CLOUD COMPUTING PERFORMANCE EVALUATION: ISSUES AND CHALLENGES

Niloofar Khanghahi and Reza Ravanmehr

Department of Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran

ABSTRACT

Cloud Computing makes the dream of computing real as a tool and in the form of service. This internet - based ongoing technology which has brought flexibility, capacity and power of processing has realized service- oriented idea and has created a new ecosystem in the computing world with its great power and benefits. Cloud capabilities have been able to move IT industry one step forward. Nowadays, large and famous enterprise have resorted to cloud computing and have transferred their processing and storage to it. Due to popularity and progress of cloud in different organizations, cloud performance evaluation is of special importance and this evaluation can help users make right decisions. In this paper, we provide an overall perspective on cloud evaluation criteria and highlight it with help of simulation. For this purpose, we present different major factors in cloud computing performance and we analyze and evaluate cloud performance in various scenarios considering these factors.

KEYWORDS

Cloud Computing, Performance evaluation, Efficiency, virtualization, load balancing

1. INTRODUCTION

The term “cloud computing” is not a new concept for the users of computer’s world and the concept dates back the last decades and when John MacCarty predicted that computers might be used one day as a public utility [1-5].

The confluence of technological advances and business development in internet broadband, web services, computer systems and applications has created complete storm for cloud computing during the past decade [1-5]. Nowadays, cloud is the best solution for people who are looking for rapid implementation methods [11].

A more accurate and scientific definition of cloud computing, models, features, performance criteria and simulation, will be reviewed in other key considerations of this paper.
The Structure of other parts of this paper is organized as follows: In Section 2, related works, simulation, evaluation methods and tool are discussed. Section 3 introduces cloud computing briefly. In section 4 the main purpose of this paper is provided. In this section, factors and criteria that is used in simulation will be introduced. Finally in Section 5, using Cloud Analyst tool, the performance of cloud in different scenarios will be evaluated. At the end, conclusions and future works are listed.

DOI : 10.5121/ijccsa.2013.3503

2. RELATED WORK

Simulation in wide environments such as cloud computing environments may be done in different categories, different environments or different criteria. The references in this study are in the following categories and have been conducted since 2009 onwards:

Performance evaluation based on different criteria, evaluation and simulation is usually performed based on criteria in accordance with goals and the results are also studied for proving goal. In studies related to cloud computing performance evaluation , some criteria such as average waiting time, load balancing and number of requests [6], cost and throughput [7, 20], workload [8], the rate of transactions [9], response time [9,18], time of allocation and release of resources [10], different scheduling algorithms [10, 12, 18], effectiveness, delays in service and productivity [12], the number of input and output operations in the network [15], etc. are studied.

Evaluation based on different methods, tools and simulations for cloud environments and other web-based environments, there are various methods and simulation tools, such as parto traffic methods[6], fuzzy systems [11], a discrete event simulation [12], a tool like CloudAnalyst [14] and etc.

Performance evaluation based on specific applications or services, Variety of services and applications are offered in cloud environments each of which can be topic for the simulation scenarios, such as scientific computing [8,17], e-learning software [11], high-performance computing [13,18], inbound and outbound applications on the network [15], services with assuring quality of service [16], Multi-Tier Cloud Applications [21] and etc.

Performance evaluation based on different environments, nowadays, large and well-known organizations like Google, Oracle, IBM and etc provide their cloud environments with their own characteristics and architectures, and on the other hand, different users use these services. So, the actual cloud environments such as Google, Amazon EC2 [9,10], GoGrid [8], Elastichosts [8], Mosso [8], Amazon Web Service [19], Azure [20] and etc. are suitable for the evaluation, or performance evaluation may be based on different providers too [7].

Evaluation in this paper is based on combination of the above categories, and is performed with the help of CloudAnalyst applications [14].

3. CLOUD COMPUTING

Cloud computing is a type of parallel, virtual, distributed, configurable, and flexible systems, which refers to provision of applications such as hardwares and softwares in virtual data centers via internet [10].cloud computing services are configurable and customers pay fees based on the use of resources and services [1-5].

The most important element of cloud structure is server which is the brain behind the whole processes in cloud. Cloud is the important model for access to distributed computing resources [1-5].

Pay per use, scalability, use the Internet technology, self-service based on the demand, high performance, quick to implement, easy to maintain and update are key benefits of cloud computing [1-5].

And the data recovery, lack of control over cloud services, service level agreements, legal problems, different architectures, audit, Reviews and evaluation of the performance cloud computing environment are the major disadvantages of cloud computing [1-5].

