I.J.Modern Education and Computer Science, **2012, 10, 59-68**
Published Online September 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2012.10.08

Identifying Key Challenges in Performance Issues in Cloud Computing

Ashraf Zia
Department of Computer Science, Abdul Wali Khan University, Mardan, KPK, Pakistan
Email: ashrafzia@awkum.edu.pk

Muhammad Naeem Ahmad Khan
Department of Computing, Shaheed Zulfikar Ali Bhutto Institute of Science & Technology, Islamabad, Pakistan
Email: mnak2010@gmail.com

***Abstract*—** Cloud computing is a harbinger to a newer era in the field of computing where distributed and centralized services are used in a unique way. In cloud computing, the computational resources of different vendors and IT services providers are managed for providing an enormous and a scalable computing services platform that offers efficient data processing coupled with better QoS at a lower cost. The on-demand dynamic and scalable resource allocation is the main motif behind the development and deployment of cloud computing. The potential growth in this area and the presence of some dominant organizations with abundant resources (like Google, Amazon, Salesforce, Rackspace, Azure, GoGrid), make the field of cloud computing more fascinating. All the cloud computing processes need to be in unanimity to dole out better QoS i.e., to provide better software functionality, meet the tenant's requirements for their desired processing power and to exploit elevated bandwidth.. However, several technical and functional e.g., pervasive access to resources, dynamic discovery, on the fly access and composition of resources pose serious challenges for cloud computing. In this study, the performance issues in cloud computing are discussed. A number of schemes pertaining to QoS issues are critically analyzed to point out their strengths and weaknesses. Some of the performance parameters at the three basic layers of the cloud — Infrastructure as a Service, Platform as a Service and Software as a Service — are also discussed in this paper.

***Index Terms*—** Cloud Computing, Key Challenges, Performance Issues, Quality of Services

I. INTRODUCTION

The cloud is a set of hardware components, network devices, storage space, software solutions and interfaces that enable the distribution of computing as a service. These solutions are provides to user on-demand and include the transmission of software, infrastructure and storage service over the Internet. The consumers need not to know anything about the actual technology behind cloud computing services. For example, in small companies, the cloud vendor becomes *de facto* information center. Cloud vendors are expected to proffer a foreseen and assured assistance level coined with assurance of complete information security to their customers. Hence, cloud service providers are solely responsible for deploying, managing and maintaining IT assets. Large many companies are discovering new avenues and windows of opportunities in cloud services.

Overall, the cloud demonstrates four primary characteristics: flexibility (elasticity) and the capability to range (scale) up and down; self-service provisioning and automated de-provisioning; program development interfaces (APIs); and, charging and metering of assistance utilization in the form of a pay-as-you-go model.

II. LITERATURE REVIEW

Salvatore et al. [1] worked on Service Level Agreements over Quality of Services (SLA-QoS) that outlines procedure to provide SLA based QoS on hazard and risky cloud environment. In these approaches, resources put on cloud that involve cost factor are in fact the mistreatment of consistent underlying infrastructures. For this very reason, another scheme is needed that can help reduce cost and improve QoS. In this context, a new scheme that uses an entirely diverse approach is known as Cloud@Home (C@H) [2]. C@H makes available computing and storage services on volunteer and unpaid basis. Since such a system operates on volunteer basis, therefore, the QoS is ensured through network reliability and resources consistency. The volunteer approach makes the scheme more reliable because it encompasses multiple services and every service is available in enhanced and improved condition. The role of C@H provider (aggregator) is to bring together the available resources, amalgamate dissimilar technologies and apply diverse administration strategies. Most of the management tasks are performed through Resource Management Module and C@H users are assured of QoS by SLA management module. For this purpose, negotiation, monitoring, recovery and termination activities are performed for better QoS.

Yanping et al. [3] propose dirichlet multinomial model based scheme for reducing the overall cost of computing in cloud and maintaining good QoS. The proposed system is based on statistical probabilities, whereas, previous approach were based on means response time. The

selection of services is based on some criteria, known as metrics. The key metrics used include service cost, response time and reputation to make resource provisioning. The key idea in this regard is to depict QoS metrics (service cost, response time and reputation) with notations and variables. Cautious design of cloud and the selection of specific cloud for specific environment enhance the reliability and increase the level of fault tolerance.

Jing et al. [4] suggest a decomposition based matrix multiplication scheme for addressing reliability and fault tolerance issues. The matrix multiplication approach can help analyze different tasks and their behavior on different clouds. The idea mainly focuses on scientific computation for cloud and is mainly based on the pretext that complex scientific computations are rewind for large matrix multiplications. The proposed scheme can also find out faulty clouds in an intelligent manner.

Xiong et al. [5] suggest a queuing network based model for analyzing the working level and throughput of cloud resources. Laplace transformation is used to determine the response time distribution of with particular reference to resource utilization in a cloud. Concentrating on the following three characteristics is prerequisite to improve performance of a typical cloud: firstly, the level of QoS needed for specific services; secondly, the resources required for maintaining response time given the number of customers/users; thirdly, the number of customers supported by a specific resources. In this regard, probability distribution function can be used to determine clear estimation of resources viz-a-viz response time.

