

Cloud Computing and Quality of Service: Issues and Developments

Isaac Odun-Ayo, *Member, IAENG*, Olasupo Ajayi, and Adesola Falade

Abstract - Cloud computing is a dynamic information technology (IT) paradigm that delivers on demand computing resources to a user over a network infrastructure. The Cloud Service Provider (CSP) offers applications which can be accessed online to users. Such applications can be shared by more than one user. CSPs provides programming interfaces that allows customers to build and deploy applications on the cloud; as well as providing massive storage and computing infrastructure to users. Users usually have no control on how data is stored on the cloud or where the underlying resources are located. With this limited control, customers' requirements and Quality of Service (QoS) expectations from CSPs are spelt out using a Service Level Agreement (SLA). It is thus imperative to have the adequate QoS guarantees from a CSP. This paper examines trends in the area of Cloud computing QoS and provides a guide for future research. A review and survey of existing works in literature is done in order to identify these Cloud QoS trends. The finding is that the ultimate expectation of any QoS metrics or model is the related to cost concern for both the CSP and user.

Index Terms— Cloud computing, QoS, SLA

I. INTRODUCTION

“CLOUD computing is a model for enabling universal, on-demand and convenient 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” [1]. Provisioning of appropriate resources to cloud workloads depends on the Quality of Service (QoS) requirements of such workloads. Due to the increasing use of the cloud, the quality of cloud services has become an increasingly important issue since there are many open challenges which need to be addressed particularly those related to trust and availability. Cloud computing is engendering massive IT infrastructure development and utilization, and has become an avenue which provides scalable on-demand, elastic and metered services to cloud users. This has led to its widespread adoption both at an organizational and

individual level. Cloud computing offers three primary services: Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). In SaaS, major cloud providers offer software products to users over the Internet. Such application can be accessed at anytime and anywhere on a device. In PaaS, the cloud user is provided with an environment to create and deploy custom applications; thus, the users need not worry about infrastructural requirements and has control over applications being developed and deployed. In IaaS, compute resources and storage are offered to users at a fee. Resources such as CPU, memory, storage and network bandwidth are made available to multiple sharing users. This sharing of pooled resources by multiple users is called multi-tenancy. There are four cloud deployment models which are: private, public, community and hybrid clouds. Private clouds are owned by an organization and the staff are the only ones allowed to manage the private cloud. It is often administered by in-house experts or out-sourced to third parties. This model is considered the most secure. A public cloud is owned and operated by major CSPs who have massive cloud infrastructure which sometimes spread across continents. Several services are offered on-demand, in a multi-tenant and virtualized manner, usually at a fee. Public clouds are considered less secure. Community clouds is owned by several organizations with common interest sharing cloud computing infrastructure. It could be managed by the community or a third party. Hybrid clouds take advantage of the benefits offered by the private, public or community clouds. Hybrid cloud is a combination of these cloud types whose entities are unique, yet sharing cloud infrastructure. Cloud users are interested in the quality of service offered by Cloud Service Providers (CSPs). The method for providing services on the cloud requires some effort, because the CSP must determine the best hardware and software configurations that will be suitable in terms of QoS for the user, while at the same time ensuring optimum use of resources [2].

Cloud consumers want to select a cloud service appropriate for them from the major service providers that can offer services with adequate QoS guarantees [3]. Consequently, cloud services have become very attractive to businesses, but commercial offerings need to deliver the QoS expected by customers. If the services being offered do not meet user's expectations, they can seek alternative CSPs. The ability to specify the QoS is an important issue for users and providers alike [3]. On the Internet, QoS is based on transmission rates, error rates and other characteristics which can be measured, improved on and to some extent

Manuscript received December 27, 2017; revised January 22, 2018. This work was supported in part by the Covenant University through the Centre for Research, Innovation and Discovery (CUCRID).

