International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 9 5666 - 5672

QoS-aware Storage Virtualization: A Framework for Multi-tier Infrastructures in Cloud Storage Systems

Mai Alfawair
Faculty of Information Technology
Al-Balqa' Applied University
e-mail: May_alfa3ori@yahoo.com

Omar Aldabbas
Faculty of Engineering
Al-Balqa' Applied University
e-mail: o.aldabbas@bau.edu.jo

***Abstract*—** The emergence of the relatively modern phenomenon of cloud computing has manifested a different approach to the availability and storage of software and data on a remote online server 'in the cloud', which can be accessed by pre-determined users through the Internet, even allowing sharing of data in certain scenarios. Data availability, reliability, and access performance are three important factors that need to be taken into consideration by cloud providers when designing a high-performance storage system for any organization. Due to the high costs of maintaining and managing multiple local storage systems, it is now considered more applicable to design a virtualized multi-tier storage infrastructure, yet, the existing Quality of Service (QoS) must be guaranteed on the application level within the cloud without ongoing human intervention. Such interference seems necessary since the delivered QoS can vary widely both across and within storage tiers, depending on the access profile of the data.

This survey paper encompasses a general framework for the optimal design of a distributed system in order to attain efficient data availability and reliability. To this extent, numerous state-of-the-art technologies and methods have been revised, especially for multi-tiered distributed cloud systems. Moreover, several critical aspects that must be taken into consideration for getting optimal performance of QoS-aware cloud systems are discussed, highlighting some solutions to handle failure situations, and the possible advantages and benefits of QoS. Finally, this papers attempts to argue the possible improvements that have been developed on QoS-aware cloud systems like Q-cloud since 2010, such as any extra attempts been carried forward to make the Q-cloud more adaptable and secure.

Keywords*-QoS-aware cloud systems; Q-Cloud; System design; Distributed storage system; Cross-tier Priority (CTP).*

*****

I. INTRODUCTION

Cloud computing is being marked as one of the most popular services being used by most companies in recent years, utilizing cloud as a storage form for multiple forms of data with various sizes and properties. Knowing that information is invariably considered one of the most valuable assets an organization can possess, cloud storage or an off-site online data storage is considered to have many benefits for organizations regardless of their sizes and nature, providing an extra storage capacity for seemingly endless data, or even more importantly, providing a redundant distributed system that may act as a failover system designed to keep their system running in case the primary system fails. Advantages of cloud storage will be explained in details later in this survey paper.

Most importantly, high performance, reliability, and availability are the three most important advantages that any user would obtain when using cloud storage. The main aim of this paper is to survey Quality of Service (QoS) provided in the context of cloud storage, and how it is managed and attributed to in cloud computing. In general, the main goal of QoS in any computing and networking service is to achieve user or customer satisfaction regarding a particular service; however, QoS in cloud computing is uniquely important due to the critical factors involved in providing an outstanding service, including reliability and security of the data stored. One can truly appreciate the importance of QoS in cloud storage when considering a company, for example, who has utilized cloud storage for their primary data after an agreement regarding the QoS, but the performance expected is not being met. As a result, such company cannot access all resources that they may require, and eventually, this would cost the company extra costs and stress, especially if this incident happens often. In practical terms, this reflects the need of a scalable, reliable, and high-performance storage infrastructure. Consequently, considering QoS and making sure that it is being met is very critical.

Considering the fact that cloud storage is a service that cloud providers may sell to their customers under certain circumstances, in order to persuade customers to use their service, such services must always be available and reliable. Cloud providers are charging their clients extra costs based on consumption and utilization, and that is an additional reason that QoS should be met, in order to provide customers the services that they have agreed on with the provider, or in other words, the services they have paid for. To provide such cloud services, technical issues always will emerge; such issues must be solved in order to keep the clients satisfied, not to mention

the creation of a good brand image that would depend on the QoS provided.

One of the techniques or services that ensure QoS is being met is Q-cloud. Q-cloud is a cloud service that is known for its QoS-aware control of the framework that alters the distribution and mitigation of the resources' performance. In order to achieve that, Q-cloud utilizes an online feedback to be able to build a Multi-input Multi-output model (MIMO) for capturing performance intervention to achieve an enhanced administration [1].