Lee et al. [6] suggest assessing SaaS in a quantifiable method. For instance, different vendors can judge the level of various services and can calculate return on investment (ROI). Additionally, quality assessment of SaaS can be performed by obtaining feedback from the service users, the number of subscribers and the way users consume SaaS. Such evaluations can indicate the level of quality of SaaS. The fundamental feature of the cloud computing is to reuse various services over the Internet. SaaS itself is a reusable component for the clients. Scalability is another key feature of SaaS and it is the core responsibility of the provider to rescale the resources.

Jeyarani et al. [7] propose a Virtual Machine scheduler, that utilizes the technique of meta-scheduler and backfill strategy, for performance optimization. Inter VM scheduler is implemented at the host level to balance the load and enhance VM Provisioner for optimized utilization of the resources. User can choose the most appropriate resource for the meta-scheduler to perform the essential jobs and the VM scheduler at the system level sends jobs for execution by optimally utilizing the best available resources. In addition, the inter VM scheduler implementation can capacitate the host machines for adaptive balancing of load.

Wang et al. [8] suggest an automatic optimization schema for cloud storage, called AOSC, which uses data chunking, placement and replication to achieve more stable and foreseeable performance. The authors report that experimental results based on Amazon Elastic Cloud Computing (EC2) data center demonstrated that AOSC could enhance the stability and help provide better performance guarantees for cloud computing.

Younge et al. [9] propose a concept of Green Cloud structure for increasing efficiency per watt within a Cloud. The proposed framework utilizes power-aware scheduling techniques over an exclusive virtual machine design. This facilitates enhancing system performance within a data center based cloud with little overall performance expense. However, the design of the VM images can also lead to a serious cost-benefits predicament.

Olson et al. [10] report the performance of cloud computing servers in cyber-physical systems (CPS) that aggregate data stored over distant locations that are spatially apart to a great extent. The proposed model is exhibited through a Community Seismic Network (CSN) comprising various sensors that provide an early warning in the event of earth shaking. CSN uses Google App Engine because of its ability to scale up in lesser amount of the time by using nominal resources.

Liang et al. [11] recommend employing RandTest technique, a probabilistic dataflow integrity checking method, for integrity attestation technique. Since many processes are implemented on different servers in multi-tenant cloud computing, therefore, RandTest can serve as a very economical and useful solution.

Du et al. [12] suggest an attestation method to validate the dataflow processing integrity in cloud infrastructures. The idea is to use consistency information among the processing components to build attestation graphs. For this purpose, per-function consistency graphs and global inconsistency graph can be used when most of the nodes are infected as these graphs can easily detect the infected nodes.

Haiyang et al. [13] present a hierarchical scheme for evaluating Quality of Experience (QoE) for Open Source Publishing in the cloud environment and ubiquities computing. The hierarchical model collects all the main components of cloud and identifies interaction between them. The hierarchical approach is built on four sub-models: cloud computing availability model, outbound bandwidth model, a latency model and a cloud computing response time model.

Luo et al. [14] present a SOA model for cloud computing that exploits the value-at-risk method to develop different mechanisms and utilizes a set of considerable metrics. The SOA register-discover-invoke triangular architecture has been extended by adding a role of insurer who communicates with the customers and service providers. It is worth mentioning that value-at-risk method is quite useful to measure valuation risk

potential and can be used for service insurance premium calculation and reimbursement.

Jha et al. [15] identify different parameters of QoS with respect to cost-effectiveness and present an architectural framework for data as a service. The analysis carried out as a result of this study is based on the hardware and network resources with particular reference to ascertaining how these resources can utilize the multimedia resources more efficiently.

Iosup et al. [16] elaborate MTC (Many-Task Computing) based cloud computing performance by analyzing it with the current computing environment. The study reports that the current cloud paradigm is limited for scientific work due to insufficient performance and cost involved for performing real time scientific work.

Yang et al. [17] highlight job scheduling algorithm in cloud computing that exploits utility based computing framework to distinguish every job on the basis of Time Utility Function that bears monetary attribute known as utility.

Vishwanath et al. [18] defines server repair/failure rates to be able to comprehend the hardware reliability for huge cloud computing infrastructures. This report is the first make an effort to research server breakdowns and hardware components fixes for large data-centers. The author has presented a specific research of failing features as well as a research on failing predictors. The author discover that (similar to others) hard disks are the best suitable replacing element, not just because it is the most major element but also because it is one of the least efficient. The author has discover that 8% of all servers can anticipate to see at least 1 hardware event in a given season and that this variety is greater for devices with plenty of hard disks. It can be estimated the IT price due to hardware repair for a huge data center (> 100,000 servers) to be over a thousand dollars money. Furthermore, upon seeing a failing, the possibilities of seeing another failing on the same server is high. It is found that the submission of subsequent failing on a machine suits an inverse contour. Our preliminary speculation is that upon seeing a failing machine it makes a changeover from a harmless condition to a new condition where there is a rich framework in failing patterns. Moreover it has been found that, place of the datacenter and the maker are the most powerful signs or symptoms of breakdowns, versus age, configuration etc. Being a data study there a number of restrictions in this research. For example, we can only review based on the time period we notice. What this means is that the results are possibly one-sided against the environmental circumstances, technology, amount of work, features etc. prevalent during that interval. The main strength of this paper involves server restoration/catastrophe charges to grasp the computer hardware durability for giant cloud computing infrastructures. This is the first perform to assess the connection between successive breakdowns on the same machine. The first predictive discovery in a datacenter to mine for aspects that describe the purpose behind failures. Main weakness of the paper is just a review can only be performed based on the time frame we notice. What this means is that the results are possibly one-sided against the ecological circumstances, technology, amount of work features etc. prevelant during that interval. The factors are not examined, that caused the system to fail or even the time.