I. Odun-Ayo is with the Department of Computer and Information Sciences, Covenant University, Ota, Nigeria. (+2348028829456; isaac.odun-ayo@covenantuniversity.edu.ng)

O. Falade is with Department of Computer and Information Sciences, Covenant University, Ota (adesola.falade@covenantuniversity.edu.ng)

O. Ajayi is from department of Computer Sciences, University of Lagos, Lagos, Nigeria. (email: olaajayi@unilag.edu.ng).

guaranteed in advance. QoS provides guarantee of availability and performance and provide a level of assurance that the resource requirements of an application are strictly supported [4]. QoS models are associated with customers and products which involves resource capacity planning through the use of schedulers and load balancers. The Service Level Agreement (SLA) provides an avenue for the customers and CSPs to agree on appropriate QoS levels which the customers are guaranteed of based on payments made. Cloud providers offer data and compute resource dynamically on the Internet based on user needs. It is thus important to maintain adequate QoS on order to ensure customer satisfaction. The purpose of this paper is to examine QoS in cloud computing. The paper will discuss QoS issues and highlight current trend in practice used to guarantee QoS to users. The rest of the paper is arranged as follows: section 2 discusses related work, while section 3 examines QoS metrics and models in cloud computing and the paper is concluded in section 4 with a suggestion for future work.

II. RELATED WORK

In [4], A QoS-driven approach for cloud computing addressing attributes of performance and security is proposed. The main focus of this paper is to examine quality of service, based on cloud security. The paper proposes a model that applies QoS to clouds with different security environments. In [2], Cloud based Video-on-Demand service model ensuring quality of service and scalability is proposed. The focus of the paper is the QoS for cloud storage of videos services. The traditional approach is to optimize performance, cost and some other parameters. The paper performs a characterization of start-up delays and used a modelling technique to arrive at a favorable conclusion. In [5], environment quality of service in cloud is proposed. The paper discussed cloud QoS in terms of cloud monitoring. The paper proposed a model that can be used as a guide for performance monitoring on the cloud. In [6], a mixed integer linear programming for quality of service optimization in Clouds is proposed. Examining QoS using certain criteria is important to cloud provider for optimization of service. The paper used two optimization algorithms on some QoS objectives to obtain various trade-offs. In [7], a quality of service based cloud resource provisioning called Q-aware is proposed. In the work, QoS was considered an important factor in the provisioning of resources by cloud providers and then proposed a QoS metric-based technique for analysis of workloads.

In [8], Quality of service approaches in cloud computing: A systematic mapping study is proposed. QoS is an issue that must be addressed properly to enhance trust in the cloud. The paper analyzed several QoS approaches to determine the area of more focus and suggested the way forward.

In [3], QoS in Software Defined Networking (SDN): A survey is proposed. Several solutions were suggested in terms of QoS. Relevant surveys were carried out in diverse areas for QoS, highlighting challenges and lessons. In [9], Secure and quality-of-service-supported service-oriented

architecture for mobile cloud handoff process is proposed. The paper focuses on QoS in terms of mobile cloud computing, energy and handoff issues. The paper proposes a four-layer model for energy efficiency and QoS in mobile Cloud computing. In [10], QoS in terms of virtualization on the cloud is discussed, while [11] focused on virtual machine provisioning based on analytical performance and QoS in cloud computing environments. It was deduced that among other things, workload, virtualization, and monitoring of applications affect performance on the cloud the most. Finally, a model for evaluating workload changes in applications to enhance QoS was presented. In [12], a PaaS architecture for real-time quality of service management in clouds was presented. The focus of the paper is QoS for real-time multimedia cloud applications. QoS metrics are applied at application and infrastructure levels to enhance resource provisioning. In [13], performance model driven QoS guarantees and optimization in clouds is presented. The focus of the paper was on optimization using different metrics for cloud applications. The proposed algorithm was used on different metrics to optimize QoS for a variety of workloads.

III. QUALITY OF SERVICE METRICS AND MODELS

QoS is the ability to provide different priority to different applications, users, or dataflow, or to guarantee a certain level of performance [14]. QoS is the totality of characteristics of a service that bear on its ability to satisfy stated and implied needs of the user of the service (service quality assurance). QoS is determined by the fulfilment of both functional and non-functional requirements. Meeting the user's requirement with regards to functionality will depend on the services description. The amount of non-functional services that has to be considered in cloud services is very high. Therefore, QoS parameter are considered to be related to non-functional properties of a cloud service [10] [3] [11]. Five key QoS attributes were identified in [15] and these are: reliability, flexibility, performance, security and usability

Reliability addresses system availability, fault tolerance, user experience levels, privacy and safety. Flexibility entails, system scalability, portability, interoperability among others.

