# Performance Degradation Factors in Cloud Computing

Rizwan Khalid, Tahir Abdullah, Imran Rashid

**Abstract**—Cloud computing presents the idea of elastic and highly scalable networking resources provisioned as a service over the internet. It is an emerging Internet based computing where shared resources, data and information are provided on demand to the end users using computers and other devices. Cloud is the significant model to access the distributed computing resources. This research is based on evaluation of performance degradation factors in cloud computing. Performance evaluation is an important factor in any application particularly in cloud networks. Traditional benchmarks are not sufficient to analyze and evaluate performance of the cloud. This research covers different strategies of performance evaluation in cloud computing. There are many factors, which covers performance degradation in cloud networks and unlike traditional systems it is difficult to reveal the degradation factors and their intensity. This research will discover different types of performance degradation factors and their magnitude in cloud computing. High performance is the essential key elements in cloud computing to make the cloud users happy. The objective of this study is to highlight the significance of parameters which can create the high performance of cloud computing. The ambition of this study is to present progressive advancement in understanding the implications of cloud computing performance parameters. The approach of this study has been to undertake an evaluation of the technological factors affecting the cloud computing performance and to present idea to build highly scalable cloud environments. This research study will also identify the metrics through which performance of the cloud network becomes unreliable and unsuitable. This research will be helpful for those who wants to deploy new cloud infrastructure.

**Index Terms**— Cloud Computing, Performance evaluation, degradation factors, parameters, Cloud Infrastructure, Traditional benchmarks, highly scalable

## 1 INTRODUCTION

The Cloud Computing is the provision of computing resources over the Internet. Cloud services authorize individuals, people, businesses and end users to utilize software, hardware and other resources that are managed by cloud providers located at remote areas. The end users are provided the infrastructures and applications by the data centers as pay per use services. Another, appropriate description of cloud computing which is considered to be the most accurate suggested by "National Institute of Standards and Technology" (NIST). The NIST presented the meaning of Cloud Computing as **"**Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics**,** three service models, and four deployment models." [1].

The people are benefitted with this remarkable technology by not investing much on the resources provided by cloud computing systems. This technology increase the capacity to use services with no investment for computer network, personnel training or getting software licenses. Users can access the cloud services which comprises of applications as a service without installing and maintaining these application and also obtain the facilities of storing data in storage databases located in remote areas.

Cloud computing is highly adoptable technology and rising rapidly in the current age because of its characteristics including security, scalability, reliability, high performance and low cost. There should service level agreement between customers and cloud service providers to avoid performance degradation obtaining the services of cloud. Violation of SLA rules will lead to performance degradation [2].

By using cloud computing technology, information and other computer resources are accessed wherever, internet connection is available. Cloud computing grants a publically shared pool of resources which are comprised of online storage space, processing power and dedicated corporate & user applications. In cloud network, the resources are ascertained and a user pays in agreement with to use these resources. Clouds can also provide deliberately varying user requests without stirring the performance and is available always for utilization without any limitation. Users can get the services of cloud from any internet enabled device, thus access to a wide group of people all over the world. This technology is considered to be very encouraging and numerous companies such as IBM, HP, Microsoft, Amazone, and Google have spent their time and invested on other resources for extra development of technologies related to cloud computing [3].

The data centers may be located at different geographical locations. But, the data centers can become burdened or overloaded while servicing the requests of excessive number of client applications located on the same location at the same time, this will degrade the performance of distributed services. Different configurations are required by the different user applications, measuring the performance of user applications at various resources is still challenging [4].

Cloud networks infrastructure have reliable security mechanisms and deliver the resources with amazing performance. The provision of hardware and other resources are delivered also as a service in cloud environment. Clients can only use their devices having facility of internet which can be smart phones, tablets or computers etc. to access the services of cloud just opening the web browsers conveniently. Consequently, cloud providers ensures the delivery of resources with entire obligations for the availability, reliability, quality of service and security of the resources delivered.

### 1.1 Cloud computing service models

Cloud networks and computing services can be elaborated in three different categories, which are arranged one by one as "Infrastructure as a Service (IaaS)", "Platform as a Service (PaaS)", and "Software as a Service (SaaS)".

**IaaS**

The IaaS is elaborated as Infrastructure as a Service and it presents the online services physical computing resources, data portioning, legacy, scaling, networking, backup, location etc. [5].