Lee et al. [19] examined the reliability of work-flows performance in the perspective of organizing and its effect on managing costs in Distributed Computing Systems (DCSs), and presented the reliability for profit assurance (RPA) algorithm as a novel work-flows scheduling heuristic. The suggested algorithm features a (operating) cost-aware duplication program to increase reliability. The development of cost attention significantly plays a role in effective duplication choices with regards to success. The work in this report is the first attempt to clearly take into account (monetary) reliability cost in work-flows organizing. The duplication program successfully determines replicability taking into account duplication cost, charge (incurred by deviating SLA targets) and failing features. The development of cost attention significantly leads to careful duplication choices in terms of success. Note that, RPA is general to deal with both interdependent projects of work-flows programs and individual separate tasks, although work-flows programs are essentially the best application model RPA can make use of for cost performance. The already complex work-flows scheduling problem in DCSs has been revisited from the point of view of reliability - for cost performance, and RPA including a novel duplication program has been setup. Replication is an attractive solution for reliable work-flows performance. However, the natural cost of duplication needs to be clearly resolved since replications have direct operating expenditures for their repetitive source utilization and oblique expenditures for service declines due to source unavailability due to such repetitive source utilization. RPA effectively details this price connotation of duplication taking the trade-off between duplication expenditures and success when making duplication choices. We have authenticated that this cost-aware duplication is a user-friendly cost enhancement technique for work-flows performance in DCSs. It demonstrates how to validate the cost-aware replication by using an intuitive cost optimization technique for workflow execution in DCSs. RPA successfully details the cost connotation of duplication taking the trade-off between duplication costs and success when making duplication choices.

Yuan et al. [20] focused on the management of resources in Cloud computing which is a major issue to be concerned. Most of the resources has low usage and high management cost due to which resource are consumed unusually. The technique used by the author utilizes several strategies including resource pre-reservation and resource borrowing. A standard shared resource model is constructed on the cloud. Main features

of the system are; routers are used to connect the clients and physical location of the tenants remains hidden. Pre-reservation strategy of the resources is used effectively to allocate and provide consumers, complex application request within a specific amount of time and also ensure clients about the requirements of the SLA and QoS. Because of the resource pre-reservation and allocation reserved resources cannot be simultaneously used by another user because it has already been reserved and those resources cannot be assigned again until they are de-allocated. The results of the experiment shows that the proposed method is an efficient method for the management of resources inside cloud computing. Because of the allocation and pre-reservation only one user can reserve the resources at a time.

Litoiu et al. [21] recommends a supreme model for the use of competent power and computing resources in cloud computing environments. It shows that how optimization of resource distribution can be able to achieve considerable cost decrease. It also emphasizes the authority that needs to be in place and the steps guarantee well-organized optimization problems. The author takes into account changeable workloads and suggests a new optimization technique along with a Service Oriented Architecture (SOA) governance approach to cloud optimization. There is a solid industrial and academic attempt for developing such a general SOA governance model. So, the emphasis is being given to the application of this model to cloud computing and resource optimization in cloud. Main strength of the paper is focus on the optimization of the cloud resources by efficient assignment of the resources which results in the reduction of the overall cost. Furthermore, this paper highlights a new system enhancement technique with SOA governance method to optimize the cloud. Numerous barriers are to be encountered in implementing the adoption of the proposed models. Instead of having several number of SOA Maturity Models there is no standard that one should be followed due to the dynamic size of the organizations. Various parameters like organizational structure, culture and education has to be kept in mind while adopting the presented scheme.

Wang et al. [22] has suggested a competent distributed metadata management scheme in cloud computing. Through different techniques it can convey high performance and scalable metadata services. The metadata server (MDS) cluster is usually adopted for the management of metadata and carrying out security strategies in a distribution system. One of the metadata management strategies is the Dcache strategy that can meet a number of goals. Firstly, Dcache takes on cooperative double layer of cache to mask slow disk performance. Secondly, by an IChord Dcache can make metadata service performance tremendously balance with the number of metadata servers in the cluster. Thirdly, Dcache is a flexible management startgey and through the adoption of an IChord can meet the addition, removal and replacement of metadata servers. It also proposes a familiar algorithm in order to control the client cooperating with server to cache metadata set. Fourthy, all the clients can easily get the uniform view. Fifthly, Dcache provide fine-grained load-balancing as it distributes metadata to multiple MDS's. Finally, in Dcache by mixing up an unchanging DPID, no object metadata migration happens when the object or its ancestor directory is renamed. The results of the discussed strategy were found to be encouraging. This paper described distributed meta-data management strategy Dcache and given the study of the efficiency evaluations, and the efficiency results are motivating.

Li et al. [23] presented an approach to find optimal deployments for huge data centers and clouds. It applies a combination of bin-packing, mixed integer programming and performance models in order to make the taken decisions affect the various strongly working together goals, which will include the pleasure of different service level harmonies for many different applications. The important thing is that it is scalable and extendable to new objects. This approach has proved to be practical in the sense of giving a solution within two minutes for the deployment of applications adding up 100-200 processes plus some models. It can optimize the energy use or the financial cost including the license cost that shows elasticity in addressing further concerns. CloudOpt is an effective and scalable algorithm for optimizing deployments in cloud. Practical results have been shown which provides a solution in under two minutes deploying various application ranging from 100 to 200 process with some replicas. It awards restrictions on individual QoS and process memory, and can boost power use or financial cost. CloudOpt can be created still more scalable with a more effective efficiency model solver.