3.1. Cloud Computing Service Delivery Models

There are three models for delivery of cloud services as follows:

Software as a Service (SaaS), in this model, users use the launched application on cloud infrastructure. Interfaces for these applications are browsers, and don't require installation. Gmail is the best known example of this model [1-5].

Platform as a Service (PaaS), in this model, users rented platforms or operating systems and they can expand their required programs on it. The most famous example of this model is Google App Engine [1-5].

Infrastructure as a Service (IaaS), this model is associated with a virtual engine [16] and users can access to infrastructures with virtual machine [12].

3.2. Cloud Computing Deployment Models

The decision on implementation of cloud is important. There are four main cloud deployment models as follows:

Public, the most common model is the cloud deployment model. Large Enterprise is Owner of a large cloud infrastructure and services to users.

Private, this model simulates a private network. It is just for an organization's infrastructure.

Community, in this model, some enterprises which have common policies, goals and concerns share infrastructure of cloud.

Hybrid, this model is a combination of two or more cloud deployment models. In this model, resource management may be internal or external [1-5].

3.3. Process of a Request in Cloud

When a request of service is given by a user to cloud, it passes a special trend until acceptance and run or rejection. This process is shown in figure 1.

As can be seen in figure 1, any entry request may be placed in one of 3 situations after entering into cloud servers which is described in following:

- Running or Serving
- Waiting in buffers
- Rejecting the request because buffer is full or inapplicable, figure 2, shows possible transmission states.

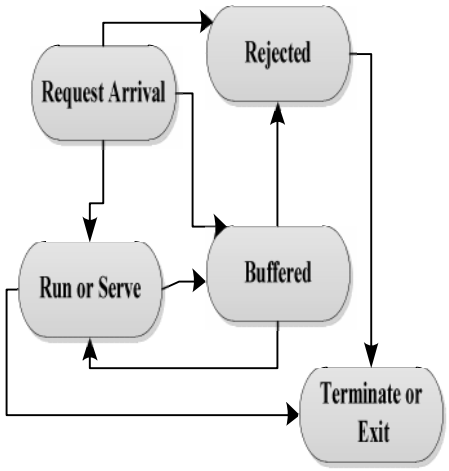


Figure 1. Possible Transmission States for any request

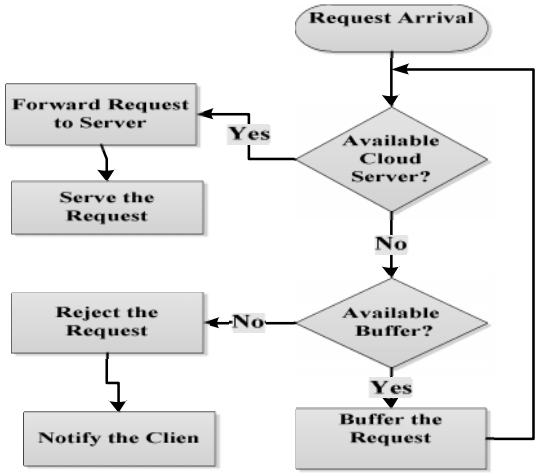


Figure 2. Process of Request into Cloud [6]

4. CLOUD COMPUTING PERFORMANCE EVALUATION

Cloud computing resources must be compatible, high performance and powerful. High performance is one of the cloud advantages which must be satisfactory for each service[1-5].

Higher performance of services and anything related to cloud have influence on users and service providers. Hence, performance evaluation for cloud providers and users is important. There are many methods for performance prediction and evaluation; we use the following methods in our evaluation:

- Evaluation based on criteria and characteristics
- Evaluation based on simulation

Another category which can be considered for evaluating cloud performance is classification of three layers of cloud services evaluation.

4.1. Factors affective on performance

Nowadays , the term “performance” is more than a classic concept and includes more extensive concepts such as reliability, energy efficiency, scalability and soon. Due to the extent of cloud computing environments and the large number of enterprises and normal users who are using cloud environment, many factors can affect the performance of cloud computing and its resources. Some of the important factors considered in this paper are as follows:

- Security, the impact of security on cloud performance may seem lightly strange, but the impact of security on network infrastructure has been proven. For example, DDoS attacks have wide impact on networks performance and if happen, it will greatly reduce networks performance and also be effective on response time too. Therefore, if this risk and any same risks threaten cloud environment, it will be a big concern for users and providers.