**PaaS**

The term PaaS explained as "Platform as a Service" and it is used for renting the cloud resources like hardware, storage, operating system and network capability on the internet. This service also delivers the platform for customers to use existing applications and develop new applications and perform test of these applications [6].

**SaaS**

SaaS is clarified as "Software as a Service" and it allows the software availability on the internet which is used by the customers as pay-on-demand or free of cost service. Most renowned example of SaaS service is Google Apps which provides different applications like email and word processing [7]**.**

### 1.2 TYPES OF CPLOUDS

There are four other types of models of cloud computing termed as Service Deployment Models which are described as under in this study.

**Private Cloud**

Private Cloud is established and controlled inside a single organization. These organizations use different software that facilitate with many cloud functionalities like virtual cloud, virtual cloud director by VMware etc. This system of cloud comprises of the applications privately hosted, and private networks take part in rivalling a cloud over the internet within the same company [8].

**Public Cloud**

Public Cloud is a place of computing resources offered by third-party organizations. Enterprises might utilize cloud usefulness from other resources, separately present their own services to the clients outside of the organization. Here is the all responsibility of maintaining the cloud goes to the cloud providers. Famous public clouds are, Examples: Google App Engine, windows Azure and Amazon [9].

**Hybrid Cloud**

Hybrid Cloud is a combination of computing resources supplied by both private and public clouds. Hybrid clouds correspond to a combined services of public and private cloud infrastructures to accomplish a thrilling decrease in cost via subcontracting as well as maintaining the preferred level of control over sensitive data by utilizing local private clouds [10].

**Community Cloud**

Community cloud in cloud computing is a mutual attempt or joint effort to share infrastructure among numerous organizations, build a particular community with mutual interests (compliance and security etc.). Whether, it is controlled and managed by third-party or internally and also it is hosted externally or internally [11].

### 1.3 Cloud Performance

Performance is generally linked to the capabilities of one of the applications within the cloud infrastructure itself. Besides all aspects of cloud computing infrastructure and resources, the most important and key point in cloud computing is its performance. Performance can be defined, as it is the execution accomplishment of a given task measured against predetermined known standards of accurateness, comprehensiveness, cost and speed. For high performance of distributed computing, all resources should be well-matched and influential [12].

High performance in cloud environments considered as it is one of the key factor that should be appropriate for each available cloud service. To get the maximum performance in cloud facilities and anything associated with cloud have great effect on cloud customers and the cloud service providers. And then evaluate the performance of the network, service providers and users are important. Testing is the activity of the league and requires new environments are created for each project. It essentially be verified for Web applications on several operating systems (OS) and their updates, platforms and various browser versions, and different kind of devices and an enormous number of simultaneous consumers to recognize their performance in actual time [13].

## 2 PROBLEM STATEMENT

The performance of cloud computing services for the users is main responsibility of cloud service provider. There are some key factors which causes degradation of performance in cloud computing. Accessibility of service that indicates available threats for distributed online services computing as a result of these aspects, like overload of frequent services, programming errors, or distributed DDoS (Distributed Denial of Service) attacks. Data transfer barrier is a huge obstacle which is transfer of data barrier and through which some points raised and these points are increasing of data, concentration of applications and these bottlenecks are effecting the rate of transfer of data as well as the cost also in cloud. Performance unpredictability specifies performance risks causing factors, like inefficient input/output division and by immense performance of distributed network systems in cloud. Storage scalability indicates the hurdles of deploying cloud computing to get solutions required extremely extensible constant storage. Scaling speedily represents the obstacles of rapidly climbing up and down in reply in terms of load without disregarding SLAs (*service-level agreements)*.

To overcome the above mentioned obstacles, there is a need of efficient load balancing algorithm which should eliminate the barriers of performance issues. In this study, our research has main focus on the evaluation of best load balancing algorithm which ensures the maximum performance of data centers in cloud computing.

## 3 MATERIAL AND METHODS

This research study is based on empirical methods using quantitative approach to conduct experiments to evaluate the cloud system performance using cloud performance evaluation tool and to find out results in the form of tables and graphs. In this research study, the following method for performance evaluation is used to investigate the factors, which are causing the performance degradation in cloud network.

### 3.1 Evaluation of performance in Cloud computing

High performance depends on the compatible and powerful resources of cloud computing. High performance must be satisfactory and adequate for each service because it is one of the benefit in cloud computing. One thing which has great influence on cloud service providers and users that is higher performance of services anything associated to the cloud. Therefore, it is very important to evaluate the performance for users and cloud providers.

