Industrial Engineering Journal
ISSN: 0970-2555
Website: www.ivyscientific.org

Industrial Engineering Journal

Volume 15 Issue 10 * October 2022

Survey of the State of Art of QoS Modeling Approaches

Ritesh Kumar Bhanu

M.Tech. Research Scholar,Department of Computer Science & Engineering

Dr. RammanoharLohiaAvadh University, Ayodhya

Dr.Ashish Kumar Pandey

Assistant Professor, Department of Computer Science & Engineering

Dr. RammanoharLohiaAvadh University, Ayodhya

Er. Dilip Kumar

Assistant Professor, Department of Electrical Engineering

Dr. RammanoharLohiaAvadh University, Ayodhya

Corresponding Author's Email: ashishkumarpandey@rmlau.ac.in

Abstract

One of the important aspects of the deployment, management and orchestration is the quality of service, security, performance, and energy consumption. It has been further stated that the inappropriate allocation of certain resources can further lead to resource contention, entailing reduced performance, damaging effects and poor energy efficiency. One of the challenges which is posed by the cloud applications is the Quality-of-Service management which basically is the problem of allocating resources to the application to guarantee service along with certain dimensions such as performance, availability and reliability. The paper basically focus on the area by providing the survey of the state of the art of QoS modelling approaches which are suitable for the cloud systems.

Keywords: Tradeoff, QoS, Modeling Techniques, Energy Consumption, Cloud

I. Introduction

Due to the increase in the technical and economical benefits of the on-demand capacity management model, cloud computing has increased over few years. There are number of cloud operators which are active in the market and consist of Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software-as-a-Service solutions. It has also become mainstream for the enterprise datacenters which includes the private and cloud architectures which are increasing day by day (St-Onge et al, 2021).

Though it has been great simplification, but the process further includes several novel challenges in the area ofQuality-of-Service management which states the levels of performance, reliability and availability which is offered by an

application or the platform or the infrastructure. QoS is important for the cloud users who expect to deliver the good characteristics and for the cloud providers who need to find the good trade-offs between the QoS levels and operational costs. However, it has been stated that it is quite a difficult decision which is often ignored by the presence of the service level agreements which states QoS targets and economical penalties which are linked to the SLA violations (Khattar et al, 2019).

The QoS properties have received major attention with the advent of the cloud computing, heterogeneity of the performance and resource isolation mechanism of the cloud platforms which consists of the prediction and assurance. It has been further stated that there are several researchers which have stated regarding the QoS management modules which can help in leveraging the high programmability of the software and hardware resources in the cloud.

Cloud computing consist of that kind of operational model which consists of different other technological advancements which includes virtualization, web services and SLA management for enterprise applications. There are different other diverse modeling techniques which helps to cope up with technological heterogeneity. Thus, it has been stated that the QoSmodeling literature is quite extensive which makes it quite difficult in the view of the available techniques and the current applications to the cloud computing problems.

II. Workload characterization

a. Deployment Environment: There are different kinds of studies which are made to show the cloud deployment environments through the process of benchmarking. There are different kinds of characteristics which uses empirical data which is used in the QoSmodeling in order to quantify the risk which needs to conduct for the ad-hoc measurement study. It is important to state the values of the QoS model parameters which consist of network bandwidth variance, virtual machine, startup times and start failure probabilities. There are different types of VM instances which includes observations of the performance. One of the major primary causes of this variability is hardware heterogeneity and VM interference. The works state regarding the variability in the VM startup times which is correlated to the operating system image size. There are different other studies on Amazon EC2 which includes high performance contention in the CPU bound jobs and network performance overheads.

b. Cloud Application Workloads: The users often face further problems which state regarding the properties of the workloads which are processes by the cloud application and further consist of the properties of the cloud deployment environment.