Alhamad et al. [24] proposed the definition provided by V.S. NIST (National Institute of Standards and Technology) which basically describes cloud computing has been given importance. In this study a number of experiments on Amazon EC2 cloud has been carried out at different time. Each time the response was judged. The main focus was the testing on the isolation across the same hardware of virtual machines that are hosted by a cloud provider. According to this article, the performance of CPU was used as the main parameter for cloud performance. The execution time of the deployed application over five types of Amazon EC2 instances was measured. The response was being recorded every two hours during many days of experimentation. These cloud solutions are becoming popular regarding allocated technology because they allow cloud customers to release well-specified sources of computing, network and storage facilities. So, consequently the customers have pay for their use of solutions without requiring to spend a lot for incorporation, servicing, or management of the IT facilities. As a outcome this need a reliable statistic technique of the reliability. This article develops a performance metrics for the measurement and comparison of the scalability of the resources of virtualization on the cloud data centers. Firstly, the need

for a reliable method of comparing the performance of the cloud services has been offered. Secondly, a different type of metrics has been proposed for suitable measurement of the scalability. The focus is being given to the visualization resources such as CPU, storage disk and network infrastructure. Ultimately, a comparison is been made between the well-known cloud providers and the proposed approach. This research as a result will help out a cloud consumer to discover the capacity and ability of a specific service before signing any kind of official contract. The main strength of this document has provided that the efficiency of Extra-large high CPU like EC2 VM, has the best balance of efficiency. Therefore, service level agreement, response time can be used as a good parameter in the contract. Only EC2 instances have been tested. Google, Salesforce and Rackspace etc instances can also be tried.

Assunção et al. [25] present an investigation that has been done regarding the benefits, that different organizations can gather by using cloud computing providers capacity of their local infrastructure for the improvement of their performance upon the requirements of its users. Six different organizing techniques have been analyzed that are considered to be suitable for a local cluster group which are handled by virtual machine centered technological innovation for the enhancement of its service level agreement (SLA) with customers. The programs aspired to use distant sources from the cloud to work out the potential of the local cluster group. So, the focus is been given to the research of the scheduling techniques which give significance to the use of sources from cloud to be able to comprehend how these techniques accomplish a stability between efficiency and utilization cost, also how much they display enhancement in the request's reaction periods. For upcoming, it has been organized to study the efficiency of the different types of programs, such as bag-of-tasks or SPMD which will be running on the local cluster group, on the cloud vendor and both at the same time. Including to it, currently working is been performed on an flexible adaptive strategy technique which is designed to boost organizing efficiency considering the customer's funds. The main strength of the paper are the tests which analyzed the cost of helping the overall performance under different strategies for organizing needs on the company group and the Cloud provider. Trial results revealed that the cost of improving the efficiency of application scheduling is higher under a situation where the site's cluster is under used. Weak point of the paper is this that local clusters have not been targeted for consideration.

Sekar et al. [26] proposed that cloud computing offers its users the possibility to reduce operating and capital expenses. This article states that for cloud computing as a model to be sustainable in the long term, a systematic approach for confirmable resource accounting shall be needed. Conformability according to this article means that the cloud customer can be assured that their purposed indeed physically consume the resources they were charged for and that this consumption was necessary on an agreed policy. So, as a primary step toward discussed vision the challenges and opportunities for the realization of such framework are expressed. The author has defined the problem of supportable resource accounting and at the same time also threw light on the challenges and all the possible choices. In order to realize this vision in practice it has been acknowledged that the given results are just the beginning points and several other practical issues still remain unresolved. But it has been planned to tackle these issues as part of a reference implementation in future. This paper described the problem of verifiable resource accounting and outlined the difficulties and potential solutions to realize this perspective in exercise. Several practical issues of efficiency effect and expense of the monitoring framework, random leakage of private information) remain uncertain.

Wang et al. [27] has emphasized on the improvement of the energy efficiency of the servers through suitable scheduling strategies therefore a new scheduling model having energy-efficient and multi-tasking based on MapReduce has been introduced. This model consist of five sections. A genetic algorithm has been designed for the solution of the model and various experiments done on the model and its effectiveness and efficiency has been presented. For the solution of the mentioned model an applied method of encoding and decoding has been designed for the entities. For the enhancement of searching ability of the algorithm and the acceleration of the convergent speed a local operator has been introduced which as a result turned out to be a very effective and efficient. Using simulations to assess the stability and capacity of cloud computing systems this article presents an imitation for a cloud computing environment. It helps in the evolution of the cloud's logical stability under different configurations without performing any experiments on the real cloud environment. The rightness of this simulation is confirmed by the theoretical calculation result of the well-known M/M/1 queuing system. The mentioned stability assessment can also help one in finding the actual service rate. It is seen that there is a restricted parameter of the number of message types i-e two currently but it is not hard at all to extend the accommodation of more types of messages. As compared to the other computing techniques in the distributed environment such as Grid computing which needs many simulators has also been proposed but this technique does not require many efforts for its stability analysis. However, it is based on the analysis of a global system that tries to simulate most aspects of a cloud computing system. So, the main focus given in this article is a special issue that may seem smaller but it is applicable to many cloud computing system because of the use of the well-known theory of M/M/1 queuing system for the verification of its correctness. There are a number of possibilities for future exploration by the use of such simulation of cloud computing environments. This paper mainly concentrates on how to improve the energy-efficiency of servers through appropriate scheduling techniques based on MapReduce, a new energy-efficient

multi-task scheduling model. The experimental tests showed that the suggested algorithm is efficient and effective.