Also discussed in this survey paper the improvements of the QoS in the cloud services and Q-cloud since 2010; has there been any succeeding innovation in this specific technology? Did it improve since then? If yes, then what are the developments that have been emerging? What are the recent modifications and is there any new feature that might be supportive? And if there has not been any improvement, then, why not?

II. MOTIVATION

This paper is directed to reflecting on a recent research under the title (Q-Clouds: Managing Performance Interference Effects for QoS-Aware Clouds) to examine its critical features, exclusively concentrating on QoS within cloud services. As the research is pointing to, there are issues that must be addressed, since not guaranteeing the most important feature in cloud computing which is QoS, is relatively disappointing, but developing and introducing another service which is Q-cloud as a solution is extremely important [2].

Self-managing virtualization is a technology that is being utilized by cloud storage services to automatically allocate and migrate data throughout the storage system lifecycle directed by user-provided QoS hints. This may bring about certain issues, mostly when consolidating multiple client applications into multicore servers, by introducing interferences in performance between collocated workloads [3].

III. QOS IN CLOUD STORAGE SERVICES

A. *The Purpose of QoS*

Quality of Service (QoS) is the general performance of a computer network as seen by the specific users of such network. Accordingly, QoS delivers the ability to provide a set of priority rules for different applications or for different individual users, therefore, the QoS has to meet the standards set between the clients and the cloud providers, to guarantee the achievement of customer's satisfaction. QoS as an aspect has to be considered all the way from the designing phase of the cloud system in order to develop an environment where the QoS requirement is met [4].

As mentioned previously, cloud services are becoming one of the most popular known services that are being utilized by numerous organizations, thus cloud as a service is gaining more and more approval day after day by organizations and individuals, and by receiving more and more acceptance, more issues will arise; more management has to be considered [5].

B. *QoS: The Technical Issues*

In infrastructures involving cloud storage, there are certain important components that need to be considered in detail, such as application management and QoS performance. Considering the fact that cloud computing is delivered through virtualization, the existence of a client who attempts to access their data in a shared virtual machine, while another previous client attempts to access an application simultaneously would cause an interference issue, as shown in Fig. 1.

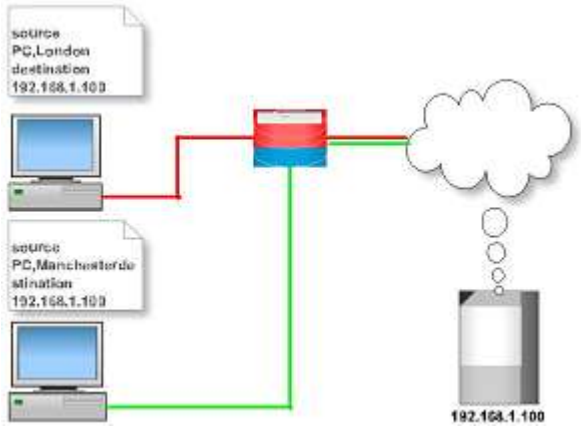

Figure 1. The concurrent attempt to access a shared storage media by two different clients.

As a consequence, the existence of multiple shared services in the cloud will lead to the creation of the issue of interference, where multiple users could attempt to access a certain allocated data concurrently at the same time.

If each client is given permission to access their data that is constrained to one machine only in the cloud without permitting another client to access the same parts of data allocated in the same machine, there would not be any interference issues since there would not be any form of data retrieval from multiple sources at the same time for the same machine. Unfortunately, such data isolation strategy does not operate in such manner. Instead, clients cannot control sets of data that are already existing at the same time, therefore, the cloud system cannot guarantee application performance and QoS performance.

IV. Q-CLOUD

A. *Q-Cloud: The Basics*

On the word of Nathuji et al, 2010 Q-Cloud is considered as a cloud storage service that controls framework to modify resources allocation in order to mitigate performance interference [2].

Q-Cloud makes use of a type of online feedback, for the purpose of constructing Multi-Input Multi-Output (MIMO) models. MIMO models help in detecting performance interference; in addition to that, MIMO is being utilized to allow the user application to specify multi-level of QoS as an application of Q-state.

