I.J.Modern Education and Computer Science, **2013, 6, 56-61**
Published Online July 2013 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2013.06.08

A Scheme to Reduce Response Time in Cloud Computing Environment

Ashraf Zia, M.N.A. Khan
Department of Computing, Shaheed Zulfikar Ali Bhutto Institute of Science & Technology, Islamabad, Pakistan
ashrafzia@gmail.com, mnak2010@gmail.com

Abstract— The area of cloud computing has become popular from the last decade due to its enormous benefits such as lower cost, faster development and access to highly available resources. Apart from these core benefits some challenges are also associated with it such as QoS, security, trust and better resource management. These challenges are caused by the infrastructure services provided by various cloud vendors on need basis. Empirical studies on cloud computing report that existing quality of services solutions are not enough as well as there are still many gaps which need to be filled. Also, there is a dire need to develop appropriate frameworks to improve response time of the clouds. In this paper, we have made an attempt to fill this gap by proposing a framework that focuses on improving the response time factor of the QoS in the cloud environment such as reliability and scalability. We believe that if the response time are communicating effectively and have awareness of the nearest and best possible resource available then the remaining issues pertaining to QoS can be reduced to a greater extent.

Index Terms— Response Time, QoS, Performance, Cloud Computing.

I. Introduction

In a broader spectrum, the cloud computing is the delivery of computing as Software as a Service, Platform as a Service and Infrastructure as a Service. Various virtual machines are deployed inside the cloud depending upon the customer requirements. Cloud computing has many important aspects such as scalability, dynamic, efficiency, cost, reliability and security. There are cloud computing companies which offer huge scalable processing infrastructures at a lower price. This can help organizations to evade the high initial cost of setting up an application deployment environment., However there are large scale software systems such as social media sites and e-commerce programs which can benefit significantly by using such cloud computing solutions to reduce their costs and enhance service quality.

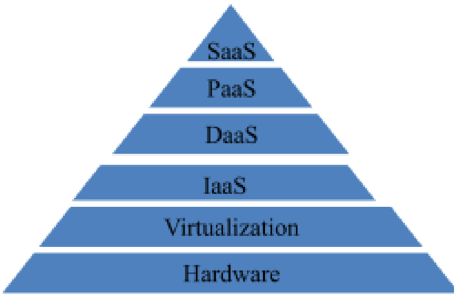


Figure1. Cloud Computing Layer Wise Distribution [7].

A. *Delivery Model of Cloud Computing.*

In the software layer, the application software is used by the customer, but it do not manage the operating systems or system utilities on which it is run. The platform layer makes use of a web host atmosphere for the deployed programs. The customer handles the programs and also manages them over the web based environment, but he/she cannot perform management of the operating system and related system utilities. In the infrastructure layer of the cloud, the clients can utilize hard drive, processors, memory, server elements or middleware. Here, the customers can manage OS, hard drive, deployed programs and possibly the underlying social network elements.

B. *Deployment Models of Cloud Computing*

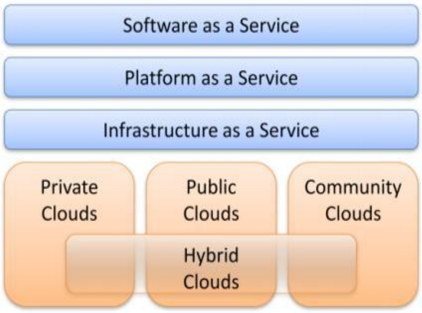


Figure2. Cloud Computing Service & Deployment Models (Source: Sriram & Khajeh-Hosseini, 2010).