In order to predict the web traffic intensity at different timescales there are different other techniques used known as Blackbox forecasting and trend analysis techniques. For almost two decades, time series forecasting has been used for web servers. There are some autoregressive models which are common in applications and include cloud application modeling for example auto-scaling. There are two other techniques which further include kernel-based methods, wavelet-based methods, regression analysis, filtering and Fourier analysis.

Khan et al (2012) stated regarding the Hidden Markov Model which is basically used to predict the temporal correlations between the workloads of the compute clusters in the cloud computing. In this paper, the author further stated the method for the workloads in the cloud environments which include provision for the cloud resources. The authors further state regarding the Co-Clustering algorithm which consists of the same kind of servers in the same workload patterns. The pattern further consisted of studying the performance on the different kind of applications which further uses hidden Markov models which is used to identify the temporary correlations between the different kind of

clusters and different kind of information which is used to know the workload variations. Di et al further in one of his study states regarding the Bayesian algorithm for the workload prediction and pattern analysis which states the results which are obtained from the data centers of the Google. The author's further states regarding the major key nine features which states regarding the workload and consist of the Bayesian classifier in order to know the posterior probability of each feature. The experiments further stated regarding the large dataset which is collected from Google data center with thousands of machines.

Gmach et al (2007) stated regarding the data center and cloud workload data. The authors further stated regarding the workload demand prediction algorithm which is basically based on trend analysis and pattern recognition which states an efficient way to use the resource pool to allocate servers to the different workloads. The pattern and the trends are firstly analyzed and consist of synthetic workloads which are basically consisted of reflecting the future behaviours of the workload. Zhu and Tung basically in one of their studies stated the use of Kalman filter which is used to model the interference on the deploying applications on the virtualized resources. The model further states regarding the time variations in the VM resource usage and consist of the VM consolidation algorithm which is further tested and competitive. The work loading modeling is basically best practice guide is basically to build empirical models. There are important issues which are treated which include relevant data and variable selection procedure. The authors further state regarding the comparative study which states the benefits of the different forecasting approaches.

III. Workload Inference

The ability of the resource demands is basically uses the QoS models in the case of enterprise applications. It consists of basically overheads of deep monitoring and the difficulty of tracking the execution paths of the individual requests. There are several works which states regarding the problem of estimating, using indirect measurements, the resource demand placed by the application on the physical resources which includes CPU requirements. From the perspective of the cloud providers and users which includes the inference techniques. This means to state the workload profile of the individual VMs running on their infrastructures include the account hidden variables due to the lack of information.

IV. Regression Techniques

One of the common workload inference approaches consist of the mean demand which is dependent on the type of the request on the resources. The technique is basically based on comparing the performance metrics which is basically predicted by the performance model against the measurements which is basically collected in controlled experimental environment. The lack of control over the system workload and configuration of the operation techniques of this type may not be applicable to the production systems for online model calibration. The methods further consist of the theory formulas to state the mean values of the set of performance metrics to a mean demand to be estimated for example CPU demand. There are regression techniques which are used to obtain demand from system measurements.

Zhang et al (2007) stated regarding the network model where the queue states regarding the tier of web application which is based on the parameters of the regression-based approximation of the CPU demand of customer transactions. It is further stated that the approximation is important for the modeling different kind of workloads whose transaction mix changes over time.

Liu et al (2006) stated regarding the service demand estimation which includes utilization and to end to end response times, the problem is further formulated as quadratic optimization programs which are based on the formulas which

results in good agreement with the experimental data. The different kind of regression methods have been developed to cope up with the problems such as outliners, data multi collinearity, online estimation, data aging and automatic definition of request types.

Kalbasi et al (2012) stated regarding the Demand Estimation with Confidence which states how to overcome the problem of multi-collinearity in the regression models. It has been implied to improve the estimation accuracy.

Cremonesi et al (2012) stated regarding the algorithm which states regarding the service demands for different system configurations. One of the algorithms known as time based linear clustering algorithm is used to identify the clusters for each service demands. The approach further proves to be robust for the noisy data. Generated data usually show the effectiveness of the algorithm.