On the contrary, Q-Cloud provisioning has not utilized sufficient resources dynamically in order to enable the increase in QoS levels for the system's efficiency to improve [6].

Consistent with S. Ferretti et al., the main goal of QoS-aware cloud storage architectures is to meet customer requirements, with such architectures possessing a functionality standard to manage resources in the virtual management environment related to the cloud application. Another very critical feature of this particular architecture is that elimination of resources is prioritized over provisioning; this is achieved by configuring the amount of resources dynamically being used on the cloud [7].

Management of virtual resources automatically for the purpose of cloud service presentation is, according to Chen et al., is mainly done to isolate application's QoS specifications from resources allocation and provisioning, utilizing the two-level architecture. Each distinct application must possess an application-specific decision module associated with it in order to study the appropriate QoS that is required by cloud users, with this particular module having a high-level performance in order to be able to make the most appropriate decision conceivable for the allocation and provision phases. QoS-aware cloud techniques should deliver the anticipated service to the user and the same proposed QoS level [8].

The previous outlines concerning QoS within cloud services are clearly indicating that QoS is critical and must be implemented in the most effective way possible, in order to provide the services that customers and cloud users require. Customer requirements refers to what was agreed upon between the customer and the cloud provider in terms of service specifications, and for which Q-cloud has been developed to tackle and solve such issues regarding QoS and client-specified cloud specifications.

B. *Why Q-Cloud?*

The significant challenge that faces any cloud service provider is QoS and application performance management that exists in the cloud infrastructure. A significant matter that might arise in the cloud environment is that application performance could alter due to the existence of other virtual machines in the cloud [6].

Q-cloud has been created with versatility to target a wide range of organizations since it might seem expedient for any business regardless of its size and requirements. Astonishingly, Q-cloud has developed a unique functionality with an ability of performing deduplication and backing up of data to the organization's devices or the cloud, knowing that the capacity of the backup is absolutely agreed upon formerly.

An access control that is recognised as "role-based control" is applied for provisioning cloud services, operating with VMware to reproduce data automatically performing data encryption when data is being transferred from source to destination. The entire process, seemingly intricate, is being done through deduplication and compression. The use of two very significant techniques, inactive data filtering and deduplication, is established for the assistance of replication where they are reducing disk capacity by huge amount to reduce costs on businesses [7].

C. *Interference in Cloud Organisation*

The central feature of any cloud platform is resource sharing and scaling, and in order to make resource sharing promising and running all the time, two forms of technology need to be recruited: virtualization and multicore processing. Virtualization is used to tackle fault isolation and manageability improvement, while multicore processing is designed to improve overall performance. However, not every accommodated workload may be a scale-up to use all cores in a server; consequently, it becomes necessary to combine several workloads so as to have an effective hardware utilization. Encapsulated in VMs, the workloads are assigned "Virtual Processors" (VPs) that are supported by individual cores, or sometimes even parts of individual cores [9].

As a result, unsatisfactory performance would occur at certain times, and might be considered as an issue that impedes the delivery of QoS-aware cloud services, but as Nathuji et al, 2010 suggested in their paper, this issue could be solved by making use of a quad-core processer capacity, which localizes consequences of caching interference [6].

It was proposed by J. Wei et al, 2013 in their paper that there would not be any sort of interference in case there is a machine dedicated to each individual client's application, allowing individual access to the network plus any storage

resources. In particular settings, similar to the existence of a heterogeneous hardware, it is common to ignore some sorts of interference, but ideally it is not a good practice in terms of overall performance, especially for the heterogeneous hardware resources that determine the QoS for applications. Hence, when there is no existence for any heterogeneous hardware, the QoS would not be as effective [11].

Similarly, when discussing cloud performance issues, one of the most common issues is the unpredictability and irregularity in performance within a virtual environment. Performance interventions between consolidated applications have been demonstrated for a variety of existing cloud systems and applications, and it has been proven that indiscretion in performance can lower the level QoS by providing inconveniency to the general users, which most of the time would prefer a consistent level of performance regardless of the QoS.