- Recovery, when data in cloud face errors and failures or data are lost for any reason, the time required for data retrieval and volumes of data which are recoverable, will be effective on cloud performance. For example, if the data recovery takes a long time will be effective on cloud Performance and customer satisfaction, because most organizations are cloud users and have quick access to their data and their services are very important for them.
- Service level agreements, when the user wants to use cloud services, an agreement will be signed between users and providers which describes user's requests, the ability of providers, fees, fines etc. If we look at the performance from personal view, the better , more optimal and more timely the agreed requests , the higher the performance will be .This view also holds true for providers.
- Network bandwidth, this factor can be effective on performance and can be a criterion for evaluations too. For example, if the bandwidth is too low to provide service to customers, performance will be low too [19].
- Storage capacity, Physical memory can also be effective on the performance criteria. This factor will be more effective in evaluating the performance of cloud infrastructure[17].
- Buffer capacity: as shown in figure 2, if servers cannot serve a request, it will be buffered in a temporary memory. Therefore, buffer capacity effect on performance. If the buffer capacity is low, many requests will be rejected and therefore performance will be low.
- Disk capacity, can also have a negative or positive impact on performance in cloud[17].
- Fault tolerance, this factor will have special effect on performance of cloud environment. As an example, if a data center is in deficient and is able to provide the minimum services, this can increase performance.
- Availability, with easy access to cloud services and the services are always available, performance will be increase.
- Number of users, if a data center has a lot of users and this number is greater than that of the rated capacity, this will reduce performance of services.
- Location, data centers and their distance from a user's location are also an important factor that can be effective on performance from the users' view.

Other factors that can affect performance which are as follows:

- Usability
- Scalability
- Workload
- repetition or redundancy
- Processor Power
- Latency [19, 22]

4.2. Performance Evaluation Criteria

There is a series of criteria for evaluation of all factors affecting performance of cloud computing some of which will be used in this paper. These criteria are under development. Some of these criteria have been selected considering the importance and criteria in simulation. It should be mentioned that all of criteria listed in pervious sections cover the factors mentioned in the previous section but some of the factors will be important in special criteria:

- Average response time per unit time, this criterion will cover all factors completely[12, 16].
- Network capacity per second (Mbps)or unit time, the most important factor associated with this criterion is network bandwidth ,availability and scalability.
- The number of I / O commands per second(IOPS)or unit time

- Average waiting time per unit time [6,18]
- Workload(requests) to be serviced per second(Mbps) or a unit of time [6]
- Throughput (Req / Sec), this criterion will be recovered recovery, buffering capacity and processing power factors [15, 20].
- The average time of processing(exe / sec)
- Percentage of CPU utilization [15, 21]
- The number of requests executed per unit time
- The number of requests per unit time buffer
- The number of rejected requests per unit time

5. SIMULATION IN CLOUD COMPUTING ENVIRONMENTS

Research and evaluation of wide environments usually associated with simulation. Also cloud computing is no exception of this rule, because research on total context of internet is too difficult and involves interaction with multiple computing and network elements, which may not be under control of developers. Moreover, network conditions may not be controllable and predictable and this will affect evaluation[15].

This section will cover evaluation and simulation of cloud environment. For simulation, we will use CloudAnalyst tool which is a graphical design based on cloudsim [14]. It is necessary to note that scenarios in CloudAnalyst are controllable and repeatable and do not require programming.

5.1. Simulation components

In simulation with CloudAnalyst tools, there are two main components which are introduced below. It should be noted that each of these two components is configurable.

Data centers or DC, this component shows the hardware configuration, which including processors, storage media, internal memory ,bandwidth etc. These data centers can be defined in different geographic areas. Besides the number of processors in data centers, VM's speed and is also configurable [14].

Users have been distributed geographically as user groups. The requests are configurable for users such as geographical area, number of requests per hour, etc. [14]. It should be noted that each user can symbolize a person, an organization or a group.

There are different policies for scheduling in these tools, such as nearest data center, optimizing response time and dynamic configuration. The assumptions used in the simulation are described in the following:

Type of requests based on data size is assumed datacenter and also 6 geographical regions are defined and both shown in Table 1. Total maximum number of users is 50, simulation time is 24 hours, and data centers policy is based on the closest connection to the to the datacenter.