Performance deals with system efficiency, response time, throughput and compliance to pre-agreed conditions of service. It often times provides metrics for drawing SLAs and measuring QoS compliance. This attribute is of high importance to both the users and the CSPs, as users are mainly interested in the response time, processing time or throughput of the applications running on top of cloud services, whilst using these performance metrics to rate the CSPs.

Security includes accountability, confidentiality, integrity, audit trails, etc. While usability focuses partly on user experience and value for money. Beyond these attributes, a major concern for cloud users and CSPs alike is that of resource consumption and techniques like monitoring utilization to identify over-provisioning or under-performance. Consumption pattern identification is a vital

step in ascertaining and maintaining certain levels of QoS. Certain relationships are essential to achieve this, which are:

- Relationship with SLA. The QoS attribute are usually specified in the SLA. The SLA is an adequate manner to specify the QoS guarantees, as it specifies the service level objectives to ensure that the delivered QoS meets the user expectation [17].
- Relationship with monitoring. The QoS attributes like response time or throughput have a strong variability and in order to implement the contract, these parameters need to be carefully controlled [17]. Consequently, continuous supervision of QoS attributes is necessary to honor SLAs by the service provider. SLA compliance is also often tested by the client. Monitoring of the various parameters of SLA is a common practice to ensure compliance with the negotiated terms. The monitoring of QoS agreements allows the customer to observe the behavior of the service and it is based on extracting metrics needed to make measurements of QoS [17]. For example, the server, applications, databases, and networks are monitored using IT technologies. The process associated with monitoring to ensure prevention, correction and control in QoS is essential to maintain proper balance between the benefits to the CSP and the satisfaction of the customer.

There are several reported efforts at trying to model QoS in clouds or to manage non-functional properties in an intelligent manner. The main objectives of these QoS models is to support user evaluation of the quality of cloud service being provided [17].

A. Metrics for Cloud Services Evaluations

Metrics to be utilized in QoS depends on the service features. According to [17], metrics that are associated with cloud services can be considered in terms of performance, economics, and security. Table 1 summarizes certain Cloud service features and their corresponding evaluation metrics. Table 1 is based on information drawn from various literature. Some aspects of these information is examined in subsequent paragraphs.

1) Performance Metrics

There are several options provided by CSPs dealing with enterprises on the cloud. Each decision has a different efficiency regarding performance, service latency and precision. Enterprises must know what their applications can do on the cloud and whether migration satisfies their goals [17]. There are issues such as response time required to process a demand. Throughput is how much transaction is possible over a period of time, while timeliness is capacity to process request in a timely manner [17]. Performance is relevant to response time, throughput and timeliness. For example, in terms of computation, metrics such as CPU load, floating point operation (FLOP) Rate and instant efficiency are important. Other features such as communication, memory and time also have their unique performance metrics [17].

2) Economic Metrics

Economic metrics is used to compare or check different costs of services because of the various option available [17]. Economics has been typically considered a driving

factor in cloud computing adoption. The economic aspect has two properties, price and elasticity. For example, in term of elasticity, metrics such as boot time, suspend time, delete time, deployment time and total acquisition time are essential [17].

3) Security Metrics

Information security is important to every enterprise on the cloud. The concept of multi-tenancy and virtualization makes cloud security important. Most enterprises must also apply relevant compliance regulations. Security concerns are numerous across the various services and deployment types. For example, in terms of data security QoS metrics such as secure socket layer (SSL) application, communication

TABLE I
 GENERAL FEATURE METRICS FOR CLOUD QOS

Features	Metrics
High Availability & Fault Tolerance	Flexibility [17]
	Backup & Disaster Recovery [3] [10]
	Response time [15]
	Service Constancy
	Hot standby / Live Migration [3]
	Fault Tolerance & Mean Time Between Failures (MTBF)
	Recoverability [17]
Efficiency	Resource utilization [2]
	Resource Allocation & Scheduling [17] [3]
	Delay & Response times [3]
Portability & Scalability	Interoperability [15]
	Modularity [15]
	Ubiquity
	Variability and Platform independence [15]
Usability	Operability
	Attractiveness [15]
	Learn ability
	User experience [15]
Modifiability	MTTC (Mean Time to Change) [17]
Sustainability	Environmental - Carbon emission & Green House Effect [18]
	Power Usage Efficiency (PUE) [18]

latency and audit ability are required [17].

