International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 4 Issue 10, October -2015

Mathematical Analysis on Quality of Service in Cloud servers

Arumugaselvi Murugan[1], Halima Aminu[2], Ganesan Subramanian[3]

[1]SRM University, Chennai, Tamilnadu, India
[2,3]Manipal University, Dubai, U.A.E.

ABSTRACT

Clouds computing are the new trend of computing where readily available computing resources are exposed as a service. A cloud is defined as both the applications delivered as services over the internet and the hardware and systems software in the data centers that provide those services. The common characteristics most interpretations share are on-demand scalability of highly available and reliable pooled computing resources, secure access to metered services from anywhere, and displacement of data and services from inside to outside the organization. While aspects of these characteristics have been realized to a certain extent, cloud computing remains a work in progress. Performance optimization is critical to its successful application. In this research, our focus is towards the Quality of Service improvement in cloud server. Mathematical analysis is done using Kendell's notation alongside Microsoft Windows Azure platform. The parameters analyzed are bandwidth, arrival rate and service time. Mozilla Firefox is the browser used throughout the analysis using Firebug to analyze the response time.

Keywords: Cloud computing, QoS, response time, bandwidth.

I. INTRODUCTION

Cloud computing is the Internet-based development and use of computer technology. It has become an IT buzzword and a style of computing paradigm in which typically real-time scalable resources such as files, data, programs, computing, hardware, and third party services can be accessible from a Web browser via the Internet to users (or called customers alternatively) [1]. Cloud computing greatly lowers the threshold for deploying and maintaining applications and infrastructure requirements since it provides Infrastructure as a Service (IaaS), Software as a Service (SaaS) and Platform as a Service (PaaS). [2], By definition, the actual term "cloud computing" originated from computer network analysts who diagram intricate connections among computers in a network. Once the internet came along and computers became connected in countless ways, network analysts would simply draw a cloud denoting that individual computers and networks were connected in an unknown way. Just as a "cloud" in the sky is diffuse and capable of hiding things, a "cloud network" is a diffuse network of computers connected in a hidden fashion. Today, the cloud is literally numerous servers residing in warehouses all around the world (these warehouses are known as "server farms") [3]. However, despite the flourishing developments; cloud computing paradigm necessitates accurate performance evaluation of cloud data centers through Quality of Service (QoS). QoS is referred to as the resource reservation control mechanisms in place to guarantee a certain level of performance and availability of a service. According to a survey conducted by IDC, the benefits versus challenges in cloud computing are: The number one concern about cloud services is **security**; with the businesses' information and critical IT resources outside the firewall, customers worry about their vulnerability to attack. The next two concerns are performance and availability. These are aspects of a broader concern about **cloud services dependability**; whether critical services in the cloud will consistently be there, when and as needed by the business. This research focuses on QoS parameters which are performance and availability parameters. M/M/C: ∞/∞ Queuing Model is used to analyze the response time for 1GB video file for one-server and multiple-server over varying bandwidth. QoS is also measured using windows Azure platform by uploading video file and analyzing the response time.

II. RESEARCH OBJECTIVES

The objective of this research is to:

- Carry out mathematical analysis of QoS using Kendell's notation.
- Identify areas where Quality of Service in cloud can be improved.

- Implementation and testing of QoS using Windows Azure.
- Ascertain and present recommendation on the performance and availability of cloud technology.
-

III. RELATED WORK

There have been a considerable number of researches in Cloud computing though only a diminutive portion of the work done so far has addressed performance issues, and rigorous analytical approach has been adopted by only a handful among these.

In [12], the authors proposed a Queuing Theory based method to predict the performance of the service exposed by the cloud. Although the correctness of the method has been demonstrated by some experiments and simulations, the model they set up is quite simple due to its presumption that a cloud only exposes one service. Actually, the authors only propose a generalized method to analyze and predict the performance of a service. They did not consider the special context of Cloud Computing. In our opinion, the structure of a cloud is like a multiple Queues but not a single Queue.

In [18], they employ the queuing model to investigate resource allocation problems in both single-class service case and multiple-class service case. Furthermore, they optimize the resource allocation to minimize the mean response time or minimize the resource cost in each case.

In addition, some researchers have undertaken the research of the performance analysis. In [19], the author proposed an M/G/m queuing system which indicates that inter arrival time of requests is exponentially distributed; the service time is generally distributed and the number of facility nodes is m.