### 3.2 Performance parameters

These days, the expression "performance" is more than an exemplary idea and incorporates more broad ideas, for example energy efficiency, reliability, scalability and so on. Because of the extending characteristics of cloud environments and the massive figure of enterprises and typical clients whose are utilizing cloud distributed services. There are numerous factors which have enormous impact on the performance of distributed computing and also on the resources it provided. Here are the most essential, significant and specific aspects affecting the cloud computing performance which are Region/Location, Data Centers, Data Center Controller, Number of users, Response Time, Processing time, Throughput, Elapsed time, Latency, Bandwidth, Servers, Memory (RAM), CPU, Storage, Security, Availability, Reliability and Scalability.

### 3.4 Performance evaluation tool

The tool that is used to evaluate the cloud performance is “Cloud Analyst” which is based on java technology. The aim of using cloud analyst as a framework tool is inspect in to prevailing simulation techniques for reviewing large scale cloud environments. Its primary purpose is to evaluate the performance of huge web applications for big organizations need to implement appropriate simulation tools and empirical methods techniques, and to fulfill all these requirements, the tool should be used cloud analyst which has the capability to evaluate large applications in cloud environment.

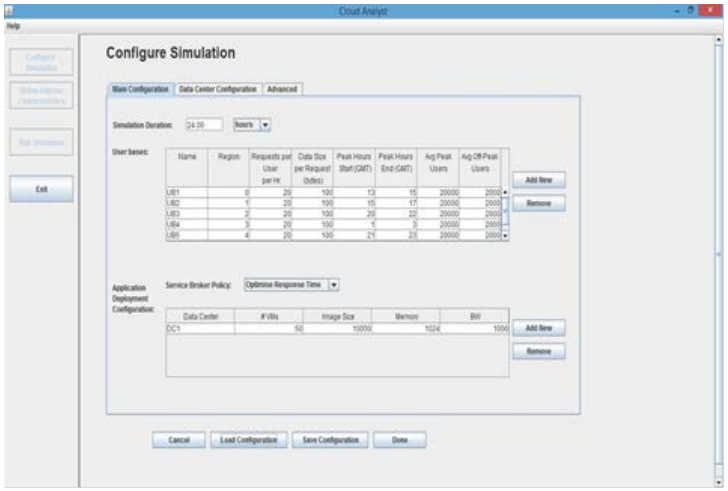


**Figure 2.** The main configuration view of Cloud Analyst

### 3.5 Load Balancing Algorithms

There are three essential, proficient and improved load balancing algorithms which are used to evaluate the performance of the cloud networks using cloud analyst are described as follows:

#### Round Robin Algorithm (RR)

This algorithm is constructed on random sampling. Means that, it randomly chooses the load in a case that some of the servers are highly loaded and some are lightly [14]. This algorithm executes by choosing virtual machines randomly. The primary job request of the client is allocated to a randomly designated virtual machine from the cluster of virtual machines and the data center controller assigns the job requests in a round manner. When the VM is assigned, it is relocated to the lower level of the group of virtual machines.

#### Throttled

This algorithm manages an index table in data center of all virtual machines and their relevant state for example (available or busy). When new request reaches, then the table is analyzed by load balancer and then that virtual machine is chosen having the state available not busy. After that, the virtual machine id having the available state is returned back to the data center controller that further allocates the request to the specific virtual machine [15].

#### Equally Spread Current Execution Load Algorithm (ESCE)

This algorithm is also called Active Monitoring algorithm for load balancing. It performs the tasks in similarities with throttled algorithm but there are some changes in respect to virtual machine index table. By using this algorithm the load balancer manages and maintains index table of all the deployed virtual machines and also maintains the presently held number of requests allocated to the virtual machines [16].