Jha et al. [28] gives a concept about minimizing the rising IT cost with the help of cloud solution. Along with it an architectural structure known as the video on-demand as an assistance , data as a assistance and speech assistance has also been suggested. The content describes OPNET MODELER as the best exercise strategy to an efficiency evaluation with cost (PCWS) for QoS Program in on–demand cloud computing. The significant elements of the structure has been outlined and its presenting execution with OPNET MODELER has been mentioned. The described framework clearly gives concern to network and hardware sources that are available and necessary in order to flow media services. Through the simulator outcomes of four different circumstances it is proven that video can be provided relatively quick in a legacy network. While in a toughest situation greater delays in the situation of first 30 minutes the response time reduces. Through these setbacks it can be said that cloud computing could certainly be an achievement but still there are justifications to be confronted despite the present growth of the power and network technology. The writer also declares the fact about the great need of study for the variation of conventional video on-demand IPTV service to cloud computing. Currently, research is been done for finding out the likelihood of applying the estimated architecture framework in the full cloud environment. This paper addressed the issue of on-demand cloud framework for demand video and explicitly proposed a framework that take network and other available hardware resources into account for streaming multimedia services. Several simulations have been performed to analyze the performance. Simulation results indicate that if number increases the response time remains constant and starts the video transmission early. Weakness of the paper is this that the simulation results do not consider different parameters like time taken for managing hardware resources i-e virtualized hardware resources. In actual implementation in virtualized hardware resources can take some time.

Bein et al. [29] has presented a problem that has been studied is the allocation of the memory servers in a data center based on online request for storage. Two efficient algorithms has been used for the selection of minimum array of servers and of the minimum overall cost. The result shows that both algorithms perform almost the best possible in case requests having entirely random values. The algorithms has been customized i-e HARMONIC m and CCHk inorder to deal with those storage space request that are bigger in number than the memory space of 1 server. Both the methods HLR (HARMONICm with Huge Requests) and CCHLR (CARDINALLITY CONSTRAINDED HARMONICk with Huge Requests) are proven to execute very well in exercise exclusively with the first half produced at unique and the second calculated from the first. For upcoming work a undertaking has been organized to examine the situation where requests on the internet are provided by two or more servers that share the bins. Another place of analysis would be the concern of the price of saving the huge item on a server reliant on the variety of already saved stored items on that specific server. Main strength of the paper is this that author has changed the algorithms of HARMONICM and CCHk (CARDINALITY CONSTRAINED HARMONICk) to deal with storage space needs that are bigger than the storage space of 1 server. The additions of the methods HLR (HARMONICMwith Huge Requests) and CCHLR (CARDINALITY CONSTRAINED HARMONICk with Huge Requests) have the same approximation rate as the unique methods and are proven that execute very well in exercise for the type of series in which the first 50 percent is produced at unique and the second is calculated from the first. Weak point of the paper is investigation of the case, where the online requests are obliged by two or more servers that shares the storage space. The price of saving the amount on a host reliant on the number of already saved items on that server.

Han et al. [30] present a Cloud service recommendation system (CSRS) that would guide the consumer to choose the best set of services which suits according to their requirements. Different ranks of services are created by the RS and presented to the consumer from which the consumer selects the best according to the requirements. For various factors of difference cloud providers RS endorses the recommendation of the service which is based on the QoS of the network and Virtual Machine (VM). Different parameters of the QoS include various timing of execution, average execution, response, average response etc of cloud services. Service ranks (S-Rank) is used for the consideration of the quality of virtualization hypervisors utilized by various cloud vendors, feedback of the consumers and for the cost of better arrangement of the services. The results of the experiments infers that the Cloud service recommender system (CSRS) would efficiently recommend a good mixture of Cloud services to customers. The proposed system contributes to the model of Green IT and for better management of the resources effectively. Weakness of the paper is dependent and independent services will create conflict in user requirements list.

Beran et al. [31] emphasis is being given to the discussion and implementation of the similar edition of a genetic algorithm and a blackboard for the solution of QoS-aware service selection problems. For the comparison of both these and probably others a cloud based framework has been introduced for the purpose of appraising and finishing them properly. The real completion has been carried out through the use of Google App Engine. As a result it permitted a quick prototyping and operation in the cloud. The author talks about two main contributions that are made; firstly the extension of two approaches through their parallelization techniques for the purpose of improving its runtime performance. The important thing to be noticed is that the

parallelization is carried out in a cloud environment. Secondly, a distributed database and deployment optimization (DD-Optimization) framework organized on the Google Application Engine helped to permit a faultless integration of other algorithms for the same kind of service selection problems. Strength of the paper is a Cloud-based framework for hosting, performing and assessing the algorithms has been presented. It was found out that as compared to the previous work the Google Application Engine (GAE) is an appropriate environment for testing and comparing the planned algorithms in both cases (sequential and parallel). Weakness of the paper is, more algorithms are to be considered for extensive benchmarking, QoS-aware service selection problems.