In [21], the authors proposed an analytical queuing based model for performance management on cloud. In their research, the web applications were modeled as queues and virtual machines were modeled as service centers. They applied the queuing theory models to dynamically create and remove virtual machines in order to implement scaling up and down.

In [22], the authors analyzed the general problem of resource provisioning within cloud computing. In order to support decision making with respect to resource allocation for a cloud resource provider when different clients negotiated different service level agreements, they have modeled a cloud center using the M/M/C/C queuing system with different priority classes. The main performance criterion in their analysis was the rejection probability for different customer classes, which can be analytically determined

In [13], Cheng et al. proposed a queuing-based model for performance management on cloud. The web applications are modeled as queues and the virtual machines as service centers and using queuing theory, dynamically scale up and down web applications on cloud. However, the experiments conducted are not enough to measure the effect of the model on usage of computing resource.

It is inevitable that there is random error between predicted status and real-time one though prediction is a feasible and effective way to optimize the usage of computing resource. So prediction is a risky method which is possible to result in a serious situation.

In [15], Xiong et al presented an approach for studying computer service performance in cloud computing by analyzing the relationship among maximal number of customers, the minimal service resources and the highest level of services. However, the analysis has not been tested on a real cloud platform but mathematically done using some arbitrary values. In this research, the performance indicators were analyzed for varying bandwidth capacity both mathematically using M/M/C: ∞/∞ Queuing system and Microsoft's Windows Azure platform.

IV. QUEUING THEORY

Queuing theory is a collection of mathematical models of various queuing systems. Queues or waiting lines arise when demand for a service facility exceeds the capacity of that facility i.e. the customers do not get service immediately upon request but must wait or the service facilities stand idle and waiting for customers.

The basic queuing process consists of customers arriving at a queuing system to receive some service. If the servers are busy, they join the queue in a waiting buffer. They are then served according to a prescribed queuing discipline, after which they leave the system.

1) Kendell's Notation

Kendall's notation for Queues A/B/C/D/E is the standard system used to describe and classify the queuing model that a queuing system corresponds to. The meanings associated with these letters are:
A - Inter-arrival time distribution
B - Service time distribution
C - Number of servers
D - Maximum number of jobs that can be there in the system (waiting and in service)
E - Queuing Discipline (FCFS, LCFS, SIRO etc.).

By default, D is ∞ for infinite number of waiting positions and E is FCFS
Kendall notation:

```
A | B | c | D | E | F
^   ^   ^   ^   ^   ^
|   |   |   |   |   |
|   |   |   |   |   +--- queuing discipline (default: FIFO)
|   |   |   |   +------- population size (default: infinite)
|   |   |   +----------- queue capacity (default: infinite)
|   |   +--------------- number of servers
|   +------------------- server process
+----------------------- arrival process
```

Fig. II: Representation of Kendell's notation

V. ANALYSIS

During the analysis, 1GB video file was considered with:

- Arrival rate of 100 and 1000 users,
- Number of servers: 1 and 10 servers,
- Bandwidth: 8Mbps and 500Mbps respectively. Assuming that a server can serve one customer at a time, the four cases are:

Case 1: M/M/1:100/∞ model (bandwidth: 8Mbps)

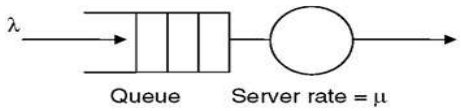

Fig III: M/M/1:100/∞ with bandwidth: 8Mbps

INPUT (unit of time: hour)	
Arrival rate (lamda)	100
Service rate (mu)	0.29
Number of servers	1
OUTPUT	
Mean time between arrivals	0.010
Mean time per service	3.45
Traffic intensity	344.83

Case 2: M/M/10:100/∞ model (bandwidth: 8Mbps)