In public cloud, only those services are identified that are required to be accessible to the customers from a third party service via the Internet. A public cloud does not mean that an individual's information is publically available; rather suppliers generally put in place an access control process for their clients. Group environment provides cost-effective means to set up solutions. In public clouds, users just pay for what they have utilized. However, a firm may opt to set up an intra IaaS or private computing to make use of the scalability advantages of a distributed data centers. A private cloud provides several advantages of a public cloud processing environment, such as being flexible and service-based. The major distinction between a private cloud and a public cloud is that in a the former model, the data and procedures are handled within the company without the limitations of system information, protection exposures and hassle of adherence to rules/regulations that govern the public clouds. In addition, private cloud solutions offer greater control over the cloud infrastructure as well as enhancing protection and resilience as the networks used are quite limited and have to use the solutions for specific purposes.

The group cloud community is managed and used by a group of firms that have specific security needs or common objective. The group members have shared access to the information and application programs in the cloud.

Hybrid cloud is a collaboration of public and private clouds that are capable to interoperate. In this model, users typically do not delegate business-critical information and computation to the public cloud; rather they keep their business-critical products and services within their control own private cloud.

II. SIGNIFICIANT OF THE PROBLEM

Cloud computing is the synthesis of a variety of technologies and innovations of various business models. Cloud computing is growing rapidly and companies such as Microsoft, Amazon, RedHat and IBM are progressively financing cloud computing infrastructure and research. Cloud computing has become a hot topic in the Internet information services. The architecture of existing cloud computing is primarily based on powerful datacenter facilities and relatively weaker clients' backgrounds to exploit its strengths truly. It is expected to bring about revolution in the entire information industry operation mode, which would have a profound impact on socio-economic structure.

III. RELATED WORK

Vishwanat *et al.* [1] propose to consider server failure rates to comprehend the hardware reliability for huge cloud computing infrastructures. The time spent on server break-down recovery and hardware components fixes could be used for this purpose as well as identifying features and predicates that lead to failure.

Lee *et al.* [2] suggested algorithm features a cost-aware duplication program to increase reliability. The duplication program successfully determines replicability by taking into account the duplication cost, charge (incurred by deviating from SLA targets) and failing features.

Yuan *et al.* [3] utilizes several strategies including resource pre-reservation and resource borrowing for managing cloud resources. 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.

Litoiu *et al.* [4] recommends that how optimization of resource distribution can help achieve considerable cost deductions. The author takes into account dynamic workloads and suggests a new optimization technique along with a Service Oriented Architecture (SOA) governance approach to cloud optimization.

Wang *et al.* [5] suggest a competent distributed metadata management scheme. Through different techniques it can express powerful 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.

Li *et al.* [6] present 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. The important thing is that it is scalable and extendable to new objects.

Alhamad *et al.* [7] develops a performance metrics for the measurement and comparison of the scalability of the resources of virtualization on the cloud data centers. Author carried out a number of experiments on Amazon EC2 cloud at different times. Each time the response was judged. The attention was examining the solitude across the same hardware components of virtual machines that are organized by a cloud vendor.

Assunção *et al.* [8] investigate the benefits that different companies can collect by using cloud computing services, capacity of their local infrastructure for the improvement of their performance upon the requirements of its users.

Sekar *et al.* [9] propose that cloud computing offers its users the possibility to reduce operating and capital expenses.

Wang *et al.* [10] emphasized on the improvement of the energy efficiency of the servers through suitable scheduling strategies. A new scheduling model having energy-efficient and multi-tasking based on MapReduce has been introduced.

Jha *et al.* [11] present 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 assistance, data as assistance and speech assistance has also been suggested.

Bein *et al.* [12] present a problem about the allocation of the memory servers in a data center based on online request for storage. Two efficient algorithms are used for the selection of minimum array of servers and of the minimum overall cost.

Beran *et al.* [14] discuss 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. The real completion has been carried out through the use of Google App Engine.

Nathuji *et al.* [15] designed Q-Clouds, a QoS-aware control structure that increases source amount to decrease efficiency disruption results. Q-Clouds uses online views to develop a multi-input multi-output (MIMO) model that records performance interruption relationships, and uses it to perform closed loop source management.