On the issue of performance irregularity, researchers have focussed on two main factors, the first of which is that during service loss due to the interference within the consolidated system, assigning a strategy called "sharing" can definitely improve the situation by such allocation strategy. The sharing allocation strategy works on a functionality that allows each individual Virtual Machine (VM) or workstation to expend up to 100% of its CPU usage, signifying that such strategy would yield an "anti-intuitive" result; consequently, the sharing allocation strategy seems as if it is averting the hypervisor scheduler from providing resource isolation between the consolidated VMs, as can be demonstrated in Fig.2.

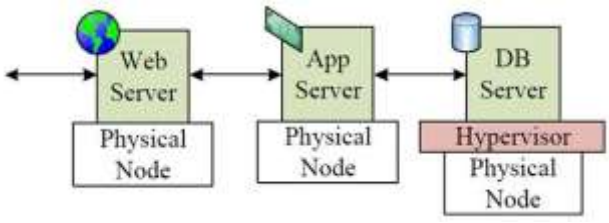

Figure 1. An illustration of the dedicated placement of one 3-tier application cloud system with three dedicated servers (A Web server, an Application server, and a database server) plus three physical nodes (Li et al, 2013).

On the contrary, the sharing allocation strategy outshines by permitting each individual VM to utilize its own CPU capacity, while other VMs remain idle, demonstrated in Fig.3. Such cases yield extreme effectiveness, especially when the utilization level is extremely high to the extent that VMs would have a queue request. For that reason, we can generally claim that a solution for the common issue of performance interference is revealed, since performance by response time can be measured through using Cross-Tier-Priority (CTP) based scheduling, discovering that CTP possess the ability to reduce performance loss due to the existence of an interference [12].

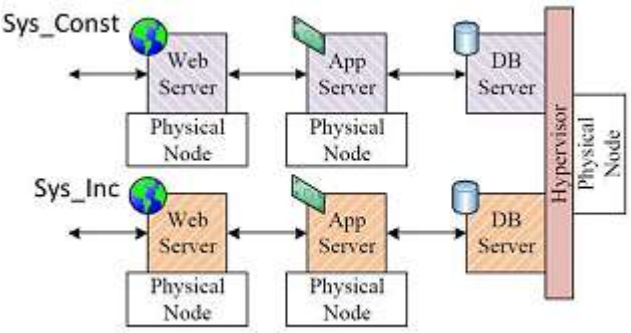

Figure 2. An illustration of the consolidated placement of the two 3-tier cloud systems with one server for each tier, plus five physical nodes as a sum. The database (DB) servers are co-positioned in specific VMs on a single shared physical node (Li et al, 2013).

'Bursty workload' constitutes the second factor that may induce performance interferences during an operation of consolidation. Such factor brought a very popular experiment of web facing application into reality, with the focus of researchers on quantifying the bursty workload to degrade the consolidation and its placements through measuring the respond time, while concurrently considering a reasonable and safe CPU utilization that is approximately 70% overall. Within this context, a negative influence exists owing to the lack of sensitivity of any VM's CPU allocation, affecting both the sharing and the isolation strategies that exist within the cloud system [12].

D. Mitigating Interference: Resource Partitioning

Enforcing partitioning within the shared system is done for the purpose of handling performance interference, the best example of which is the partitioning of Logical Link Control (LLC) within Multiple Virtual Storage (MVS). According to this method, it seems that the matter of performance interference is more likely to be resolved as can be seen in Fig.4.

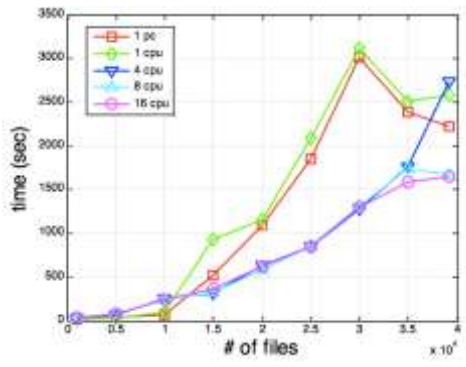

Figure 3. An illustration of the performance influence of interference involving cache on consolidated VMs (Nathuji et al, 2010).