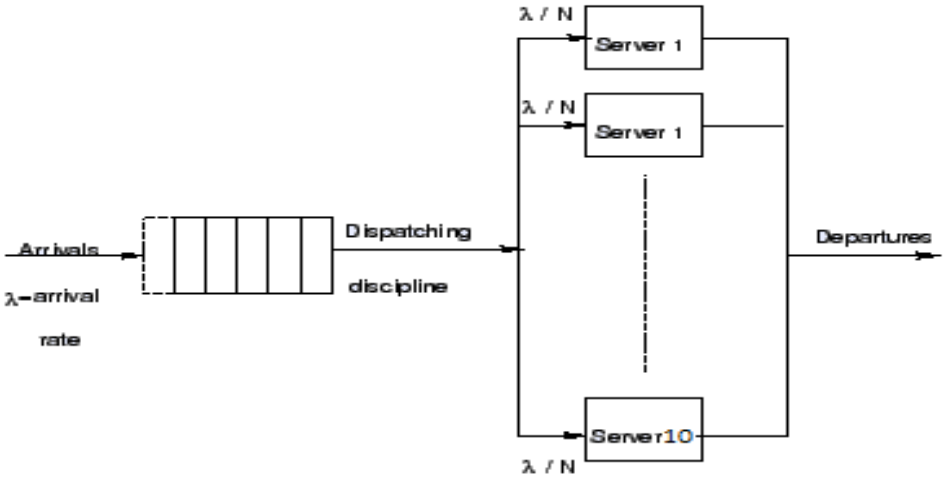

Fig IV: M/M/10:100/∞ with bandwidth: 8Mbps

INPUT (unit of time: hour)	
Arrival rate (lamda)	100
Service rate (mu)	2.86
Number of servers	10
OUTPUT	
Mean time between arrivals	0.010
Mean time per service	0.35
Traffic intensity	3.50

Case 3: M/M/1:100/∞ model (bandwidth: 500Mbps)

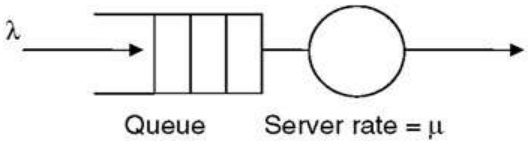

Fig V: M/M/1:100/∞ with bandwidth: 500Mbps

INPUT (unit of time: hour)	
Arrival rate (lamda)	100
Service rate (mu)	1800
Number of servers	1
OUTPUT	
Mean time between arrivals	0.010
Mean time per service	0.00056
Traffic intensity	0.0556

Case 4: M/M/C: ∞/∞ model (bandwidth: 500Mbps)

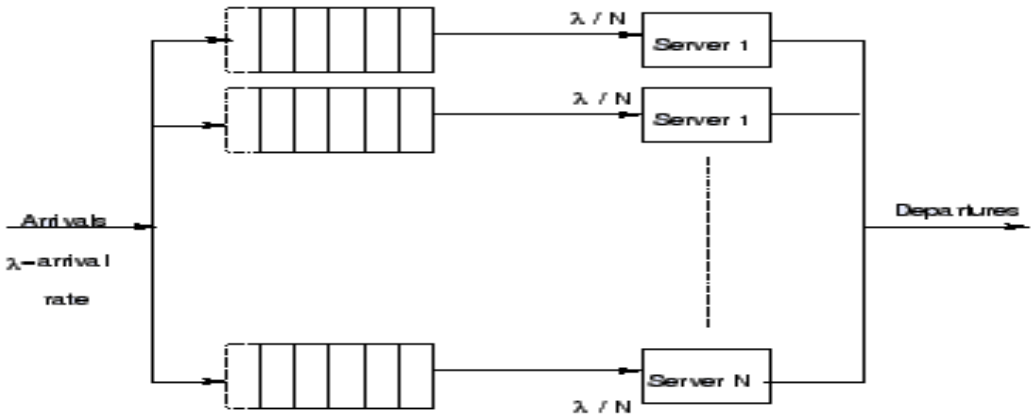

Fig VI: M/M/10:100/∞ with bandwidth: 500Mbps

INPUT (unit of time: hour)	
Arrival rate (lamda)	1000
Service rate (mu)	1800
Number of servers	10
OUTPUT	
Mean time between arrivals	0.001
Mean time per service	0.00056
Traffic intensity	0.0556

Interpretation of results:

Case 1:
Based on Hall's queuing rule of thumb, it is suggested that

- The number of servers is larger than 364
- The absolute minimum number of servers based on Teknomo's Queuing rule of thumb is 345 (at worst level of service)

Case 2:
Based on Hall's queuing rule of thumb, it is suggested that

- The number of servers is larger than 41
- The absolute minimum number of servers based on Teknomo's Queuing rule of thumb is 35 (at worst level of service)

Case 3:
Queuing Intensity = 0.056
Queuing Utilization = 5.556%
Queue Length in Queue = 0.003
Queue Length in System = 0.059
Delay in Queue = 0.000
Delay in System = 0.001
Probability of idle server = 94.444%
Probability distribution that there are n customers in the system:

Case 4:
Queuing Intensity = 0.556
Queuing Utilization = 5.556%
Queue Length in Queue = 0.000
Queue Length in System = 0.556
Delay in Queue = 0.000
Delay in System = 0.001

Probability of idle server = 57.375%
Probability distribution that there are n customers in the system:

The next analysis was done by uploading a video file in Microsoft's Azure. By varying bandwidth capacity, the file was accessed and the response time was recorded. The browser used throughout the experiment is Firefox.