IV. FACTORS FOR RESPONSE TIME IMPROVEMENT

A. *Optimum Bandwidth*

Bandwidth plays an important role in the response of the internet application service that we wish to have. If we wish to run a high application service like for example an online game but don't have enough bandwidth to get proper response then eventually our performance will be effected. Therefore, optimum bandwidth plays a vital role in the improvement of response time.

B. *Best Protocol Selection*

Various internet application require specific protocols to run them. If we had such a system that should suggest about the best protocols suited for the required application then it will definitely improve the response time and will remove the extra overhead.

C. *Best Medium Selection*

In Medium Test if we choose the wired media such as fiber optic which is very reliable and have higher data transfer rates then it will also improve the response time due to high transmission of data packets. Despite of the fact that the fiber optic medium is very expensive.

The wired media results in higher transfer rate and reliability is high rather that we choose wireless media.

In wireless media the various electric radiations, grass and weather not only effects the signal's strength but also open to security risks.

The wireless medium can easily hack through backtracking and other software's available in the market.

V. THE OVERALL SYSTEM ARCHITECTURE

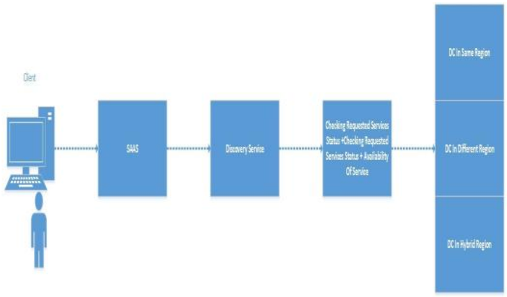

Client
SAAS
Discovery Service
Checking Requested Services Status +Checking Requested Services Status + Availability Of Service
DC In Same Region
DC In Different Region
DC In Hybrid Region

Figure3. Services selection w.r.t to response time in same, different & hybrid regions.

The end user selects respective services based on their requirements in the software and checks them. The services are forwarded to the discovery module for best response time and available services in the data center on the cloud provider. The data centers are targeted according to the availability and best response time with respect to data center in the same region, different region and data centers in the hybrid region.

VI. THE PROPOSED FRAMEWORK

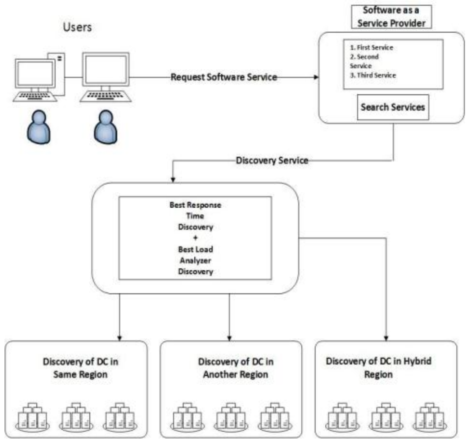

Users
Request Software Service
Software as a Service Provider
1. First Service
2. Second Service
3. Third Service
Search Services
Discovery Service
Best Response Time Discovery + Best Load Analyzer Discovery
Discovery of DC in Same Region
Discovery of DC in Another Region
Discovery of DC in Hybrid Region

Figure4. A high level system framework for best services discovery using multiple data centers in the same, different and hybrid regions in the Cloud.

The cloud provider is a central position where all the entire data centers are registered with their services and privileges. The CP works like a central server where all the data centers communicate and register their services.

The CP also has a central place where all decisions are made and policies are implemented. When a new request came for services it needs a specific service, it makes a connection to the CP. The CP provides a list of data centers with services including other services.