Nathuji [32] et al. suggests that the cloud should provide extra sources according to the requirements to get the enhanced performance that customers would have observed if they were operating in isolation. Accordingly, the author have designed Q-Clouds, a QoS-aware control framework that improves source percentage to reduce performance interruption effects. Q-Clouds uses online opinions to build a multi-input multi-output (MIMO) model that records performance interruption relationships, and uses it to perform closed loop source management. Moreover, the efficiency is used to allow applications to mention various levels of QoS as program Q-states. Q-Clouds with dynamism delivers under used sources to allow raised QoS stages, thereby enhancing program performance. Trial assessments of the solution using standard benchmark programs show the advantages: performance disturbance is alleviated completely when possible, and system usage is enhanced by up to 35% using Q-states. To overcome the challenges charged by efficiency disturbance effects, author recommends an alternative approach: QoS-aware cloud that actively compensate for efficiency disturbance using closed loop source control. Q-Clouds have been presented, a QoS-aware control theoretic control framework for multicore cloud servers. Q-Cloud servers manage disturbance among combined Virtual Machines by dynamically adjusting resource allocations to programs based upon amount of workload SLAs. Q-Clouds ensures that the efficiency experienced by programs is the same as they would have obtained if there was no performance disturbance. This paper focused on Q-Clouds system that is developed to offer promises that the efficiency knowledgeable by programs is separate of whether it is combined with other workloads. A MIMO model is presented that records disturbance effects to drive a closed loop resource management controller. Q-Cloud servers can create use of the declared states to provision lazy sources dynamically, thereby enhancing cloud efficiency and utilizations. A strategy to examine how Q-Clouds can be extended to better deal with disturbance in the I/O paths. How concerned issues around applications with phased behaviors should be treated. Checking out the incorporation of overall performance disturbance aware control for dynamic work location using live migration techniques.

Hill et al. [33] presented various results from the experiments conducted on Windows Azure platform. An exhaustive performance evaluation of the integral parts from various platform is shown. The paper provides recommendations for performance improvement for azure storage services, dynamic scaling, azure sql services and for testing and development.

Wang et al. [34] proposed a distributed file system known as ASDF for meeting the requirements of data-intensive applications, users, developers and administrators. Main features of the system are compatibility, extensibility and autonomy.

Zhao et al. [35] presented a new additional extra efficient query algorithm to deal with SQL query. The algorithm utilizes various techniques of divide and conquer, scheduling algorithms to get load balance and the pipeline technique to process result return.

Larumbe et al. [36] presents Cloud Location and Routing Problem (CLRP), a mathematical problem aiming to solve all issues existing inside a multi-layer using convex integer programming formulation. Cloud computing is geographically distributed on multiple regions, data centers location, servers and software elements. The results highlight importance of location and routing when different information is disseminated on the internet and the resources utilized effect the overall network performance.

Barboza et al. [37] proposes and evaluates the feasibility of constructing a plain off-the-shelf architecture for on demand gaming service. The proposal depends on running appropriate remote server in cloud that enables the client to perform a few basic tasks like reading user input and displaying the required game screens. Low-power computing systems like Mobile devices, digital TV setup boxes can also be used because it uses rich graphics and most of the processing is performed remotely. Only user input processing and execution of the game logic was focused. Game sounds and their transfer over network were not focused and addressed.

Breskovic et al. [38] introduced a method for cost-efficient use of SLA templates in autonomic Cloud computing. An approach is presented for autonomic amendments and creation of SLA mappings to decrease the cost of creating SLA mappings for market contributors, allowing them to exploit new public SLA templates representing the present market tendencies without any exertion. Author investigated autonomic creation and management of public SLA templates by analyzing their structures and SLO values. The results of the evaluation model based on the simulation framework shows that the approach increases the overall net utility of traders and market in general and lowers the cost of market maintenance.

Gao et al. [39] provides the performance assessment and scalability issues and needs in various clouds. Then, it provides a set of formal and graphic designs along with

metrics, that can be used to assess SaaS performance and scalability in clouds. Furthermore, it reports our example in the EC2 cloud environment. The results show the good potential application and efficiency of the suggested design models in assessing SaaS performance and scalability.

Kumar et al. [40] discuss the three scheduling techniques Min-Min, Min-Max and Genetic Algorithm have been mentioned as well as analytics of Min-Min and Min-Max have been shown. The performance of the standard Genetic Algorithm and the suggested Improved Genetic Algorithm have been examined against the example data. New scheduling idea is also suggested in which the Min-Min and Max-Min can be mixed in the Genetic Algorithm.

III. KEY CHALLENGES IDENTIFIED

In this section we summarize gap analysis of the different key challenges, methodologies, tools and techniques that have been observed during the course of literature survey of this study for performance improvement. The different aspects of cloud computing that have been accounted for in this analysis include: architectural management, efficiency, reliability, response time, quality of services employing software development methodologies, addressing collaboration issues between service provider and consumer while maintaining trusts in cloud computing. The analysis also accounts for the need to develop and promote cloud computing tools to exploit true benefits of computing by following defined set of guidelines and policies that exclusively target the cloud computing services.