A. Windows Azure

Windows Azure is an integrated operating system for cloud computing that facilitates the management of scalable Web applications over the Internet. Because the platform offers a wide range of services, all of these things and more are possible. The most basic thing a cloud platform does is run applications. Azure provides four options for doing this: Virtual Machines, Cloud Services, Web Sites, and Mobile Services. Azure Cloud Services technology is designed on purpose to support reliable, scalable and low-admin applications, and it is an example of PaaS. The hosting and management environment is maintained at Microsoft data centers. [22]

Windows Azure servers are located in across seventeen regions namely: East US, East US2, US Gov lowa, US Gov Virginia, North Central US, South Central US, West US, East Asia, Southeast Asia, Brazil South, North Europe, West Europe, Japan East, Japan West, Australia East and Australia Southeast.

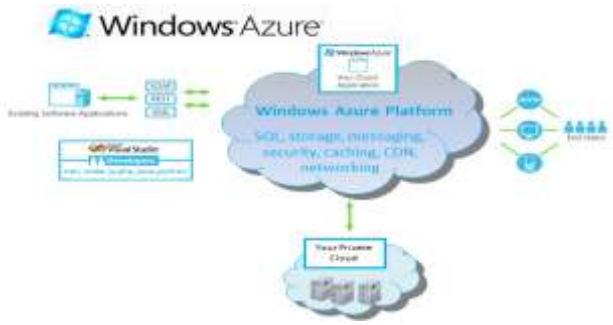

Fig VII: Windows Azure Architecture

For the purpose of this research, the server region used was North Europe and the researcher's location is Dubai, Middle East. Thirty readings were taken and the average of each case was recorded.

RESPONSE TIME FOR 1-COMPUTER (WIFI) WITH 8Mbps Bandwidth (Time in sec)				
20.55	26.03	23.17	15.79	25.36
19.02	14.05	14.05	10.67	10.73
14.09	16.42	09.09	15.97	18.57
11.45	11.26	13.76	19.61	10.98
14.56	26.01	15.57	11.41	13.62
21.17	16.13	10.81	13.58	8.02
RESPONSE TIME FOR 5-COMPUTERS SIMULTANEOSLY(WIFI) WITH 8Mbps Bandwidth (Time in sec)				
15.63	19.92	24.34	20.68	24.74
22.46	24.62	16.31	18.43	19.14
17.05	16.23	19.06	23.03	21.42
23.18	22.32	18.22	19.47	18.35
19.04	20	19.47	21.03	11.57
20.33	20.78	15.17	17.76	17.85

Reading (Response Time I)

RESPONSE TIME 1-COMPUTER(WIFI) WITH 156kbps Bandwidth (Time in sec)				
1:38	1:30	1:34	1:28	1:13
1:19	1:51	1:29	1:34	1:26
1:54	1:47	1:48	1:35	1:21
1:28	1:23	1:48	1:13	1:25
1:46	1:44	1:30	1:35	1:30

1:39	1:59	1:44	1:48	1:57
RESPONSE TIME FOR 5-COMPUTERS SIMULTANEOSLY(WIFI) WITH 156kbps Bandwidth (Time in sec)				
2:11	2:15	1:49	1:54	2:12
1:55	1:51	2:07	2:12	2:16
2:06	2:08	1:54	2:16	2:01
2:05	1:53	2:05	1:57	2:23
2:20	2:18	2:09	2:07	1:41
RESPONSE TIME 1-COMPUTER(LAN) WITH 156kbps Bandwidth (sec)				
57.55	1:03	54.78	57.94	57.92
58.52	59.46	1:08	46.44	1:05
1:07	1:11	55.06	1:09	56.28
53.88	58.13	54.61	56.32	55.38
1:10	57.95	56.36	1:11	1:03
59.49	1:12	1:08	54.45	58.97
RESPONSE TIME FOR 5-COMPUTERS SIMULTANEOSLY (LAN) WITH 156kbps Bandwidth (min:sec)				
1:33	1:27	1:40	2:00	1:50
1:26	1:54	1:26	1:34	1:34
1:47	1:22	1:21	1:26	1:25
1:23	1:40	1:38	1:49	1:23
1:43	1:25	1:20	1:33	1:37
1:51	1:38	1:45	1:24	1:39