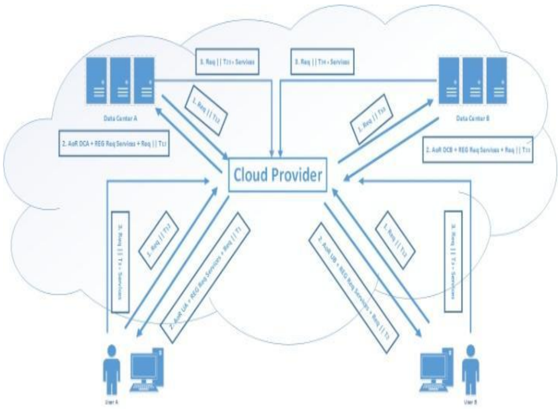


Figure5. The System Architecture.

The algorithm clearly mentions the step by step flow of the overall system. In below data flow diagram, the data center request for services with the other data centers having the same service. The cloud provider implements the overall policy of the scheme.

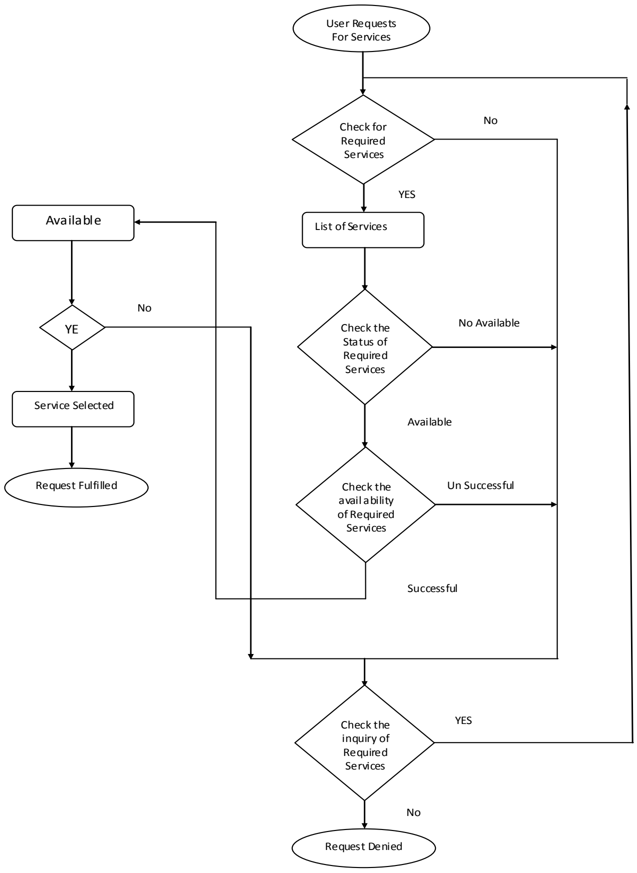
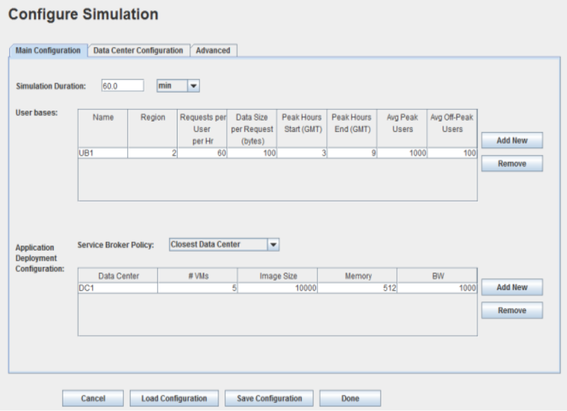

User Requests For Services
Check for Required Services
No
YES
List of Services
Check the Status of Required Services
No Available
Available
Check the availability of Required Services
Un Successful
Successful
Available
YE
No
Service Selected
Request Fulfilled
Check the inquiry of Required Services
YES
No
Request Denied

Figure6. Data flow diagram of the Service Discovery on the Data Centers in the Cloud Provider for Services

VII. RESPONSE TIME SIMULATIONS

For checking the response time we have drawn different scenarios inside the cloud analyst simulator. The data centers are divided in different regions with different users based in a particular region. These scenarios are explained below:



Configure Simulation

Main Configuration | Data Center Configuration | Advanced

Simulation Duration: 60.0 min

User bases:

Name	Region	Requests per User per Hr	Data Size per Request (bytes)	Peak Hours Start (GMT)	Peak Hours End (GMT)	Avg Peak Users	Avg Off-Peak Users
UB1	2	60	100	3	9	1000	100

Application Deployment Configuration: Service Broker Policy: Closest Data Center

Data Center	# VMs	Image Size	Memory	BW
DC1	5	10000	512	1000

Figure7. Configuration of the Simulator.

In the above figure the configuration of the simulator is shown. In the main configuration you can add different users based in different regions and you can also add different data centers to a specific region. Similarly you can also create copies of different virtual machines in a specific data center. In the service broker policy you select the desired policy such as closest data center, optimal response time. In the advanced tab you can select the no of simultaneous user accessing the data center and no of data centers simultaneously providing services to no of users. Six different regions can be selected around the whole world with different data centers and users in different locations depending upon the situation.

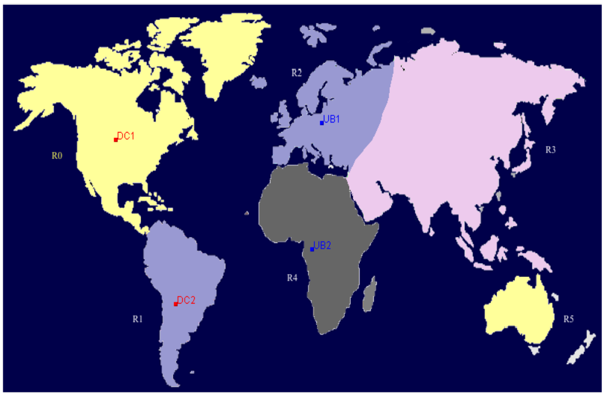


Figure8. Users & Data Centers based in different regions.

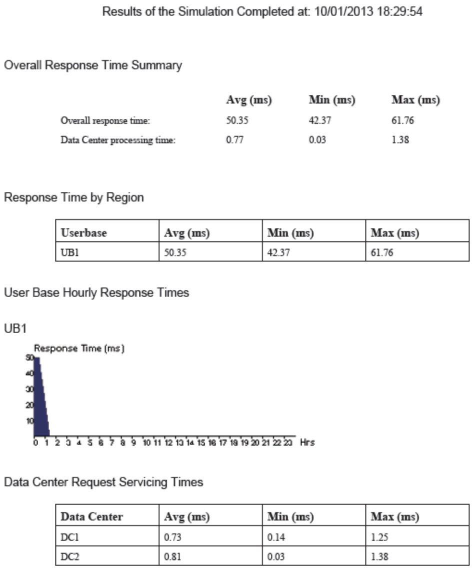
For simulation the following parameters are set:

TABLE1. Parameters List for Simulation

Parameter		Values
Virtual Machine	Image Size	10000
	Memory	1Gb
	Bandwidth	100
Data Center	Architecture	X86
	OS	Linux
	VMM	Xen
	Number of Machines	25
	Memory per Machine	2Gb
	Storage per Machine	100000Mb
	Bandwidth per Machine	10000
	Number of processors per machine	5
	Processor Speed	100MIPS
	VM Policy	Time Shared
Grouping Factor	User Grouping Factor	1000
	Request Grouping Factor	100
	Executable Instruction Length	250

A. *Availability of Services & Location of Data Centre in the Same Region*

In this particular scenario the user and data centers are located in the same region with the following results.