Nevertheless, there are still certain issues that arise when implementing such approach, making such method not reasonable to use:

- Firstly, when implementing this approach, costs regarding the complexity of the system would add to the bill of the hosting organization, as it is obvious that there are several types of performance interferences, including a shared I/O and memory bandwidth. Hence, in certain scenarios additional software or hardware requirements need to be placed, adding to the overall cost, in order to mitigate certain interferences.
- Moreover, depending on the partitioning technique implemented, the occurrence of inefficient resource utilization is very likely to take place. For instance, in cases of allocating all of the colours present to two VMs, this will lead to the aversion of consolidation, unless dynamic page recolouring takes place.

As a result, regardless of hardware support for cache partitioning, there will still be unproductive cache utilization. This is due to the fact that storage capacities hardly used by one core could not be used by another owing to the strict partitioning [6].

E. Cloud Management Via Performance Service Level Agreements (SLAs)

Data request interference within the cloud environment is not a well-rated performance, therefore, the association between the resources that should be definite and the QoS that is set as the standard is getting to a radical state. In other words, such association might be demolished, leading to quite a few scenarios that should be well considered in order to quantify the amount that a client is charged and the aspects of metric that is agreed to be met as shown as shown in Fig.5.

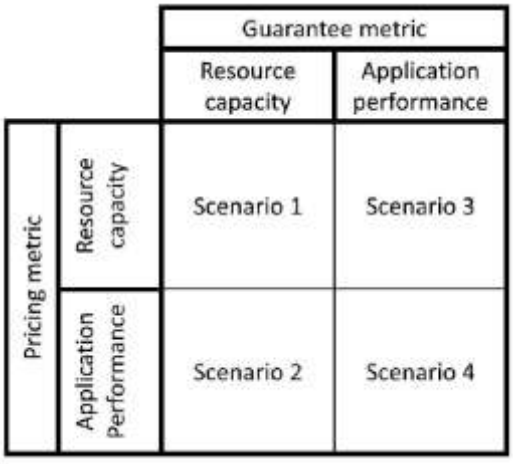

Figure 4. An illustration of the scenarios that result from the overlap of different pricing and guarantee measures for a cloud system (Nathuji et al, 2010).

Taking a look at the scenarios, the first possible scenario demonstrates a case where the entire performance depends on resource utilization, consequently, a reliable performance is not definite as far as the client is concerned, i.e. QoS is not met or ensured. Such approach is easy and simple to implement, and that is why it is preferred and widely used by cloud systems. Nevertheless, due to the presence of a performance interference, the module would become insufficient to operate by clients, principally owing to the existence of other capacities beyond the control of clients, causing serious technical issues.

The second scenario, where also QoS and the performance are still not guaranteed, demonstrates a probable advantage of price adjustment depending on the actual performance. As clearly stated, performance and QoS are not guaranteed so reasonably price adjustment would not be appropriately advantageous to the client since the performance is initially not as expected. As a consequence, the client might end up paying extra costs due to the lack of reliability or poor QoS of the cloud system, thus, such approach would not be expedient to utilize either.

The third scenario where the QoS is guaranteed yet pricing is a determinant in the need of additional resources; in such a specific case, terms and conditions of the cloud provider would most probably not benefit the cloud users when considering additional charges. Accordingly, and as demonstrated in Fig.6, this is yet another scenario where practical effective performance is not achieved as required.

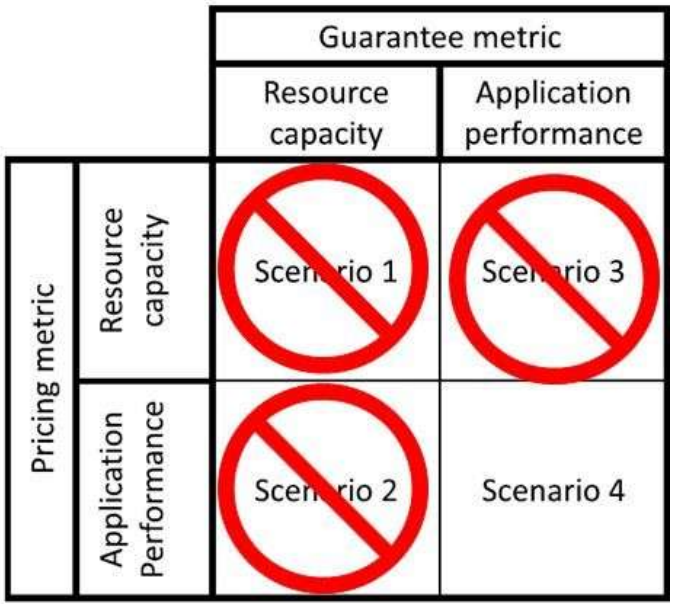

Figure 5. An illustration of the possibilities of performance success regarding the scenarios of the different pricing and guarantee measures for a cloud system.