V. System Models

In order to improve the effectiveness of the QoS management models, explicit models of the logic are used. There are different kinds of models which are used to model QoS in the cloud systems. In the case of reviewing the queue models, Petrinets and there are other specialized formalisms used for reliability evaluation. There are different other models which exist such as stochastic reward nets, stochastic activity networks, stochastic process algebras. There are some pros and cons of the stochastic formalisms which is found in the highlights of the author where he stated regarding particular method which works on one of the system models but not on others, making it difficult to the absolute recommendations on the best model.

a. Performance Models: In the case of the queueing models, these models are basically used to model the single resources subject to contention and queueing networks which are basically used to capture the interaction among number of resources and application components. There are performance models which are used to survey the queueing models and networks. LQNs are basically used as model key interaction between the application mechanisms which include finite connection pools, admission control mechanisms or synchronous request calls. The feature basically includes in depth knowledge of the application behaviour. On the other hand, it includes other closed form solutions which include classes of queueing systems and networks, solutions of other models which basically rely on numerical methods.

b. Dependability models: There are different kind of methods which include Petri nets, Reliability Block Diagrams and Fault Trees. One of the flexible and expressive modelingapproaches is the petri nets which include general interactions between the system components, and which include synchronization of the event firing times. There are different kinds of large applications also which are used in performance analysis RBDS and fault trees are used to obtain the reliability of the overall system. The interactions of the components state regarding the faulty state and component results which lead to the possible failure of another components.

c. Black- Box Service Models: There are different kinds of models which are being used in the optimization of the web service composition, and these kinds of models are further helping in the cloud-based business process execution, SaaSapplications and IaaS resource orchestration. One of the major ideas behind this method is basically it describes the

response side, lack of any other information which is required and internal kind of characteristics which includes contention level from concurrent requests.

There are different kinds of non-parametric black box service models which include methods which are totally based on the average execution time values. There are several other works which include standard deviations or the finite ranges of the variability for the time execution. There are parametric service models instead of Markovian distributions or exponential methods. In the case of heavy tailed execution times, there are general distributions with Laplace transforms.

Huang et al (2012) stated regarding the graph theoretic model which includes QoSaware service composition in the cloud platforms which include handling network virtualization. Hence, it can be stated that there are different other authors which stated regarding the service provisioning in the cloud platforms which are basically consisting of the virtual network services. A system model basically state the characteristics of the cloud service provisioning behaviours and the exact algorithm is basically used to optimize the experience of the users which are under QoS requirements. A comparison with the state-of-the-art QoSstates regarding the proposed algorithm which is considered to be both light weight and cost effective.

Klein et al (2012) stated regarding the handling of the network latencies which uses QoS aware service composition. The authors state regarding the network model which consist of the latencies between the locations and the genetic algorithm which is used to achieve network aware and service provisioning.

d. Simulation Models: In the case of the cloud system stimulation there are different other stimulation packages which exist. There are many other solutions which are basically based on the toolkit known as CLOUDSIM [93] toolkit which further helps to make use of the potential cloud resources and helps the user to set up a stimulation model which is located in different kind of data centers in the case of hybrid deployments. One of the extensions of the CLOUDSIM is CLOUDANALYST which is known to be well known for the geographically distributed workloads which are served by the applications which are deployed on the number of virtualized data centers.

In the CLOUDSIM, there is one more thing known as EMUSIM which helps in the process of emulation and consist of Automated Emulation Framework. The major importance of the emulation is to understand the application behavior which helps in extracting the profiling information. The information basically is used as input for CLOUDSIM, which helps in the process of cloud deployment.

There are different other tools which are further developed in the case of data center energy consumption. For example, GREENCLOUD is basically known as that extension which is used for the packet level stimulator which basically focuses on the energy consumption of the data center, where the application is deployed and consists of server, links, and switches (St-Onge et al, 2021).