Table 1. Data size assumptions and Geographic regions

Data size assumptions		
100 – 500 Byte	Hybrid requests (computing, networking, I / O, memory)	
500 – 1000 Byte	Requests related to memory	
1000 – 1500 Byte	Requests related to CPUs and computing	
1500 – 2000 Byte	Requests related to transmission and network	
2000 – 2500 Byte	Requests related to storage and retrieval and data access	
Geographic regions		
Region Number	Region	Time Zone
0	N.America	GMT - 6.00
1	S.America	GMT - 4.00
2	Europe	GMT + 1.00
3	Asia	GMT + 6.00
4	Africa	GMT + 2.00
5	Oceania	GMT + 10.00

5.2. Simulation Metrics

The following metrics of CloudAnalyst are used in this simulation:

- Minimum, maximum and average overall response time
- Minimum, maximum and average processing time in the overall data center
- Minimum, maximum and average response time per user
- Minimum, maximum and average time per data center
- The total cost of the virtual machine
- Cost per VM of Data Center
- Cost of data in each data center
- Total cost in each data center

5.3. Simulation Category

There are three following categories in this simulation and evaluation. It should be mentioned that, this category has been selected based on major components in cloud environment. Specific metrics are used in each section and reviewed in chart. These categories have been selected because data centers, users and geographic region are important in cloud computing environments.

Simulation and evaluation based on data centers, in this section, evaluation is done by modifying the virtual machine, memory and bandwidth.

Simulation and evaluation based on users, in this section, we evaluate the results with change in number of users and volume of work.

Simulation and evaluation based on geographical region; in this section, we study geographical location of users and data centers and want to determine how effective they will be on criteria if the data centers and users are in the same region or are far from each other in different regions .

It should be mentioned that only results of some metrics are studied and mentioned for each case of simulation scenario.

5.4. Simulation and Evaluation Based on Data Centers

Scenario 1: in this scenario, the number of users is assumed constant and equal to 50, they are distributed in different regions, and requests are hybrid. All settings related to the data centers are fixed and only the number of data centers is changed from 1 to 20 Centers. The simulation time is 24 hours. Table 2 shows the configuration of the data centers. The results of the simulations are presented below. As can be seen in Table 3, the virtual machine operating system is Linux. This table is shown as some additional information such as some type of cost.

Table 2. Profile of Scenario 1

#VM	Memory	BW	VM	Arch
5	512	1000	Xen	X86
OS	Cost per VM $Hr	Memory Cost $/s	Storage Cost $/S	Data Transfer Cost $/s
Linux	0.1	0.05	0.1	0.1

In Figure 3, the average, minimum and maximum response time is shown. There is little reduction in response time after 10 data centers so putting more than that only increases the cost. It can be concluded that increasing number of data centers is not sign of minimization of response time.

Figure 4 shows the processing time in data centers. As shown in Figure 4, the rate has been fixed in maximum and average state after 10 data centers the rate is fixed and will not have positive impact and only will increase cost. It should be mentioned that, interaction between 5 Data Center is due to the chart level rise for this state, as can be shown in Figure 4.

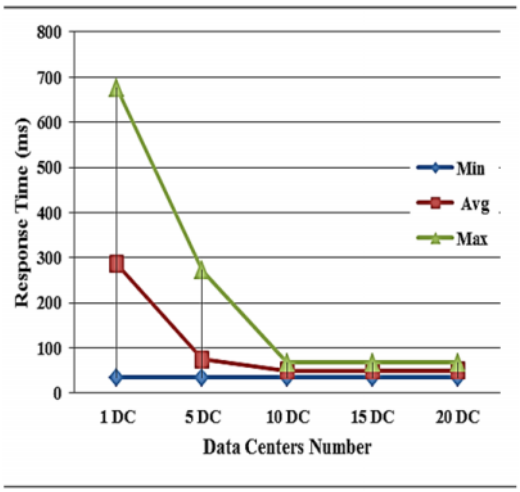


Figure 3. Overall response time in Scenario 1

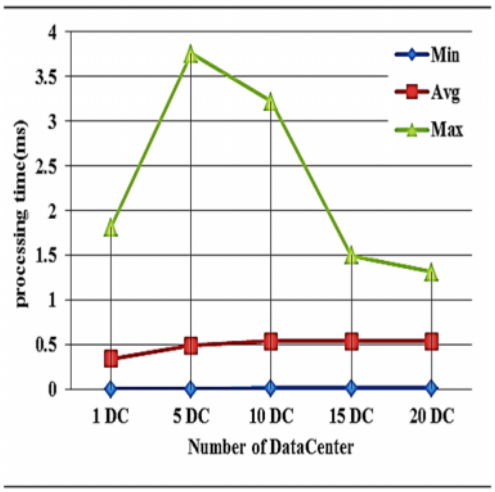
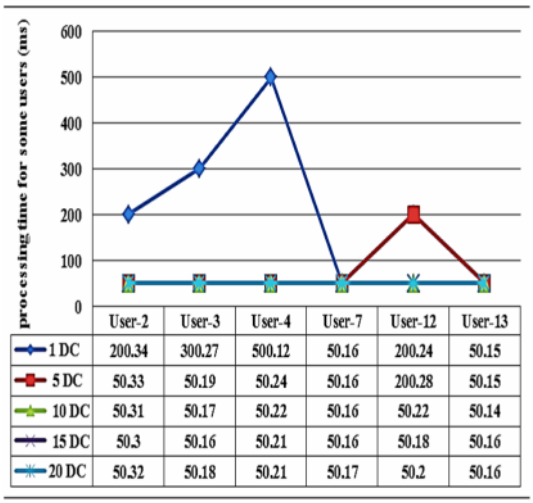