Scenario four, being considered as the most standard and practical procedure to be implemented, has QoS that is ultimately guaranteed, with pricing being definite and not adjustable as well [6].

V. Recommendations and Suggested Improvements

This section discusses the role Q-cloud has taken to improve QoS-aware cloud storage according to schemes presented in a paper entitled (Q-Clouds: Managing Performance Interference Effects for QoS-Aware Clouds) [2].

There are issues that emerged preventing the delivery of an effective cloud service, or in other words delivering an agreed service with clients, being provided under high QoS. Such issues where performance interference occurs include two or more cloud users requesting resources from a common VM server. Thus, the outcome would be an interference that generally affects or even disturbs performance [6].

Resource partitioning, a method created to avert performance interference by means of page colouring, does not solve the issue of interference completely, since in non-dynamic page colouring the colours that are allocated to the different resources will eventually come to an end.

The implementation of Service Level Agreements (SLAs) in the management of cloud services logically leads to the presence of four different scenarios, providing the four types of QoS that the cloud provider could agree with the client on any of them. The first three scenarios discussed previously are not considered effective to be implemented by clients; however, the fourth scenario would be the best one in terms of the effectiveness of implementation [2].

In the year 2011, there was not any improvement to the status of research that was considered in 2010 regarding QoS-aware cloud systems, yet, there were still some arguments regarding QoS instead of Q-Cloud, which is a stream of thought discussed in this paper.

The year 2012 witnessed an improvement in the field since two types of methodologies to support the Q-Cloud in providing its services adequately have emerged: Q-Cloud Protect and Q-Cloud as a backup. Such additional services were considered as methods towards substantial adoption of cloud services by clients through utilizing cloud services even more. Other than that, there were not any major improvements that are worth discussing.

In 2013, a major accomplishment has been introduced, which is the adoption of the sharing allocation strategy instead of the consolidated cloud system. According to Li et al, 2013, the sharing allocation strategy permits up to 100% CPU optimization for each individual VM, while on the other hand, the consolidated cloud system permits only up to 50% CPU optimization to each VM. Such sharing allocation strategy is considered very effective, mostly at higher utilization levels when the VMs are in a queue request.

In theory, the key for resolving performance interference issues has been discovered by implementing the Cross-tier-priority (CTP) based scheduling, knowing that CTP can reduce performance loss [12].

VI. Conclusion

Cloud computing is a technological web-based phenomenon that became one of the mostly used services that are running on the global network. Cloud computing constitutes a major technological service that most organizations in recent days cannot operate without. This paper has explained the criteria that would be required to deliver cloud storage in a convenient QoS to provide apposite and favourable service to clients to utilize cloud services with technical methods and modules mitigating any sort of technical matter that would halt the delivery of the agreed service.

This paper aims to survey QoS considering how to achieve it in the required parameters within a cloud environment. QoS in the technical and networking environment is aimed to satisfy the needs of clients regarding a particular service. Nevertheless, this paper has described how QoS in cloud storage is even more important than in any other service, due to the critical role QoS would tackle and determine for the cloud service users. When tackling the performance expected by cloud users, if the QoS is not being met and is a reflection of a poor technical state, client dissatisfaction and additional unwanted costs to the hosting organization would be the end result, especially if the incident of poor QoS occurs often. Thus, considering QoS and pertaining that it is operating in an adequate level of performance is very critical.

The importance of service availability, reliability and security of the cloud service was previously referred to throughout the paper since, when considering cloud services, a cloud service provider is a business that needs to offer high quality services to attain a unique outstanding brand image compared to other competing providers in the market. At the present time, several general cloud providers exist such as Dropbox, Box, Google Drive, and many others, with customers of different cloud providers being charged according to the provider terms of use and the client's data usage. Hence, in order to provide clients the services that they are requiring and expecting, QoS need to be met. Undoubtedly, in technical environments like the cloud environment, issues might arise and for any issue that occurs, a solution can be put in place to eliminate such issues. The paper discussed some solutions for

such issues in order to provide the most apposite services required.