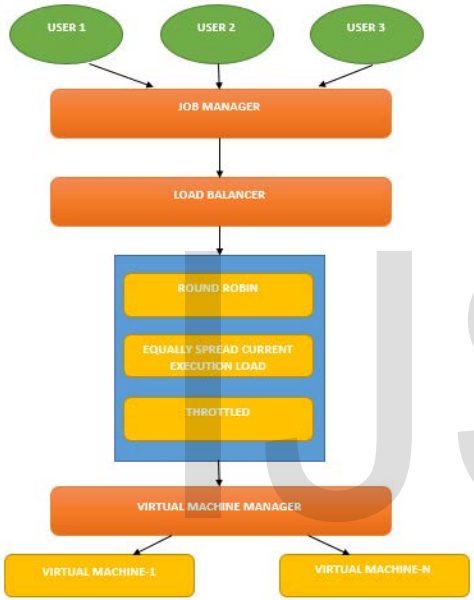


**Figure 3.** Execution of Load Balancing Algorithms

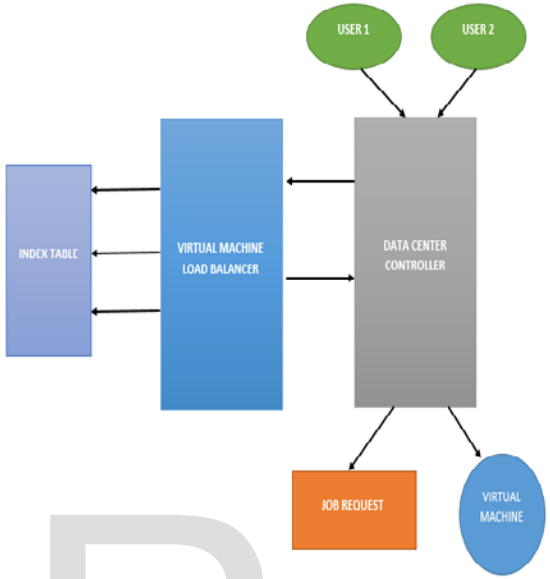


**Figure 4.** Working of Round Robin Algorithm

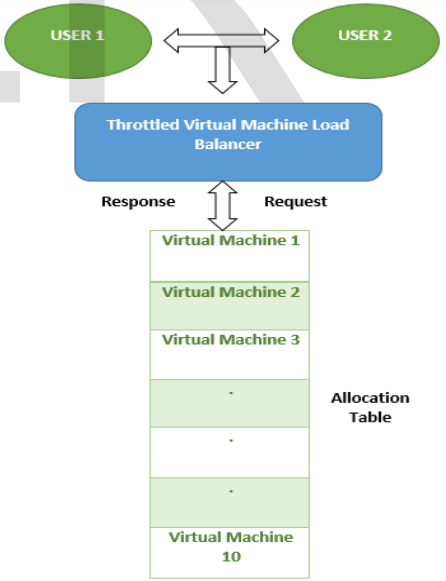


**Figure 5.** Throttled Load Balancing Algorithm

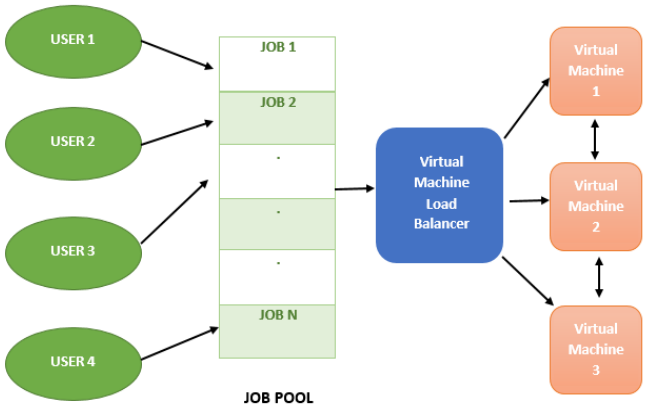


**Figure 6**. Equally Spread Current Execution Load Algorithm

### 3.6 Parameters used for experiments (variables and constants)

There are different parameters used to perform experiments on Cloud Analyst toolkit and the values of some of the parameters are kept constant and variables. Parameters used as variables in each user base are as follows.

*Variable Parameters*

Average Peak Users, Average Off-Peak Users, Load Balancing Algorithms

*Constant Parameters*

*Table of constant parameters is mentioned below.*

**Table 1.** The profile of parameters and their values which are kept constant.

| Parameters | Values | Parameters | Values | Parameters | Values | Parameters | Values |
|---|---|---|---|---|---|---|---|
| VM Image Size | 10000 | Data Center – Storage per machine | 100000 Mb | Data Center – VMM | Xen | User Grouping Factor | 1000 |
| VM Memory | 1024 Mb | Data Center – Available BW per Machine | 10000 | Data Center – Number of Machines | 20 | Request Grouping Factor | 100 |
| VM Bandwidth | 1000 | Data Center – Number of processors per machine | 4 | Data Center – Memory Per Machine | 2048 Mb | Executable Instruction Length in (bytes) | 250 |
| Data Center – Architecture | X86 | Data Center – Processor speed | 100 MIPS | Data Center – OS | Linux | Data Center – VM Policy | Time Shared |