Results of the Simulation Completed at: 10/01/2013 18:29:54

Overall Response Time Summary

	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	50.35	42.37	61.76
Data Center processing time:	0.77	0.03	1.38

Response Time by Region

Userbase	Avg (ms)	Min (ms)	Max (ms)
UB1	50.35	42.37	61.76

User Base Hourly Response Times

UB1

Data Center Request Servicing Times

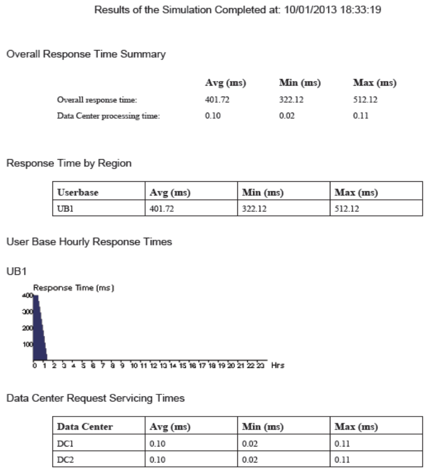
Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	0.73	0.14	1.25
DC2	0.81	0.03	1.38

Figure9. Response Time of User & Data Center in the Same Region

As you can see in the above results that users and data centers located in the same region has low response time.

B. *Availability of Services & Location of Data Centre in the Different Regions*

In this particular scenario the user and data centers are located in the hybrid regions with the following results.



Results of the Simulation Completed at: 10/01/2013 18:33:19

Overall Response Time Summary

	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	401.72	322.12	512.12
Data Center processing time:	0.10	0.02	0.11

Response Time by Region

Userbase	Avg (ms)	Min (ms)	Max (ms)
UB1	401.72	322.12	512.12

User Base Hourly Response Times

UB1

Data Center Request Servicing Times

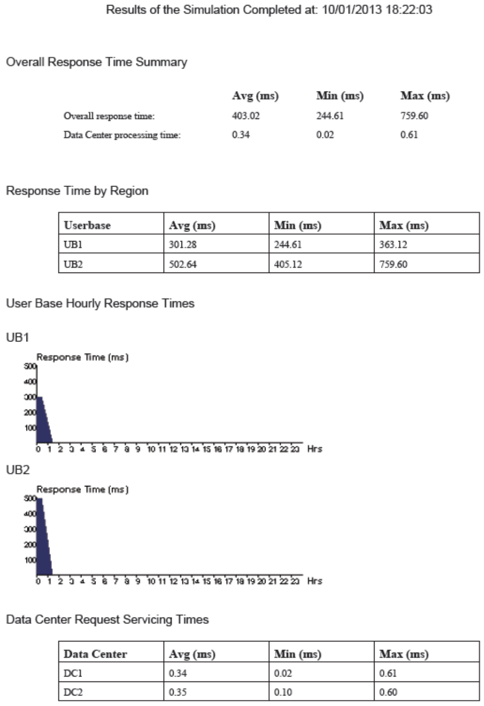
Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	0.10	0.02	0.11
DC2	0.10	0.02	0.11

Figure10. Response Time of User & Data Centers in different Region

In the above results that users and data centers are located in different regions with greater response time.

C. *Availability of Services & Location of Data Centre in Hybrid Regions*

In this particular scenario the user and data centers both are located in the hybrid regions with the following results.



Results of the Simulation Completed at: 10/01/2013 18:22:03

Overall Response Time Summary

	Avg (ms)	Min (ms)	Max (ms)
Overall response time:	403.02	244.61	759.60
Data Center processing time:	0.34	0.02	0.61

Response Time by Region

Userbase	Avg (ms)	Min (ms)	Max (ms)
UB1	301.28	244.61	363.12
UB2	502.64	405.12	759.60

User Base Hourly Response Times

UB1

UB2

Data Center Request Servicing Times

Data Center	Avg (ms)	Min (ms)	Max (ms)
DC1	0.34	0.02	0.61
DC2	0.35	0.10	0.60

Figure11. Response Time of Service & Data Center in the Hybrid Regions.

In the above results that users and data centers are located in hybrid regions with an average response time.

