing demand of application. For horizontal scaling, another VM can be added and the load balancer re-distributes the load among all of the VMs. horizontal scaling is less expensive then vertical scaling.

B. Vertical Scaling

Vertical scaling also called as resizing is a method in which the change in workload is handled by migration of the application to a different VM which might be on a different physical host. Although migration is a costly operation and incurs some penalty in terms of availability and performance during migration, but for applications which can't scale horizontally, resource scaling is achieved by migration only.

Static VM sizes forces users to characterize their resources on coarse-grained level leading to under-utilized resources or under-performing applications as Cloud users pay even for the idle resources and idle resources cannot be utilized by the cloud providers.

6. Approach

There are two approaches available to find elastic solution. First one is Manual in which cloud provider provides application programming Interface with which user interacts with system and he is responsible for monitoring his virtual environment and applications and for performing all elasticity actions. User can add or remove resources as per his requirement. GoGrid [7], Rackspace [8] and Microsoft Azure [9] are few cloud service providers to mention in this category. Another solution is Automatic in which according to user-defined rules and settings, cloud system automatically takes decision about scaling. This automatic action can be reactive or predictive.

Reactive solutions are based in if-then mechanism. If certain conditions are satisfied then predefined action is taken. Amazon [10], Rightscale [11], Scalr [12], André et al. [13] use this approach. The predictive approach uses heuristics or mathematical & analytical techniques to predict the system load behavior, and based on that results decides when and how to scale in or scale out resources. The works of PRESS [14], Roy et al. [15], Shen et al. [16], Ali et al.[17] use predictive techniques to automatically scale resources.

Amazon Web Services is a popular commercial cloud service provider which provides replication mechanism called Auto-Scaling as a part of EC2 services. CloudWatch is the monitoring service utilized by EC2 which provides the performance metrics based on CPU, network bandwidth, memory etc and balances the workload among active instances. Apart from commercial solutions, a vast literature is available which provides various reactive/predictive solutions to manage the computation resources.

Predictive Elastic Resource Scaling (PRESS) [14] is a predictive elasticity algorithm which identifies repeating patterns, called signatures, by implementing Fast Fourier Transform (FFT) and predicts workload based on the signatures. If no signature is found, it employs a statistical state-driven approach and a discrete-Markov-chain to predict the near future workload. PRESS pads 5-10% of predicted value to avoid risk of Service Level Objective (SLO) violations and avoids possible under-provisioning.

N. Roy et al. [15] implements a look-ahead resource allocation algorithm based on autoregressive-moving-average (ARMA) model to predict the future workload and provisions VMs to the users. Minimizing a cost function based on the time interval, it optimizes resource usage and minimizes idle resources hence keeping operational costs low. The model is fit for stationary processes and can't be directly applied to non-stationary processes.

Z. Shen et al. [16] proposed a system derived from PRESS whose architecture has four major components: Resource Demand Prediction module was already presented in PRESS. Prediction Error Correction module minimizes SLOs by scaling up the predicted resources. Scaling conflict handling manages the conflicts when the local resources are not sufficient to satisfy all of the VMs' requirements based on local conflict handling priorities, or decide for migration. Predictive frequency voltage scaling adjusts the hosts' CPU frequency to save energy consumption without affecting application SLOs.

André et al. [13] proposes CloudDReAM (Dynamic Resource Allocation Middleware) for supporting large scale and highly interactive applications such as Massively Multilayer Online Games (MMOG) which require a large infrastructure capable of processing and disseminating a large amount of data with high performance and constant availability. These gaming applications may require as much as 10000 servers at a time but most of the time the resource usage is low. Rather than globalized load balancing, this approach utilizes local load balancing using the middleware and hence it is lightweight. This algorithm is based on reactive approach and sets corresponding over provisioning and under provisioning threshold values and when a threshold value is reached, appropriate actions are taken depending on whether the system is under-loaded or overloaded. Since the approach is not predictive, the provisioning always lags the demand in time.