Some of the key challenges identified in different areas for performance improvement include storage services, scaling, network services, scheduling, service level agreement templates, optimal location of data centers and software components, efficient SQL query processing, architecture and process improvement.

As in persistent storage the data is stored on multiple locations and replicated thereby increasing concurrency and therefore it's a key challenge to be addressed. Design and development of scalable cloud applications would enhance utilization of numerous cloud applications. The requirement for substantial space for storing on data-intensive applications and high-performance computing keeps growing. Downloading a file having size less than 64MB takes more delay rather than a file having size greater than 64MB. Improving query algorithm efficiency in cloud data management system, especially query on deliberate files turn out to be an increasingly important challenge. When the number of client nodes (slaves) increases the query processing algorithm should be flexible enough to scale & improve performance by reducing cost. Executing social gaming purposes, i.e. purposes that want real-time responsiveness in the whole cloud is challenging and still not exactly regular in the cloud environment. Basic problem in the area is to create UDP network protocol and to reduce the delay of communication by frequent requests.

IV. CONCLUSION

Cloud computing presents a new period of computing that provides additional ways for resource allocation and uses. This paper is an effort to study the current state of the affair with respect to quality of services in the cloud computing environment. The paper also observed into the key challenging areas that how resources are allocated to clients and what are the roles of cloud providers. Similarly this paper also investigated how the performance can be increased by improving various components in a scalable way with low cost, better performance and QoS. Some technical and functional issues in cloud that affects the performance of a cloud are also pointed out. The literature review reveals many underlying issues in cloud that degrade system performance with particular reference to scalability and cost issues for a specific resource. The paper also discusses many of the deployed systems with their pros and cons.

REFERENCES

[1] S. Distefano, A. Puliafito, M. Rak and S. Venticinque, "QoS Management in Cloud @ Home Infrastructures", *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, IEEE (2011)*.

[2] "CloudHome project," 2010, https://cloudathome.unime.it/.

[3] Y. Xiao, C. Lin, Y. Jiang, X. Chu and X. Shen, "Reputation-based QoS Provisioning in Cloud Computing via Dirichlet Multinomial Model", *ICC 2010 proceedings, IEEE (2010)*.

[4] J. Deng, S. C.-H. Huang, Y. S. Han and J. H. Deng, "Fault-Tolerant and Reliable Computation in Cloud Computing", *Globecom 2010 Workshop on Web and Pervasive Security, IEEE (2010).*

[5] K. Xiong and H. Perros, "Service Performance and Analysis in Cloud Computing", *Congress on Services – I, IEEE (2009)*.

[6] J. Y. Lee, J. W. Lee, DW. Cheun and S. D. Kim, "A Quality Model for Evaluating Software-as-a-Service in Cloud Computing", *Seventh ACIS International Conference on Software Engineering Research, Management and Applications, IEEE (2009)*.

[7] R. Jeyarani, R. Vasanth ram and N. Nagaveni, "Design and Implementation of an efficient Two-level Scheduler for Cloud Computing Environment", *10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, IEEE (2010)*.

[8] J. Wang, P. Varman and C. Xie, "Avoiding Performance Fluctuation in Cloud Storage", *IEEE (2010)*.

[9] A. J. Younge, G. Laszewski, L. Wang, S. Lopez-Alarcon and W. Carithers, "Efficient Resource Management for Cloud Computing Environments", *IEEE (2010).*

[10] M. Olson and K. M. Chandy, "Performance issues in cloud computing for cyber-physical applications", *4th International Conference on Cloud Computing, IEEE (2011).*

[11] Y. Liang, Z. Hao, N. Yu and B. Liu, "RandTest: Towards More Secure and Reliable Dataflow Processing in Cloud Computing", *International Conference on Cloud and Service Computing, IEEE (2011).*

[12] J. Du, W. Wei, X. Gu, and T. Yu, "Runtest: assuring integrity of dataflow processing in cloud computing infrastructures," *in Proceedings of the 5th ACM Symposium on Information, Computer and Communications Security, ser. ASIACCS '10. New York, NY, USA: ACM, 2010, pp. 293–304.*

[13] H. Qian, D. Medhi and K. Trivedi, "A Hierarchical Model to Evaluate Quality of Experience of Online Services hosted by Cloud Computing", 12th IFIP/IEEE International Symposium on Integrated Network Management, IEEE (2011).

[14] M. Luo, L. Zhang and F. Lei, "An Insuanrance Model for Guranteeing Service Assurance, Integrity and QoS in Cloud Computing", IEEE International Conference on Web Services, IEEE (2010).

[15] R. K. Jha and U. D. Dalal, "On Demand Cloud Computing Performance Analysis With Low Cost For QoS Application", International Conference on Multimedia, Signal Processing and Communication Technologies, IEEE (2011).

[16] A. Iosup and R. Prodan, "Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing", IEEE Transactions On Parallel And Distributed Systems, Vol. 22, No. 6, June (2011).

[17] B. Yang, X. Xu, F. Tan and D. H. Park, "An Utility-Based Job Scheduling Algorithm for Cloud Computing Considering Reliability Factor", 2011 International Conference on Cloud and Service Computing, IEEE (2011).

[18] K. V. Vishwanath and N. Nagappan, "Characterizing Cloud Computing Hardware Reliability", SoCC'10, USA (June 10–11, 2010).