One of the implemented technologies that ensures QoS is being taken into the operational account is Q-cloud, which is a cloud service identified as QoS-aware control of the framework that carries mitigation and allocation for the resource's performance to ensure the required QoS level. Q-cloud operates an online feedback utility in order to build a Multi-Input Multi-Output model (MIMO) for the purpose of capturing performance interference, performing better cloud management as a result.

Improvements on the QoS in the cloud service environment and Q-cloud since 2010 have been clarified previously, yet several queries and questions remain regarding the subject matter, including: has there been any subsequent recent work in this particular field? Has there been any improvements since then? If yes, then what are these improvements and what do they include? What are the current changes in the cloud storage field? If there are no recent improvements in the field, then why not? Is technical work redirected to another field?

ACKNOWLEDGMENT

We are indebted to thanking our university BAU for their continuous support and dedication to the completion of this research. Their assistance proved to be a milestone in the accomplishment of this paper.

REFERENCES

[1] Al-Shehri, S. and Li, C. (2014). "Quality of Service for Cloud Computing." AMR, 905, pp.683-686.

[2] Nathuji, R., Kansal, A. and Ghaffarkhah, A. (2010). "Q-Clouds: Managing Performance Interference Effects for QoS-Aware Clouds." 1st ed. paris.

[3] Buyya, R., Yeo, C., Venugopal, S., Broberg, J. and Brandic, I. (2009). "Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility." Future Generation Computer Systems, 25(6), pp.599-616.

[4] Wang, X., Du, Z., Xue, Y., Fan, L. and Wang, R. (2012). "Self-adaptive QoS-aware resource allocation and reservation management in virtualized environments." International Journal of Computational Science and Engineering, 7(4), p.308.

[5] Singh, S. and Chana, I. (2014). "QRSF: QoS-aware resource scheduling framework in cloud computing." The Journal of Supercomputing.

[6] R. Nathuji, A. Kansal, and A. Ghaffarkhah, "Q-clouds: managing performance interference effects for qos-aware clouds", in Proc. European Conference on Computer Systems, EUROSYS 2010, pp. 237–250.

[7] S. Ferretti, V. Ghini, F. Panzieri, M. Pellegrini, and E. Turrini, "Qos–aware clouds," in Proc. Cloud Computing, CLOUD 2010, pp. 321–328.

[8] CHEN, L., QI, W. and QI, Y. (2012). "Atmospheric monitoring network system based on cloud computing." Journal of Computer Applications, 32(5), pp.1415-1417.

[9] Chen, T., & Bahsoon, R. (2013, May). "Self-adaptive and sensitivity-aware QoS modeling for the cloud." In *Proceedings of the 8th International Symposium on Software Engineering for Adaptive and Self-Managing Systems* (pp. 43-52). IEEE Press.

[10] HUANG, J., LIU, Y., YU, R., DUAN, Q. and TANAKA, Y. (2013). "Modeling and Algorithms for QoS-Aware Service Composition in Virtualization-Based Cloud Computing." IEICE Trans. Commun., E96.B(1), pp.10-19.

[11] Jenn-Wei Lin, Chien-Hung Chen, and Chang, J. (2013). "QoS-Aware Data Replication for Data-Intensive Applications in Cloud Computing Systems." IEEE Transactions on Cloud Computing, 1(1), pp.101-115.

[12] Li, Y. (2014). "QoS-Aware Dynamic Virtual Resource Management in the Cloud." AMM, 556-562, pp.5809-5812.

[13] Lin, J. W., Chen, C. H., & Chang, J. M. (2013). "QoS-aware data replication for data-intensive applications in cloud computing systems." *Cloud Computing, IEEE Transactions on*, *1*(1), 101-115.

[14] Tang, X., & Xu, J. (2005). "QoS-aware replica placement for content distribution." *Parallel and Distributed Systems, IEEE Transactions on*, *16*(10), 921-932.