From the above three scenarios it is quite clear that the users and data centers in the same region has less response time, while users and data centers in the different regions has high response time and users & data centers in the hybrid data centers have an average response time.

VIII. CONCLUSION

Cloud computing changed the era of traditional computing by providing on the fly computation services. Cloud environments are colonized with thousands of smart data center. The cloud resources and services have been changing with user modes and circumstances. All these heterogeneous data centers communicate each other in ad- hoc manner and construct an ad hoc network. Any mobile or static service comes and become as part of this ad hoc network. These services provide and use resources while at the same time many other process and service roaming in this ad hoc network. For availability and reliable use of services (better QoS), availability of data centers, services, users and process are significant. A lightweight, portable and availability scheme always needed for cloud environment which enhanced the availability level.

Availability of all included services is important for better quality of services (QoS). Here we analyzed some of well- known and currently deployed schemes for availability. We point out strengths and weaknesses of each scheme and also suggest further solution for improvements.

REFERENCES

[1] K. V. Vishwanat and N. Nagappan, “Characterizing Cloud Computing Hardware Reliability”, *SoCC’10, USA (June 10–11, 2010).* DOI:10.1145/1807128.1807161.

[2] 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).* DOI:10.1109/GRID.2010.5697959.

[3] 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).* DOI:10.1109/ICSSEM.2011.6081285.

[4] M. Litoiu and M. Litoiu, “Optimizing Resources in Cloud, a SOA Governance View”, *GTIP, USA (Dec. 7, 2010).* DOI:10.1145/1920320.1920330

[5] Y. Wang and H. T. LV, “Efficient Metadata Management in Cloud Computing”, *Proceedings of IEEE, (2011).* DOI:10.1109/ICCSN.2011.6014777.

[6] J. Z. 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).*

[7] M. Alhamad, T. Dillon, C. Wu and E. Chang, “Response Time for Cloud Computing Providers”, *WAS2010, France, (8- 10 November, 2010).* DOI:10.1145/1967486.1967579.

[8] Marcos Dias de Assunção, Alexandre di Costanzo and RajkumarBuyya, “Evaluating the Cost-Benefit of using Cloud Computing to Extend the Capacity of Clusters”, *HPDC’09, Germany, (June 11–13, 2009).* DOI:10.1145/1551609.1551635.

[9] V. Sekar and P. Maniatis, “Verifiable Resource Accounting for Cloud Computing Services”, *CCSW’11, USA, (October 21, 2011).* DOI:10.1145/2046660.2046666.

[10] 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). DOI:10.1109/CIS.2011.21.*

[11] 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). DOI:10.1109/RAICS.2011.6069264.*

[12] 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). DOI:10.1109/ITNG.2010.31.*

[13] S. Han, M. M. Hassan, C.W. Yoon and E.N. Huh, “Efficient Service Recommendation System for Cloud Computing Market”, *ICIS 2009, Korea, (November 24 -26, 2009). DOI:10.1145/1655925.1656078.*

[14] P. P. Beran, E. Vinek and E. Schikuta, “A Cloud-Based Framework for QoS-Aware Service Selection Optimization”, *WAS2011, Vietnam, (5-7 December, 2011). DOI:10.1145/2095536.2095584.*

[15] R. Nathuji, A. Kansal and A. Ghaffarkhah, “Q-Clouds: Managing Performance Interference Effects for QoS-Aware Clouds”, *EuroSys’10, France, (April 13–16, 2010). DOI:10.1145/1755913.1755938.*

Ashraf Zia is a student of Computer Science at the Department of Computing, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad. His research interests are in the field of Software Engineering, Global Software Development, Requirement Engineering and Cloud Computing.

M.N.A. Khan obtained D.Phil. degree in Computer System Engineering from the University of Sussex, Brighton, England, UK. His research interests are in the fields of Software Engineering, Cyber Administration, Digital Forensic Analysis and Machine Learning Techniques.