4) General Metrics

There are QoS metrics relating to general features. For example, in terms of availability, the QoS metric of flexibility, accuracy and response time may be considered [17].

B. Quality of Service Modelling techniques

Even though the cloud has greatly simplified the capacity provisioning process, it possesses several novel challenges in the area of QoS management [13]. QoS denotes the level of performance, reliability and availability offered by an application and by the platform or infrastructure that host it [17]. QoS is fundamental for cloud users who expect providers to deliver the advertised quality characteristic and for cloud providers, who need to find the right trade-off between QoS levels and operational cost [17]. However, finding the trade-off is a difficult decision often exacerbated

by the presence of SLAs specifying QoS targets and economic penalties associated to SLA violations [13]. While QoS properties have received constant attention well before the advent of cloud computing, performance heterogeneity and resource isolation mechanism of cloud platforms have significantly complicated QoS analysis and prediction. There are two primary modelling techniques for interactive cloud service: workload modelling and system modelling.

- **Workload modelling:** This deals with the determination of rates of arrival of request and of the demand for resources such as CPU demands of an application in terms of the infrastructure and also considering the QoS as required by such workloads [15].
- **System modelling:** This aims at evaluating the performance of a Cloud system either at design time or at run time. Models are used to predict the value of specific QoS metrics such as response time, reliability and availability.

C. Cloud Workload Modelling

The definition of accurate workload models is essential to ensure good predictive capability for QoS models. Workloads modelling involves workload inference and workload characterization.

1) Workload characterization

Several studies have attempted to characterize the QoS on the cloud deployment environment. Through benchmarking, statistical characterization of empirical data are useful in QoS modelling to quantify risks without the need to conduct an adhoc measurement study. They are vital to estimate realistic values for QoS model parameters such as networks bandwidth variance, virtual machine start up times and start failure probabilities [15].

Having described the properties of the cloud deployment environment users are faced with the additional problem of describing the characteristics of the workloads processed by a cloud application. Black box forecasting and trend analysis techniques are commonly used to predict web traffic intensity at different times scales. Time series forecasting has been extensively used for web servers for almost two decades [15]. Auto regression models are quite common in applications and they are already exploited in cloud application modelling. Other techniques include wavelet – based methods, regression analysis, filtering analysis and kernel – based methods.

2) Workload Inference

The ability to quantify resource demands is a pre-requisite to parameterize most QoS models for enterprise application. Inference is often justified by over heads of deep monitoring and by the difficulty of tracking execution paths of individual requests. It is possible to estimate, using indirect measurements, the resource demand placed by an application on physical resources such as CPU requirements [15]. From the perspective of cloud provider and users, inference techniques provide a means of estimating the workload profile of individual VMs running on their infrastructure, taking into account hidden variables due to lack of information [15].

Regression techniques is a common workload inference approach which involves estimating only the mean demand placed by a given type of request on the resource [17]. The standard model calibration technique is based on comparing the performance metrics such as response time, throughput and resource utilization predicted by a performance model against measurements collected in a controlled experimental environment [15]. These methods exploit queuing theory formulas to relate the mean values of a set of performance metrics to a mean demand to be estimated such as CPU performance. Several other regression techniques have been proposed such as queuing network model, demand estimation with confidence approach, online demand estimation approach, optimization based inference techniques and dynamic estimation of CPU demands.

D. System Models

Several classes of models can be used to model QoS in cloud systems. They include the performance models, dependability models, Black-box service models and simulating models.

1) Performance models

There a performance models such as queuing systems, queuing networks and layered queuing networks (LQN). While queuing systems are widely used to model single resources subject to contention, queuing networks are able to capture the interaction among a number of resources and applications components [15]. LQNs are used to better model key interaction between application mechanisms such as infinite connection ports, admission control mechanism, or synchronous request calls. Modelling these features usually require an in-depth knowledge of the application behaviors. On the other hand, while closed-form solutions exist for some classes of queuing systems and queuing networks, the solution of other models including LQNs rely on numerical methods [15].