**Table 2.** Delay Matrix (in Mbps) and Bandwidth Matrix (in Mbps) by region

| Region/ Region | Delay Matrix | | | | | | Bandwidth Matrix | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 |
| 0 | 25 | 100 | 150 | 250 | 250 | 100 | 2000 | 1000 | 1000 | 1000 | 1000 | 1000 |
| 1 | 100 | 25 | 250 | 500 | 350 | 200 | 1000 | 800 | 1000 | 1000 | 1000 | 1000 |
| 2 | 150 | 250 | 25 | 150 | 150 | 200 | 1000 | 1000 | 2500 | 1000 | 1000 | 1000 |
| 3 | 250 | 500 | 150 | 25 | 500 | 500 | 1000 | 1000 | 1000 | 1500 | 1000 | 1000 |
| 4 | 250 | 350 | 150 | 500 | 25 | 500 | 1000 | 1000 | 1000 | 1000 | 500 | 1000 |
| 5 | 100 | 200 | 200 | 500 | 500 | 25 | 1000 | 1000 | 1000 | 1000 | 1000 | 2000 |

### 3.7 Regions

The world is divided into six Regions geographically valued from (0 to 5) which correlated with the six main continents.

**Table 3.** User bases and related regions/continents

| User Bases | Region | Continents | Distance Between Regions | Kilometers |
|---|---|---|---|---|
| UB 1 | 0 | N. America | --- | --- |
| UB 2 | 1 | S. America | N. America and S. America | 8,418 km |

| | | | | |
|---|---|---|---|---|
| UB 3 | 2 | Europe | N. America and Europe | 6,725 km |
| UB 4 | 3 | Asia | N. America and Asia | 9,853 km |
| UB 5 | 4 | Africa | N. America and Africa | 13,802 km |
| UB 6 | 5 | Oceania | N. America and Oceania | 13,623 km |

### 3.8 EXPERIMENT PARAMETERS DESCRIPTION

Similar with most of the real-world web application let us undertake primarily the application is installed in a single site, in Region 0 (North America). Simulator is configured with a single data center located at region 0 in N. America, having 50 virtual machines, each of which has 1024 Mb of memory and other parameters which are configured and mentioned in the form of tables and screen shots. There are three load balancing algorithms used to perform experiments respectively which are Throttled, Round Robin and equally spread current execution load. Simulation time is set to 24 hours for each scenario in this experiment. For ease, a single time zone is defined with each user base and it is supposed that majority of users of the application use the applications after working hours in the evening for 2 hours. Also it is assumed that 5% users remain online concurrently during peak hours and the number of users in off peak hours is assumed to be one tenth of the peak hour users. In this experiment, the request time per user per hour is set to be 20, which means that each user makes request after every 3 minutes.

**Table 4.** Configuration details of different parameters

| User base | Regions | Requests Per User Per Hour | Data Size Per Request (bytes) | Time Zone | Peak Hours Start (GMT) | Peak Hours End (GMT) |
|---|---|---|---|---|---|---|
| UB1 | 0 – N. America | 20 | 100 | GMT–6.00 | 13:00 | 15:00 |
| UB2 | 1 – S. America | 20 | 100 | GMT –4.00 | 15:00 | 17:00 |
| UB3 | 2 - Europe | 20 | 100 | GMT + 1.00 | 20:00 | 22:00 |
| UB4 | 3 - Asia | 20 | 100 | GMT + 6.00 | 01:00 | 03:00 |
| UB5 | 4 - Africa | 20 | 100 | GMT + 2.00 | 21:00 | 23:00 |
| UB6 | 5 - Oceania | 20 | 100 | GMT + 10.00 | 09:00 | 11:00 |

It is supposed that the web applications are deployed on single location only i.e. Region 0 (North America). Similarly, 6 user bases are defined in different regions starting from region 0 to 5. Peak hours of each user base are set to 2 hours, as mentioned in Table 3.6. Number of users in peak and off-peak hours are set with fixed ratio for all user bases separately according to the geographical location.