Reading (Response Time II)

The **average response time** using **one computer** with a bandwidth of **8Mbps (wifi)** was **15.72sec** while the **minimum** and **maximum** were **8.02sec** and **26.03sec**. The time it takes to **download** the video file (200MB) was **4mins 48sec** and using **five computers** simultaneously with the same bandwidth, the average response time was **19.59sec** and the **minimum** and **maximum** were **11.57** and **24.74** respectively.

In the second analysis, the **average response time** using one computer with a bandwidth of **156kbps (wifi)** was **1min 36sec** while the **minimum** and **maximum** were **1min 13sec** and **1min 59sec** respectively. The time it takes to download the video file (200MB) was **10min 8sec** and using **five computers** simultaneously with the same bandwidth, the average response time was **1min 44sec** and the **minimum** and **maximum** were **1min 41sec** and **2min 23sec**.

Lastly, the average response time using **one computer** with a bandwidth of **156kbps (LAN)** was **1min 01sec** while the **minimum** and **maximum** were **53.88sec** and **1min 12sec** respectively. Using **five computers** simultaneously with the same bandwidth, the average response time was **1min 35sec** while the **minimum** and **maximum** were **1min 20sec** and **1min 54sec**.

VI. SUMMARY

In this research, the concept of Queuing theory is being used for understanding the input process and using M/M/C: model of Kendell's notation the mathematical process of queuing is done for single and multiple servers. Microsoft windows azure is also used to upload and publish a video file which was accessed for varying bandwidth and the response time recorded.

CONCLUSION

The main aim of the cloud service providers is to ensure maximum usage of the resources with minimal waiting time [4]. Scheduling criteria should be in such a way that the waiting time can be minimized and depending upon the number of servers and bandwidth capacity. This research analyses QoS in cloud and proposes ways of improvement.

RECOMMENDATIONS

The recommendations hold pertinence to all actors of the Cloud ecosystem but are of most relevance to Cloud infrastructure providers, Internet Service Providers and application developers wishing to minimize the impact that poor QoS provisioning can bring.

1. **Multiple servers:** The key benefit of having numerous servers in cloud computing is, the system performance increases efficiently by reducing the mean queue length and waiting time than compared to the conventional approach of having only single server so that the consumers need not wait for a long period of time and also queue length need not be bulky.

2. **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, eventually our performance will be affected. Therefore, optimum bandwidth plays a vital role in the improvement of response time hence reducing queues.
3. **Best Protocol Selection:** Various internet applications 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.

4. **Best Medium Selection:** In this, 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. The wired media results in higher transfer rate and reliability is high rather the wireless media. In wireless media the various electric radiations and weather not only affects the signal's strength but also open it to security risks. Wireless medium be can easily hacked through backtracking and other software's available in the market

REFERENCES

[1]. H. P. Kaiqi Xiong, "Service Performance and Analysis in Cloud Computing," 12 October 2014. [Online]. Available: http://www4.ncsu.edu/~hp/Kaiqi10.pdf.

[2]. T. Y. S. Z. C. J. Lizheng Guo, "Dynamic Performance Optimization for Cloud Computing Using M/M/m Queuing System," Hindawi Publishing Corporation Journal of Applied Mathematics, p. 8, 2014.

[3]. B. Gabriel, "NASA Turns to Online Giant Amazon for Cloud Computing Services for Mars Rover Curiosity," 12 August 2012. [Online]. Available: http://biginscience.com/big-in-science-articles/2012/8/12/nasa-turns-to-online-giant-amazon-for-cloud-computing-servic.html.

[4]. BusinessVibes, "Global Cloud Computing Service Market Expects to Reach $555 Billion Worth by 2020," BusinessVibes, 2014 .

[5]. H. P. Kaiqi Xiong, "Service Performance and Analysis in Cloud Computing," IEEE Xplore Digital Library, pp. 693 - 700, 2009.