2) Dependability Models.

Petri nets, reliability block diagrams (RBD) and fault trees are probably the most widely known and used for dependability analysis. Petri nets are a flexible and expansive modelling approach, which allows a general interaction between system components, including synchronization of event times [15]. RBD and fault trees aim at obtaining the overall system reliability from the readability of system components. The interaction between components focus on how the faulty state of one or more components result in the possible failure of another component [15].

3) Black Box Service Models.

Black box service models have been used primarily in optimizing web service composition, but they are now becoming relevant also in the description of SaaS applications, IaaS resource orchestration and cloud-based business process execution [15]. Service models were described in terms of response time, assuming the lack of any further information concerning its internal characteristics such as contention level from concurrent request.

4) Simulation Models

Several simulation packages exist for cloud system simulation. Many solutions are based on the CLOUDSIM toolkit that allows the user to set up a simulation model that explicitly considers virtualized cloud resources, potentially located in different data centers as in the case of hybrid deployment CLOUDANLYST is an extension of CLOUDSIM that allows the modelling of geographically-distributed workloads served by applications deployed on a number of virtualized data centers [15].

The most important application of QoS model is optimal decision – making for cloud system management, such as capacity allocation, load balancing and admission control.

5) Capacity Allocation.

The infrastructure provider capacity allocation problem arising at the provider side involves deciding the optimal placement of running applications on a suitable number of VMs, which in turn has to be executed on appropriate physical servers [15]. The essence is for resource sharing that will minimize cost associated with energy consumption, while guaranteeing the SLA stipulated with a customer. The infrastructure capacity allocation arises in IaaS and PaaS scenario where the user is in charge, with control of a number of VMs or application containers running on the system [15].

6) Load Balancing.

Request load balancing is an increasingly supported feature of cloud offerings. A load balancer dispatches request from users to servers according to a load dispatching policy. This process is for infrastructure-provider load balancing [15]. In the infrastructure-user load balancing, the load balancer is installed and managed transparently by the cloud provider. A cloud user can also decide to install its own load balancer for cloud application. This may be helpful in jointly tracking capacity allocation and load balancing.

7) Admission Control.

The infrastructure-provider admission control is an overload protection mechanism that rejects requests under peak workload conditions to prevent QoS degradation [15]. The infrastructure – user admission control mechanism is used as an extreme mechanism, which is helpful when additional resources are obtained with some significant delay. For example, during a cloud burst, if the public cloud resources are not provided timely one can drop new incoming request to preserve QoS for users already in the system, avoiding application performance degradations [15].

IV. CONCLUSION

Cloud computing provides scalable, on-demand, elastic and metered services to cloud users over the Internet. There are various service types and deployment models to ensure that appropriate services are delivered by the CSP. In this paper, QoS as it relates to cloud computing was discussed. Various performance metrics were highlighted. Various Cloud QoS models and applications were also discussed. Finally, it can be concluded that QoS could mean completely different things when viewed from the CSPs and the users' perspectives, there should be a balance such that the CSP can minimize cost (maximize profit) by efficiently utilizing

resources while ensuring that their customers' (users) satisfaction is guaranteed. This means that the user must have a perception of "value" from the service being received and paid for through a CSP.

ACKNOWLEDGMENT

We acknowledge the support and sponsorship provided by Covenant University through the Centre for Research, Innovation and Discovery (CUCRID).