Ali et al. [17] proposes proactive autoscaling based on Hidden Markov model (HMM) to capture and model stochastic time-series when the actual sequence of the states is not observable as compared to Markov Models. The algorithm utilizes HMM Weka tool to implement HMM along with TPC-W benchmark as a load generator. Although the model is quite accurate but the time taken to collect historic data for training and establishing the model parameters is large.

7. Conclusion

In this paper, we have surveyed various algorithms/methods of elasticity for resource allocation to virtual machines for improvement of Quality of Service (QoS) and minimize Service Level Agreement (SLA) violations. We concluded that for variable workload conditions, one needs to predict the future workloads and allocate the resources accordingly. The limitations and advantages of various algorithms is also presented which in general concludes that manual elastic resource provisioning policy demands user intervention and inefficient in resource utilization. Reactive policy is better but provisioning always lags the demand in time and hence SLA violation may occur. Forecasting based policy is superior to them and will be the component of future cloud resource provisioning technologies.

References

[1] M. Armbrust, A. Fox, R. GriFFIth, A. D. Joseph, R. Katz, A. Konwinski, G. Lee,D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing,"Commun. ACM, vol. 53, no. 4, pp. 50{58, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1721654.1721672

[2] P. Mell and T. Grance, The nist definition of cloud computing," National Institute of Standards and Technology [Online]. Available:http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf
[3] S.K. Sowmya, P. Deepika, J. Naren,"Layers of Cloud – IaaS, PaaS and SaaS: A Survey", *International Journal of Computer Science and Information Technologies*, Vol. 5 (3) , 2014, 4477-4480
[4] Thomas Erl, Zaigham Mahmood and Ricardo Puttini,"Cloud computing concepts, technology and architecture"
[5] Jens Myrup Pedersen, M. Tahir Riaz, Bozydar Dubalski, Damian Ledzinski, Joaquim Celestino Júnior, Ahmed Patel,"Assessing Measurements of QoS for global Cloud Computing Services" *Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing,2011,pp. 682-689*
[6] "Open data center alliance: Compute infrastructure as a service rev, 1.0," 2012. [Online]. Available: http://www.opendatacenteralliance.org/docs/DCACompute IaaS MasterUM v1.0 Nov2012.pdf
[7] GoGrid." [Online]. Available: http://www.gogrid.com/
[8] "Rackspace." [Online]. Available: http://www.rackspace.com/
[9] "Microsoft Azure." [Online]. Available: http://www.windowsazure.com/
[10] "Amazon Web Services." [Online]. Available: http://aws.amazon.com/
[11] RightScale." [Online]. Available: http://www.rightscale.com/
[12] Scalr." [Online]. Available: http://scalr.net/
[13] André Pessoa Negrão, Miguel Adaixo, Luís Veiga and Paulo Ferreira, "On demand Resource Allocation Middleware for Massively Multiplayer Online Games", IEEE 13th International Symposium on Network Computing and Applications,2014 pp. 71-74
[14] Z. Gong, X. Gu, and J. Wilkes, "Press: Predictive elastic resource scaling for cloud systems," in *Proceedings of the 6th Intl. Conference on Network and Service Management*, ser. CNSM 2010. IEEE, 2010, pp.9-16
[15] N. Roy, A. Dubey, and A. Gokhale, "Efficient autoscaling in the cloud using predictive models for workload forecasting," in *Proceedings of the 4th Intl. Conference on Cloud Computing*, ser. CLOUD 2011. IEEE, 2011, pp. 500–507.
[16] Z. Shen, S. Subbiah, X. Gu, and J. Wilkes, "Cloudscale: elastic resource scaling for multi-tenant cloud systems," in *Proceedings of the 2nd Symposium on Cloud Computing*, ser. SOCC 2011. ACM, 2011, pp. 5:1–5:14.
[17] Ali Yadavar Nikravesh, Samuel A. Ajila, Chung-Horng Lung, "Cloud Resource Autoscaling System based on Hidden Markov Model (HMM)", IEEE International Conference on Semantic Computing, 2014 pp. 124-127