[6]. E. C. I. L. Bhaskar Prasad Rimal, "A Taxonomy and Survey of Cloud Computing Systems," Fifth International Joint Conferences on INC, IMC and IDC, 2009.

[7]. F. T. Y.-S. D. S. G. Bo Yang, "Performance Evaluation of Cloud Service Considering Fault Recovery," in Cloud Computing, First International Conference, CloudCom , Beijing, China, 2009.

[8]. K. S. N. Ani Brown Mary, "Performance Factors of Cloud Computing Data Centers using [(M/G/1):(∞/GDModel)] Queuing Systems," International Journal of Grid Computing and Applications, pp. Vol. 4, No. 1, 2013.

[9]. D. D. Yao, "Refining the Diffusion Approximation for the M/G/m Queue," Operations Research, vol. vol. 33, pp. pp. 1266 - 1277, 1985.

[10]. A. N. J. Martin, "On Service Level Agreements for IP Networks," IEEE INFOCOM, vol. vol. 2, 2002.

[11]. M. C. a. O. Boxma, "BMAP modelling of a correlated queue," in Network Performance Modeling and Simulation, J. Walrand, K. Bagchi and G.W. Zobrist (eds.), 1998, pp. 177-196.

[12]. L. R. Y. Simmhan, "Comparison of resource platform selection approaches for scientific workflows," 19th ACM Intl. Symposium on High Performance Distributed Computing, HPDC, pp. 445-450, 2010.

[13]. S.-c. L. Hao-peng Chen, "A Queuing-based Model for Performance Management on Cloud," IEEE Xplore, 2010.

[14]. T. G. Peter Mell, "The NIST Definition of Cloud," 5 11 2014. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[15]. EMC, "Cloud Computing Services," 25 12 2014. [Online]. Available: http://www.emc.com/corporate/glossary/cloud-computing-services.htmEMC.

[16]. T. G. Peter Mell, "National Institute of Standards and Technology," 09 2011. [Online]. Available: http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[17]. K. S. Mather T, Cloud Security and privacy: An nterprise Perspective on Risks and Compliance, O'Reilly Media, 2009.

[18]. G. Reese, Cloud Application Architectures: Building Applications and Infrastructure in the Cloud, O'Reilly Media, April, 2009.

[19]. J. F. R. John W. Rittinghouse, Cloud Computing Implementation, Management, and Security, Broken Sound Parkway NW: Taylor and Francis Group, LLC, 2010 .

[20]. R. B. M. K. F. H. Hurwitz, Cloud Computing For Dummies, For Dummies, 2010.

[21]. P. D. Juran, "Introduction to Queueing Theory," [Online]. Available: http://pages.stern.nyu.edu/~djuran/queues.doc.. [Accessed 29 12 2014].

[22]. M. Rouse, "TechTarget," 02 May 2009. [Online]. Available: http://searchcloudcomputing.techtarget.com/definition/Windows-Azure.

[23]. E. I. E. Z. R. Mohamed Eisa, "Enhancing Cloud Computing Scheduling based on Queuing Models," International Journal of Computer Applications (0975 – 8887), p. Volume 85 – No 2, 2014.

[24]. L.-T. C. W.-L. C. K.-C. W. Chao-Tung Yang, "Implementation of a Medical Image File Accessing System on Cloud Computing," 13th IEEE International Conference on Computational Science and Engineering, 2010.

[25]. S. H. S. S. a. R. S. Sinung Suakanto, "Performance Measurement of Cloud Computing Services," International Journal on Cloud Computing: Services and Architecture, pp. Vol. 2, No. 2, 2012.

[26]. J. M. V. B. M. Hamzeh Khzaei, "A Fine-Grained Performance Model of Cloud Computing Centers," IEEE Transaction on Parallel and Distributed Systems, pp. Vol X, No Y, 2012.

[27]. Z. A. Soumya Ranjan Jena, "Response Time Minimization of Different Load Balancing Algorithms in Cloud Computing Environment," International Journal of Computer Applications, pp. Vol 69, No. 17, 2013.

[28]. M. N. A. K. Ashraf Zia, "A Scheme to Reduce Response Time in Cloud Computing Environment," International Journal on Modern Education and Computer Science, pp. No 6, 56-61, 2013.

[29]. M. M. K. M. Nataraja Suresh, "Improved load balancing model based on partitioning in Cloud Computing," International Journal of Computer Science and Mobile Computing, pp. Vol. 3, Issue 8, pg 411-416, 2014.