Figure 4. Processing time in data centers

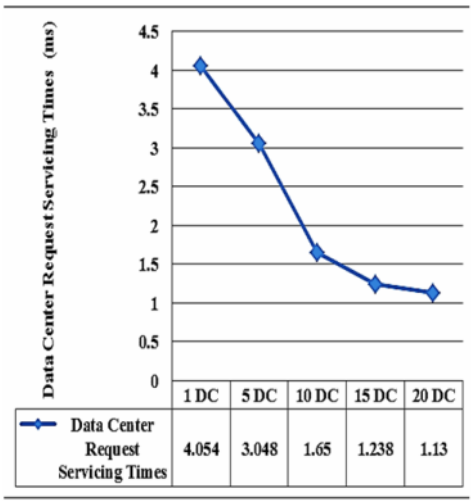
Figure 5, show the response time for some users. As can be seen in figure 5 too, increasing 5 Data Center, will be improved processing time, bud additional Data Center is only extra cost and also for low volume requests, additional Data Center will not Cause significant changes in processing time.

Figure 6, shows the average of maximum service time per request in data centers. It also proved to be similar to previous values after the same number of data centers. This reviews shows that the average service time is decreasing with increasing number of centers, but reduction is lower after some additional data center and costs is too high.



	User-2	User-3	User-4	User-7	User-12	User-13
1 DC	200.34	300.27	500.12	50.16	200.24	50.15
5 DC	50.33	50.19	50.24	50.16	200.28	50.15
10 DC	50.31	50.17	50.22	50.16	50.22	50.14
15 DC	50.3	50.16	50.21	50.16	50.18	50.16
20 DC	50.32	50.18	50.21	50.17	50.2	50.16

Figure 5. response time for some users



	1 DC	5 DC	10 DC	15 DC	20 DC
Data Center Request Servicing Times	4.054	3.048	1.65	1.238	1.13

Figure 6. Servicing time in data centers

Figure 7 shows the total cost and the cost of the virtual machine. As can be seen in Figure 7, with the increase of data centers, cost increases and since rate of other criteria remains fixed after special number, this cost is useless.

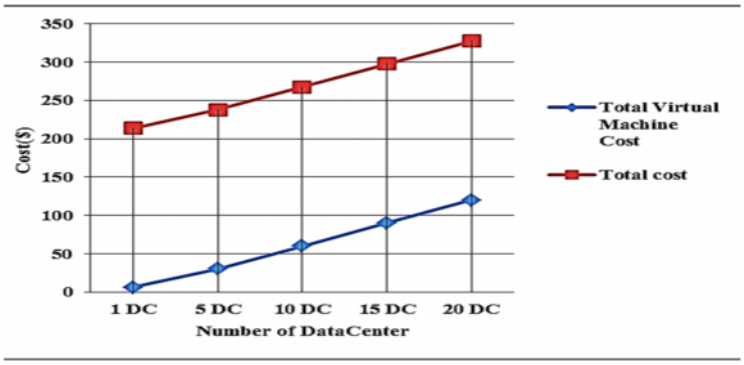


Figure 7. Total and VM Costs

Scenario2: In this scenario, the goal is to evaluate criteria by considering a fixed number of data centers and changing the number of processors, memory, and storage media in simulation. The number of users is 50 and the number of data centers is 6. At first, we increased memory from 1DC to 5DC, then, we did the same work for media storage and processing of 1DC to 5DC. The results are shown in Figure 8 and its interpretation is as follows:

Change in number of processors of data centers has the greatest impact on processing time, and has the greatest impact on cost too, as can be shown in Figure 9.