## 4. RESULTS AND DISCUSSION

Table 5 shows the results of three scenarios in an experiment performed. In each scenario, 8 difference values of variables as number of users and three different load balancing algorithms are used. All scenarios in experiments are performed on cloud analyst simulation tool. This tool generates output in the form of response time and processing time. The output results report generated on cloud analyst shows minimum, maximum and average response time and processing time for each experiment separately. But in this research, only overall average values of response time and processing time are used for evaluation of performance. All scenarios performed in the experiment are combined in the form of Table 5. Also two graphs are generated on the bases of the above combined table which presents response time and processing time in the form of line graphs.

**Table 5.** Combined table of results values of all scenarios with different algorithms

| Load balancing algorithms | User bases | Number of avg. peak users | Overall avg. response time (ms) | Overall avg. data center processing time (ms) |
|---|---|---|---|---|
| **Scenario 1** | | | | |
| **Throttled** | UB1, UB2, UB3, UB4, UB5, UB6 | 20000 | 671.49 | 379.05 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 50000 | 1074.91 | 777.07 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 100000 | 1799.62 | 1499.36 |

| | | | | |
|---|---|---|---|---|
| | UB1, UB2, UB3, UB4, UB5, UB6 | 200000 | 3102.21 | 2805.19 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 370000 | 5538.60 | 5237.74 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 450000 | 6823.11 | 6518.89 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 600000 | 9136.38 | 8830.14 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 850000 | 13325.67 | 13013.60 |
| **Scenario 2** | | | | |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 20000 | 836.60 | 543.98 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 50000 | 1640.50 | 1342.73 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 100000 | 3016.18 | 2720.33 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 200000 | 5996.79 | 5696.87 |
| **Round Robin** | UB1, UB2, UB3, UB4, UB5, UB6 | 370000 | 11335.52 | 11029.16 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 450000 | 14528.07 | 14218.24 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 600000 | 20424.18 | 20107.69 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 850000 | 29310.05 | 28983.29 |
| **Scenario 3** | | | | |
| **Equally** | UB1, UB2, | 20000 | 836.79 | 544.17 |
| **Spread Current Execution Load** | UB3, UB4, UB5, UB6 | | | |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 50000 | 1587.61 | 1293.66 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 100000 | 3015.74 | 2719.91 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 200000 | 5996.91 | 5696.99 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 370000 | 11340.88 | 11034.53 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 450000 | 14535.57 | 14225.74 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 600000 | 20424.23 | 20107.73 |
| | UB1, UB2, UB3, UB4, UB5, UB6 | 850000 | 29341.16 | 29014.38 |

Table 5 clearly shows the overall average response time and overall average processing time of all user bases with randomly increasing numbers of users. There are three portions of the above table and each portion shows different values of response time and processing time according to three different load balancing algorithms and load of users.

Firstly, in this experiment in scenario 1, Throttled load balancing policy is used. Similarly, Round Robin and Equally Spread Current Execution Load algorithms in Scenario 2 and Scenario 3 respectively. This table shows the overall average response time and processing time of all user bases UB1, UB2,….to…, UB6 by varying number of users 20000, 50000…to…, 850000 in all scenarios as mentioned in the above table. For example in scenario 1, for 20000 number of users, the response time is 671.49 ms and processing time is 379.05 ms. On the other hand, in scenario 2, for same number of users response time is 836.60 ms and processing time is 543.98 ms and in scenario 3, response time is 836.79 ms and processing time is 544.17 ms. It is observed that by using Throttled load balancing algorithm in scenario 1 there is huge difference in response time and processing time as compared with other 2 scenarios. But it is found that the response time and processing time are almost the same in scenario 2 and scenario 3 with Round

Robin and Equally Spread Current Execution Load algorithms respectively. Similarly, other variations in response time and processing time can be seen by changing the load as increase in number of users. It is obvious from the above table that Response time and Processing Time is less by using Throttled load balancing algorithm which means that the performance of cloud network is better with Throttled load balancing algorithm.