VI. Conclusion

In order to reduce the over provisioning of the required resources and consumption of energy one of the major factors is the consolidation of services. It has been stated that whenever there are multiple jobs or resources running on multicore CPU which include number of shared resources such as networking, caches, memory controllers, memory buses, perfecting hardware, and disks etc. The energy consumption and QoS violation can be increased by the invoking of performance delegation which reduces the benefits of the consolidation.

In the recent years, it has been stated that the process of cloud computing has matured which means from early stage to the mainstream operational model. However, there are current approaches in the workload and system modeling and the early applications to the cloud QoS management. The diversity in the technology basically makes it very difficult to analyze the QoS and their service level guarantees.

In the case of QoS management, there has been number of techniques applied but one of the techniques which are known as white box system modeling technique is basically limited and popular in the software performance engineering community. The technique basically create gap between the knowledge that is available for an application and which the designers and techniques uses to manage it. One of the research questions is basically that point where it has been stated regarding the availability of the application internals which can provide major benefit in QoS management. It has been further stated that there is major trade off which exist between the available information, QoS complexity, computational cost of decision making and accuracy of predictions and that trade off requires investigation by the research community.

One of the major roles is played by the pricing models in the case of the cloud systems. In the case of computing resources there is growing interest between the cloud spot markets where certain types of bidding strategies are developed for procuring computing resources. There are different other approaches which states regarding the cloud resources selection and dynamic pricing. It has been expected that in the upcoming years, the models that has been stated plays major role of importance in the capacity allocation frameworks.

References

1. Cremonesi P, Sansottera A: Indirect estimation of service demands in the presence of structural changes. *In Proceedings of Quantitative Evaluation of Systems (QEST)*. IEEE, London, UK; 2012:249–259.
2. Gmach D, Rolia J, Cherkasova L, Kemper A (2007) Workload analysis and demand prediction of enterprise data center applications. In: Proceedings of the 2007 IEEE 10th International Symposium on Workload Characterization, IISWC ’07, 171–180, Boston, MA, USA.
3. Huang J, Liu Y, Duan Q (2012) Service provisioning in virtualization-based cloud computing: Modeling and optimization. *In: Proceedings of 2012 IEEE Global Communications Conference, GLOBECOM 2012*, 1710–1715, Anaheim, CA, USA.
4. Kalbasi A, Krishnamurthy D, Rolia J, Dawson S. (2012). DEC: Service demand estimation with confidence. *IEEE Trans SoftwEng*, 38(3), 561–578.
5. Khan A, Yan X, Shu T, Anerousis N (2012) Workload characterization and prediction in the cloud: A multiple time series approach. *In: Proceedings of the 2012 IEEE Network Operations and Management Symposium, NOMS 2012*, 1287–1294, Maui, HI, USA.
6. Khattar N, Sidhu J, Singh J. (2019). Toward energy-efficient cloud computing: a survey of dynamic power management and heuristics-based optimization techniques. *The Journal of Super Computing*, 75(8), 4750-4810.
7. Klein A, Ishikawa F, Honiden S (2012) Towards network-aware service composition in the cloud. *In: Proceedings of the 21st International Conference on World Wide Web, WWW ’12*, 959–968, Lyon, France.

8. Liu Z, Wynter L, Xia C, Zhang F. (2006). Parameter inference of queueing models for it systems using end-to-end measurements. *Perform Eval,* 63(1), 36-60.
9. St-Onge, C., Benmakrelouf, S., Kara, N. et al. (2021). Generic SDE and GA-based workload modeling for cloud systems. Retrieved from https://doi.org/10.1186/s13677-020-00223-5.
10. Zhang Q, Cherkasova L, Smirni E (2007) A regression-based analytic model for dynamic resource provisioning of multi-tier applications. *In: Proc. of the 4th ICAC Conference*, 27–27, Jacksonville, Florida, USA