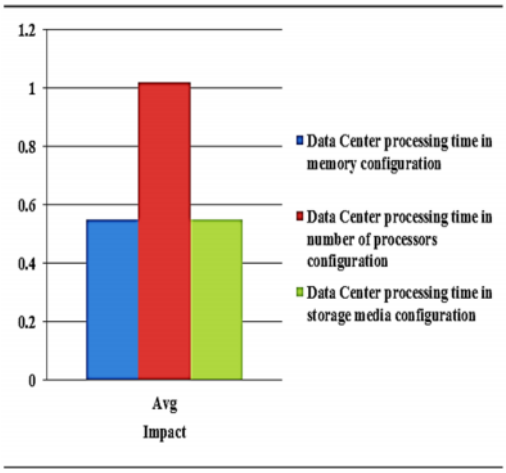


Figure 8. Data Center Processing time in Changes in Scenario2

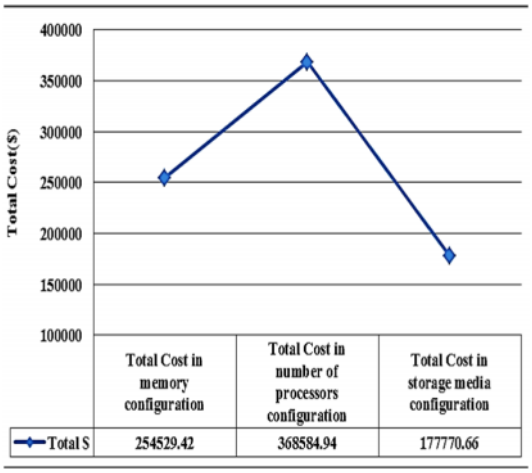


Figure 9. total Cost in Scenario2

5.5. Simulation and Evaluation Based on Users

Scenario3: In this scenario, the goal is to change the number of users and evaluate the results. In order to get more accurate results for the scenario, only one data center was considered and the number of users will increase to 10, 20, 30, 40 to 50, who are located in the same area with data center. It is necessary to note that capacity of authorized user was considered 30 for each data center. Figure 10 shows the processing time of the data center based on number of users. As can be seen, increasing the number of users increases processing time. Figure 11shows the total cost and data transmission costs, the cost will be increased strongly with increasing number of users which are not affordable. It can be concluded that if a data center user is overrated capacity, not only it will not be profitable but also it will lower efficiency of that center.

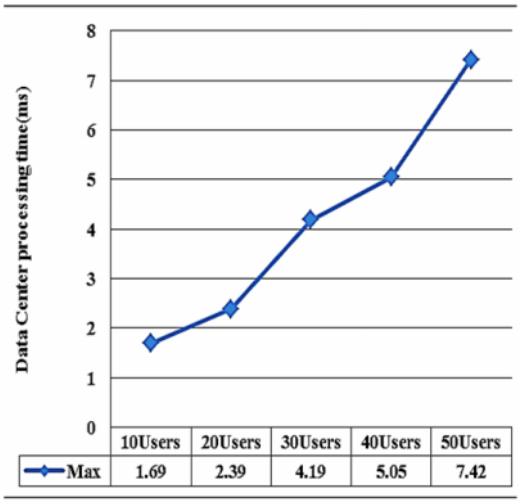


Figure 10. Max Data Center processing time

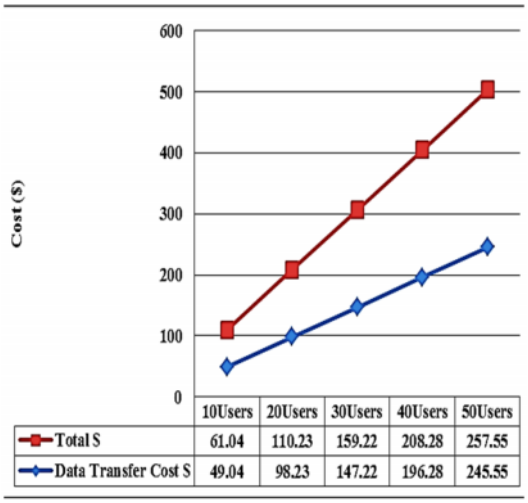


Figure 11. Total Cost in Scenario3

Scenario4: in this scenario,12 users are assumed and every 2 users have 100, 500, 1000, 1500, 2000 or 2500 requests per unit time. Users with the same number of requests are connected to one data center. Other configurations are fixed. The results show that increasing the number of requests per unit times have little impact on response time, on processing time, data centers. But unlike other measures, it is effective on the data transfer and thus the cost of data transfer. More information can be transferred with the increasing number of requests and therefore, costs will also increase. This result is shown in Figure 12.