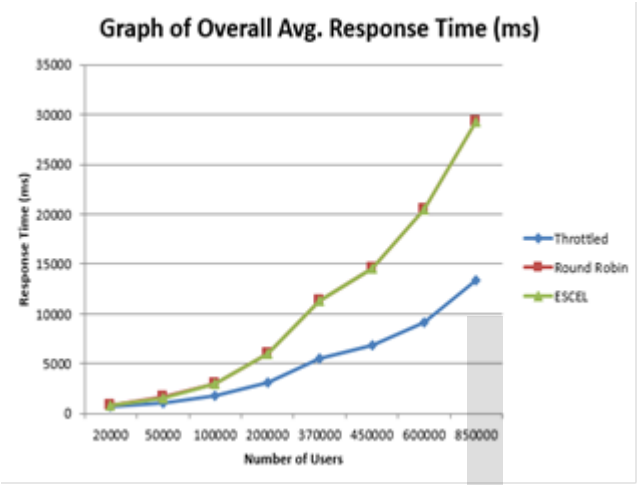


**Figure 7.** Performance in case of response Time

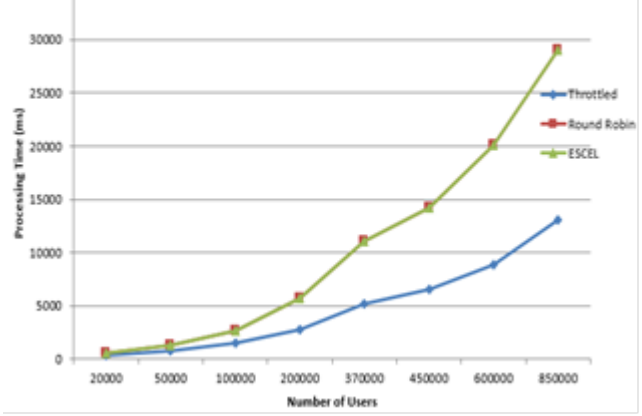


**Figure 8.** Performance in case of Processing Time

The Graph in Figure 7 shows the overall average response time in accordance with number of users and with three load balancing algorithms. The gradual upward movement of graph curves showing increase in response time with the increase in number of users as workload on data centers. The effect of three different load balancing algorithms can also be clearly observed from the above graph curves. The performance of using Throttled load balancing algorithm can be seen clearly with the decrease in response time as compared to other two load balancing algorithms, the curve of which of them goes side by side and showing response time nearly doubled than Throttled algorithm. Round Robin and Equally Spread Current Execution load balancing algorithm's curve in the graph shows that the response time of using these algorithms is nearly same looking slight difference in response time. Performance is depends on distribution of loads on data centers and load balancing algorithms are responsible to distribute the load and to optimize the cloud resources to increase performance.

The Graph in Figure 8 shows the overall average processing time of data centers to process the user's requests using different load balancing algorithms. There is also an upward movement of graph curve gradually showing the increase in data center processing time with the increase in load as number of users. From the above graph shown in Figure 8, the effect of three different load balancing algorithms can be seen clearly, which shows the processing time is less with same load on data centers as compared to other two load balancing algorithms which are Round Robin and Equally Spread Current Execution load. The processing time is almost doubled by using Round Robin and Equally Spread Current Execution load as load balancing algorithms than Throttled algorithm. Throttled algorithm showing better performance in terms of processing time as compared to other two load balancing algorithm. By using Throttled algorithm, resources are optimized in better way rather than Round Robin and Equally Spread Current Execution load algorithms. These two algorithms showing performance in terms of processing time nearly same shown with the graph curve.

## 5. CONCLUSION

The users which are closer to the data center showing the better performance in form of response time. Performance can be enhanced using different load balancing algorithms as using Throttled algorithm, the response time is better than using others like Round Robin and Equally Spread Current Execution Load. For attaining the improvements to be more effective, data center capacity must be increased to fulfill the demand of users in peak hours.

By observing the above mentioned results and discussion, it is concluded that the response time of user requests and processing time of data center are better for the user bases which are located closer to the region where data center is located. It is also observed and concluded from this experiment, UB1 has better response time which is located in the same region as the data center. By using different load balancing algorithms, it is also experimented and achieved results and on the basis of these results we can find out the better algorithm with respect to better performance of cloud. Also the number of requests, having time out issues, experienced by the increase in number of users on data centers. Experiments are performed with most widely used load balancing algorithms for performance evaluation of cloud network. These algorithms are Throttled, Round Robin and Equally Spread Current Execution Load. Response time for every user base is directly related to two things, one is the distance of user base from data center and the other is the load balancing algorithm. The purpose of load balancing algorithm is to maximize the utilization of resources and satisfaction of the clients and significantly enhance the performance of the cloud network. Furthermore, by implementing the load balancing algorithm, the goal of minimizing the response time and to reduce number of rejection of jobs requests can be achieved also that will definitely enhance the performance of the cloud. It is concluded that proposed Throttled load balancing algorithm is found to be the best algorithm among the others to get maximum performance by reducing response time.