REFERENCES

- [1] Peter Mell, Timothy Grance "The NIST Definition of Cloud Computing", NIST Special Publication 800-145, 2011
- [2] Carlos Barba-Jimenez, Raul Ramirez-Velarde, Andrei Tchernykh, Ramón Rodríguez-Dagnino, Juan Nolasco-Flores, Raul Perez-Cazares, "Cloud based Video-on-Demand service model ensuring quality of service and scalability", Journal of Network and Computer Applications vol. 70, no. 10, 2016
- [3] Murat Karakus, Arjan Durresi "Quality of Service (QoS) in Software Defined Networking (SDN): A survey", Journal of Network and Computer Applications, vol. 80, pp. 200-218., 2017
- [4] Bruno Guazzelli Batista, Carlos Henrique Gomes Ferreira, Danilo Costa Marim Segura, Dionisio Machado Leite Filho, Maycon Leone Maciel Peixoto, "A QoS-driven approach for cloud computing addressing attributes of performance and security", Future Generation Computer Systems, vol. 68, pp. 260-267, 2017
- [5] Manjusha Kalekuri, Kolasani Ramchand Rao, "Environment Quality of Service in Cloud", 7th International Conference on Communication, Computing and Virtualization 2016, Procedia Computer Science, vol. 79 pp. 118 – 126, 2016
- [6] Tom Guérou a, Yacine Gaoua, Christian Artigues, Georges Da Costa, Pierre Lopez, Thierry Monteil, "Mixed integer linear programming for quality of service optimization in Clouds", Future Generation Computer Systems vol. 71, no. 1, 2017
- [7] Sukhpal Singh, Inderveer Chana, "Q-aware: Quality of service based cloud resource provisioning", Computers and Electrical Engineering, vol. 47, pp. 138–16, 2015
- [8] Abdelzahir Abdelmaboud, Dayang N.A. Jawawi, Imran Ghani, Abubakar Elsafi, Barbara Kitchenham, "Quality of service approaches in cloud computing: A systematic mapping study", The Journal of Systems and Software, vol. 101, pp. 159-179, 2015
- [9] Abdul Razaque a, Syed S. Rizvi, Meer J. Khan, Qassim B. Hani, Julius P. Dichter, Reza M. Parizi, "Secure and quality-of-service-supported service-oriented architecture for mobile cloud handoff process", Computers & Security vol. 66, pp. 169–184, 2017
- [10] Django Armstrong, Karim Djemame, "Towards Quality of Service in the Cloud", School of Computing, University of Leeds, United Kingdom.
- [11] Rodrigo N. Calheiros, Rajiv Ranjan, and Rajkumar Buyya, "Virtual Machine Provisioning Based on Analytical Performance and QoS in Cloud Computing Environments", International Conference on Parallel Processing. DOI 10.1109/ICPP.2011.17, 2011
- [12] Michael Boniface, Bassem Nasser, Juri Papay, Stephen C. Phillips, Arturo Servin, Xiaoyu Yang, Zlatko Zlatev, Spyridon V. Gogouvis, Gregory Katsaros, Kleopatra Konstanteli, George Kousiouris, Andreas Menychtas, Dimosthenis Kyriazis "Platform-as-a-Service Architecture for Real-time Quality of Service Management in Clouds", Seventh Framework Programme FP7/2007-2011, ICT2007.1.2., 2007
- [13] Jim (Zhanwen) Li, John Chinneck, Murray Woodside, Marin Litoiu Gabriel Iszlai, "Performance Model Driven QoS Guarantees and Optimization in Clouds", CLOUD '09 Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing, Pp. 15-22
- [14] Congduc Pham, "QoS for Cloud Computing", PIREGRID THEMATIC DAY, available at www.univ-pau.fr, 2011
- [15] Petcu D. (2016) Service Quality Assurance in Multi-clouds. In: Altmann J., Silaghi G., Rana O. (eds) Economics of Grids, Clouds, Systems, and Services. GECON 2015. Lecture Notes in Computer Science, vol 9512. Springer, Cham

- [16] Amid Khatibi Bardsiri, Seyyed Mohsen Hashemi, "QoS Metrics for Cloud Computing Services Evaluation", *I.J. Intelligent Systems and Applications*, vol. 12, pp. 27-33, 2014
- [17] Danilo Ardagna, Giuliano Casale, Michele Ciavotta¹, Juan F Pérez and Weikun Wang, "Quality-of-service in cloud computing: modelling techniques and their applications" *J. of Internet Services and Applications*, vol. 5, no. 11, available at www.jisajournal.com/content, 2014
- [18] Berl Andreas, Gelenbe Erol, Di Girolamo Marco, Giuliani Giovanni, De Meer Hermann, Dang, Minh Quan, Pentikousis Kostas. "Energy-efficient cloud computing", *The Computer Journal*, vol. 53, no. 7, pp. 1045–1051, 2010