In addition, increasing the size of the overall response time is also relatively effective. This result is shown in Figure 13.

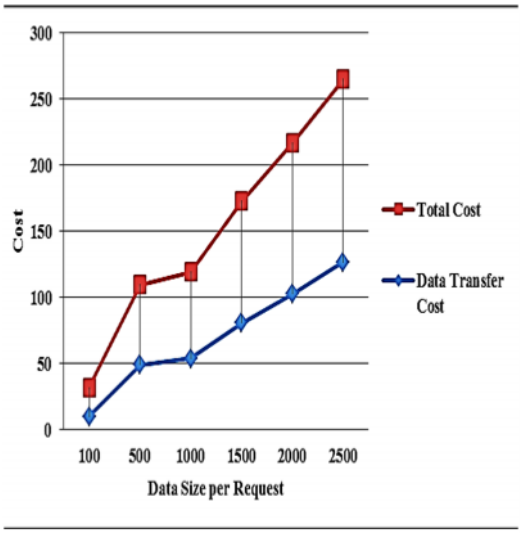


Figure 12. Total Cost in Scenario4

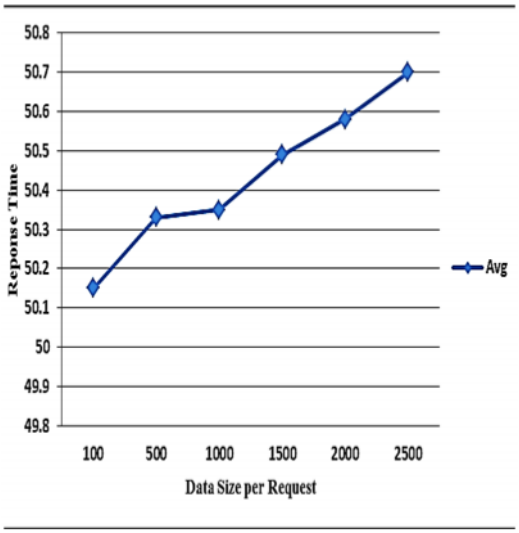


Figure 13. Avg Response time based on request size

5.6. Simulation and Evaluation Based on Geographical Region

Scenario5: In this scenario, we can evaluate the position of the data center and users. For this section, we have considered 18 users and 3 data centers, which are placed in three cases.

In the first case, all users are placed in a single region and all three data centers are in another region separate from the users. In the second cases, users and data centers are located in the same region and distributed users and data centers have been distributed in the third case.

The results show that these changes affect cost and other measures. The overall response time is shown in Figure 14. This result shows that it is better for the users and data centers to be in the same region or have the least distribution. As can be seen in Figure 15, Processing rate in data centers will be reduced when user is away from the center, because the response time increases, so users may have fewer requests from the data center.

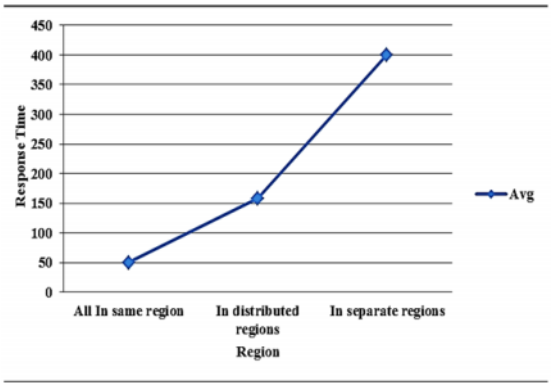


Figure 14. Response time based on geographical Region

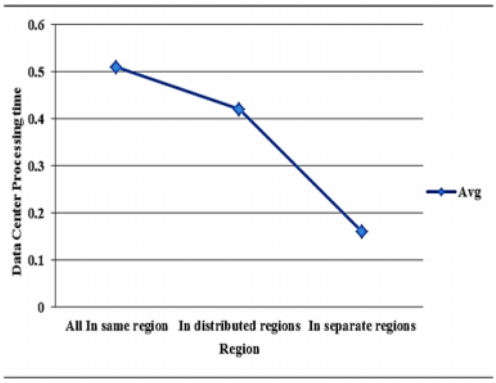


Figure 15. Process time of data center based on geographical region

6. CONCLUSION AND FUTURE WORK

According to prediction and evaluation of cloud computing performance, we can reach different conclusions. As an example, increasing power and speed of the data center is not always efficient, and sometimes it only has additional costs. So, one should not expect to increase efficiency more than what was required and should find standard based on requests and user types. Distribution of data centers and use of the closest data center is better and more optimal. Due to increase in development of cloud computing, it is predicted that storage and computing on personal computers will be forgotten and all of these things will be transferred into the Clouds. Therefore, architecture and evaluation of the optimal and efficient data centers should be performed for the future of computing world through suitable prediction.