## ACKNOWLEDGMENT

First and foremost, I must acknowledge my limitless thanks to **ALLAH**, the Ever-Magnificent, the Ever-Thankful, for His help and bless. I owe a deep debt of gratitude to my university for giving me an opportunity to complete this work. I owe a great many thanks to a great many people who helped and supported me during this research.

I would like to show my deepest gratitude to my supervisor, **Mr. Tahir Abdullah (Lecturer, Govt. College University Faisalabad)**, who's encouragement, guidance and support enabled me to develop an understanding of the subject and to carry out the research.

I take immense pleasure in thanking **Mr. Tahir Abdullah (Lecturer, GCUF)** for all his valuable cooperation in research work. Finally, yet importantly, I would like to express heartfelt thanks to my beloved parents for their blessings and my friends and classmates for their help and wishes.

## REFERENCES

[1] Peter Mell, T. G, "The NIST Definition of Cloud Computing," *National Institute of Standards and Technology Special Publication 800-145 7 pages (September 2011)*.

[2] Pritam Fulsoundar, R. I, "Prediction of Performance Degradation in Cloud Computing," *International Journal of Advanced Computer Research Volume-3 Number-4 Issue-13 December-2013*, 126-129.

[3] Pradeep Singh Rawat, G. S, "Performance Analysis of Cloud Computing Environment and Comparision with IaaS model using CloudAnalyst," *Proc. of the Intl. Conf. on Advances in Electronics, Electrical and Computer Science Engineering – EEC 2012* (pp. 550-554). New York: Institute of Research Engineers and Doctors.

[4] Ahmad M. Manasrah, T. S, "A Variable Service Broker Routing Policy for data," *Journal of King Saud University – Computer and Information Sciences, 2016*.

[5] Gorelik, E, "Cloud Computing Models," *Working Paper CISL# 2013-01 January 2013 Composite Information Systems Laboratory (CISL) Sloan School of Management, Room E62-422 Massachusetts Institute of Technology Cambridge, MA 02142*.

[6] Khan, F. A, "Web Server Performance Evaluation in Cloud Computing and Local Environment," *School of Computing Blekinge Institute of Technology SE – 371 79 Karlskrona , Sweden*.

[7] Arfan Shahzad, N. A, "Opportunity and Challenges using the Cloud Computing in the case of Malaysian Higher Education Institutions," *Proceedings of 6th Annual American Business Research Conference, At Sheraton LaGuardia East Hotel, New York, USA,*. New York: ResearchGate, 2014.

[8] Niloofar Khanghahi, R. R, "Performance monitoring and analysis of cloud computing environment," *International Conference on Electrical, Communication, Computer, Power, and Control Engineering (ICECCPCE), 2013* (pp. 185-190). New Jersey, USA: IEEE.

[9] Nelson L. S. da Fonseca, R. B, "*Cloud Services, Networking, and Management,*" New Jersey: Wiley, 2015.

[10] Mohammed Maqsood Ali, M. H, "Cloud Computing for Retailing Industry: An Overview," *International Journal of Computer Trends and Technology (IJCTT) – Volume 19 Number 1 – Jan 2015*, 51-56.

[11] Kees Blokland, J. M, "Testing Cloud Services: How to Test SaaS, PaaS

& IaaS," *San Rafael: Rocky Nook, Inc, 2013*.

[12] Alexandru Iosup, S. O, "Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing," *IEEE Transactions on Parallel & Distributed Systems, 2011* .

[13] Elisabeth Stahl, L. D, "Performance Implications of Cloud Computing," *ibm.com/redbooks*, 1-28, 2012.

[14] Veerawali Behal, A. K, "Comparative Study of Load Balancing Algorithms in Cloud Environment using Cloud Analyst," *International Journal of Computer Applications (0975 – 8887)*, 36-40, 2014.

[15] Rajeev Kumar, T. P, "Performance Analysis of Load Balancing Algorithms in Cloud Computing," *International Journal of Computer Applications (0975 – 8887)*, 19-27, 2015.

[16] Pooja, N. K, "Performance Evaluation of Cost-Time Based Workflow Scheduling Algorithms in Cloud Computing," *International Journal of Advanced Research in Computer Science and Software Engineering*, 437-439, 2013.