[19] Y. C. Lee, A. Y. Zomaya and M. Yousif," Reliable Workflow Execution in Distributed Systems for Cost Efficiency", 11th IEEE/ACM International Conference on Grid Computing, IEEE (2010).

[20] Y. Yuan and W. Liu, "Efficient resource management for cloud computing", 11th IEEE/ACM International Conference on System Science, Engineering Design and Manufacturing Informatization, IEEE (2011).

[21] M. Litoiu and M. Litoiu, "Optimizing Resources in Cloud, a SOA Governance View", GTIP, USA (Dec. 7, 2010).

[22] Y. Wang and H. LV, "Efficient Metadata Management in Cloud Computing", Proceedings of IEEE, (2011).

[23] J. Z. W. Li, M. Woodside, J. Chinneck and M. Litoiu, "CloudOpt: Multi-Goal Optimization of Application Deployments across a Cloud", 7th International Conference on Network and Service Management (CNSM), IEEE, (2011).

[24] M. Alhamad, T. Dillon, C. Wu and E. Chang, "Response Time for Cloud Computing Providers", WAS2010, France, (8- 10 November, 2010).

[25] M. D. Assunção, A. Costanzo and R. Buyya, "Evaluating the Cost-Benefit of using Cloud Computing to Extend the Capacity of Clusters", HPDC'09, Germany, (June 11–13, 2009).

[26] V. Sekar and P. Maniatis, "Verifiable Resource Accounting for Cloud Computing Services", CCSW'11, USA, (October 21, 2011).

[27] X. Wang and Y. Wang, "Energy-efficient Multi-task Scheduling based on MapReduce for Cloud Computing", Seventh International Conference on Computational Intelligence and Security, IEEE, (2011).

[28] R. K. Jha and U.D. Dalal, "A performance comparison with cost for QoS application in on0demand cloud computing", International Conference on Recent Advances in Intelligent Computational Systems (RAICS), IEEE, (2011).

[29] D. Bein, W. Bein and S. Phoha, "Efficient data centers, cloud computing in the future of distributed computing", Seventh International Conference on Information Technology, IEEE, (2010).

[30] S. Han, M. M. Hassan, C. Yoon and E. Huh, "Efficient Service Recommendation System for Cloud Computing Market", ICIS 2009, Korea, (November 24 -26, 2009).

[31] P. P. Beran, E. Vinek and E. Schikuta, "A Cloud-Based Framework for QoS-Aware Service Selection Optimization", WAS2011, Vietnam, (5-7 December, 2011).

[32] R. Nathuji, A. Kansal and A. Ghaffarkhah, "Q-Clouds: Managing Performance Interference Effects for QoS-Aware Clouds", EuroSys'10, France, (April 13–16, 2010).

[33] Z. Hill, J. Li, M. Mao, A. Ruiz-Alvarez and M. Humphrey, "Early Observations on the Performance of Windows Azure", HPDC'10, Chicago, USA, (June 20-25, 2010).

[34] C-M. Wang, C-C. Huang, & H-M. Liang, "ASDF: An Autonomous and Scalable Distributed File System", 2011. 11th IEEE/ACM International

Symposium on Cluster, Cloud and Grid Computing, 485-493. IEEE. doi:10.1109/CCGrid.2011.21

[35] J. Zhao, X. Hu & X. Meng, "ESQP : An Efficient SQL Query Processing for Cloud Data Management", 2010, CloudDB'10, Canada, (October 30, 2010).

[36] F. Larumbe & B. Sans, "Optimal Location of Data Centers and Software Components in Cloud Computing Network Design", 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2012, doi:10.1109/CCGrid.2012.124

[37] D. C. Barboza, H. L. Junior, E. Walter, G. Clua & V. E. F. Rebello, "A Simple Architecture for Digital Games On Demand using low Performance Resources under a Cloud Computing Paradigm", Brazilian Symposium on Computer Games and Digital Entertainment, 2010, doi:10.1109/SBGAMES.2010.34

[38] I. Breskovic, M. Maurer, V. C. Emeakaroha, I. Brandic & S. Dustdar, "Cost-Efficient Utilization of Public SLA Templates in Autonomic Cloud Markets", Fourth IEEE International Conference on Utility and Cloud Computing, 2011, doi:10.1109/UCC.2011.38

[39] P. Pattabhiraman, X. Bai & T. Tsai, "SaaS Performance and Scalability Evaluation in Clouds", Proceedings of The 6th IEEE International Symposium on Service Oriented System Engineering (SOSE 2011), 2011, 61-71.

[40] P. Kumar, A. Verma, "Scheduling Using Improved Genetic Algorithm in Cloud", ICACCI'12, Chennai, India, (August 03-05, 2012).

Mr. Ashraf Zia is a lecturer in Computer Science at the Department of Computer Science, Abdul Wali Khan University, Mardan. In teaching, he has been focusing on applying web engineering concepts and problem based learning approaches in Computer Science Education. In research, his current interests include Cloud Computing, Global Software Development and Requirement Engineering. Mr. Zia received his Bachelor's degree in Computer Science from University of Peshawar, Pakistan.

Dr. Muhammad Naeem Ahmed Khan obtained D.Phil. degree in Computer System Engineering from the University of Sussex, Brighton, England, UK. Presently, he is affiliated with Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad. His research interests are in the fields of software engineering, cyber administration, digital forensic analysis and machine learning techniques.