According to the review and evaluation performed in the field of performance, cloud computing still has shortages in performance evaluation and special measures are required for this work. It is better to consider delay in evaluations or implement a criterion for evaluation of service level agreement because these agreements are the most important for the users and one can present more accurate evaluation in future by specifying type of user's request or specifying and distinguishing all users.

REFERENCES

[1] Borko Furht & Armando Escalante,(2010) "Handbook of Cloud Computing", Springer,
[2] Abah Joshua & Francisca N. Ogwueleka,(2013) "Cloud Computing with Related Enabling Technologies," International Journal of Cloud Computing and Services Science (IJ-CLOSER), Vol.2, No.1, pp. 40~49
[3] NIST Advisory Working Group, (2011) "NIST Cloud Computing Standards Roadmap", NIST Special Publication 500-291
[4] Peter Mell & Timothy Grance, (2011) "The NIST Definition of Cloud Computing", NIST Special Publication 800-145
[5] M.Malathi, (2011) "Cloud Computing Concepts" IEEE
[6] AYMAN G. FAYOUMI, (2011) "PERFORMANCE EVALUATION OF A CLOUD BASED LOAD BALANCER SEVERING PARETO TRAFFIC" Journal of Theoretical and Applied Information Technology, Vol. 32 No.1
[7] Sergej Poltorak, (2011) "Cloud Computing: Meet the Players. Performance Analysis of Cloud Providers", BASEL UNIVERSITY COMPUTER SCIENCE DEPARTMENT

[8] Alexandru Iosup & Simon Ostermann & Nezih Yigitbasi (2010) "Performance Analysis of Cloud Computing Services for Many-Tasks Scientific Computing", IEEE TPDS, MANY-TASK COMPUTING,
[9] Vladimir Stantchev, (2009) "Performance Evaluation of Cloud Computing Offerings"; Third International Conference on Advanced Engineering Computing and Applications in Sciences IEEE
[10] Nezih Yigitbasi, (2009) "C-Meter: A Framework for Performance Analysis of Computing Clouds", IEEE/ACM International Symposium on Cluster Computing and the Grid
[11] Mohammed Alhamad, (2011) "A Trust-Evaluation Metric for Cloud applications", International Journal of Machine Learning and Computing, Vol. 1, No. 4
[12] Ioannis A. Moschakis & Helen D. Karatza , (2011) "Performance and Cost evaluation of Gang Scheduling in a Cloud Computing System with Job Migrations and Starvation Handling", IEEE
[13] Keith R. Jackson & Krishna Muriki, (2010) "Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud", 2nd IEEE International Conference on Cloud Computing Technology and Science
[14] Bhathiya Wickremasinghe, (2010) "CloudAnalyst: A CloudSim-based Visual Modeller for Analysing Cloud Computing Environments and Applications"; 24th IEEE International Conference on Advanced Information Networking and Applications
[15] Yiduo Mei & Ling Liu, (2011) "Performance Analysis of Network I/O Workloads in Virtualized Data Centers", IEEE TRANSACTIONS ON SERVICE COMPUTING
[16] Kaiqi Xiong.(2009) "Service Performance and Analysis in Cloud Computing"; IEEE
[17] Simon Ostermann & Alexandru Iosup,(2010) "A Performance Analysis of EC2 Cloud Computing Services for Scientific Computing", Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (LNICST)
[18] Ioannis A. Moschakis & Helen D. Karatza, (2010) "Evaluation of gang scheduling performance and cost in a cloud computing system" , Springer
[19] Keith R. Jackson & Krishna Muriki, (2010), "Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud", IEEE International Conference on Cloud Computing Technology and Science
[20] Donald Kossmann & Tim Kraska & Simon Loesing, (2011), "An Evaluation of Alternative Architectures for Transaction Processing in the Cloud", ACM
[21] Arshdeep Bahga, Vijay K. Madisetti, (2013) "Performance Evaluation Approach for Multi-Tier Cloud Applications", Journal of Software Engineering and Applications
[22] Miguel G. Xavier & Marcelo V. Neves & Fabio D. Rossi, (2012), "Performance Evaluation of Container-based Virtualization for High Performance Computing Environments", IEEE PDP