ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **9** (2020), no.10, 8217–8225
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.9.10.51 Spec. Iss. on AOAOCEP-2020

CLOUD COMPUTING AS A QUEUE MODEL WITH SERVER BREAKDOWN

G. ARUL FREEDA VINODHINI

ABSTRACT. Data Storage and sharing has become very easy because of the concept of cloud computing. There exist many service providers and hence the efficiency of their service should be assessed to compete in the field. The efficiency can be achieved by reducing the waiting time of the customer. To calculate the waiting time of a customer queue models are used. Here we consider cloud computing as a single server queue model which undergoes a breakdown. Immediately all customers are cleared from the system and repair starts. During repair period customers who arrive to the system initiate a timer and abandon the system once the timer is completed. This cloud can be developed as a queue model and the performance measures like waiting time of customers, abandon rate etc., are evaluated. We know that the transient solution gives more appropriate results than a steady state solution.

1. INTRODUCTION

Cloud Computing, the future is going to change the way we function. Storing and sharing are the two main feature of cloud computing. Initially we used floppy disks to save data, later came the CD's, flash drives etc....File sharing and saving has taken a mass development over the period of time. The size of the storage device also become smaller and smaller but with higher capacity. We were running out memory by storing all apps, documents in our device. Now the life style has completely changed in such a way that all are saved in a remote place and any one can access it from anywhere. Simultaneous editing

2020 *Mathematics Subject Classification.* 60J22, 68M10, 60K30.
Key words and phrases. Cloud, Queuing Theory, QoS, Catastrophes, Transient Solution.

of documents, power point presentations or excel sheets are also made possible with cloud computing. Especially during the pandemic days we are able to work from anywhere is the main advantage of it. Cloud computing provides variety of services depending upon the needs of the society. The costs are charged according to the usage. As an individual user the service providers offer 5GB storage for free. Whenever we require more storage it is offered at an extra cost. The payment can be made monthly or yearly depending on your choice.

Businesses large scale or small scale need to spend more money to maintain their data. They cannot afford for backup support or a massive storage hub. For them cloud computing is the best solution. The risk of system breakdown will be taken care of by the service providers. Cost saving is the main reason why many bigger organizations choose cloud. Moreover it's enough if we pay for what we use. The other benefits of cloud computing includes low cost, less infrastructure, efficient performance, recovery made easier, increased data safety. There are four types of cloud models to suit the business requirement. Private cloud is one designed for the use of that particular organization. This is highly used for intra business purpose. Community cloud provides service to community and organizations. Public cloud is used for business to consumer interactions. Hybrid cloud covers both business - business and business – consumer interactions. There are three types of cloud computing services: SaaS, PaaS and IaaS. Some organizations find difficult to invest in the required software. For them SaaS will be an advantageous one. PaaS provides service to those who need platforms to construct new web page, software, etc. Besides PaaS, the next fundamental cloud service model is IaaS. It delivers computing set-up like network connections, load balancers, virtual server space, IP addresses and bandwidth. The hardware resources from many multiple servers are mined and distributed across networks and numerous data centers. This gives redundancy and consistency to IaaS. The other smaller versions are grid computing and utility computing. Utility computing is well suited for small industries where as grid computing provides only limited service in comparison with cloud.

The demerits of cloud computing includes technical issues that crop up now and then. Even good service providers struggle to avoid these technical issues. When one works in a server shared with others, any sudden attack on their system will also affect all the members sharing the system. Security threat is the biggest demerit in cloud. All important documents of the organization when

stored in a common server user by many there is a chance that hackers might take this information. Cloud providers may face downtime due to power failures, service maintenance and low internet availability. This downtime should also be considered while calculating the waiting time of the customer. From the user's side also the network connectivity should be good. In case of natural calamities you cannot even access the data in the cloud. Based on the bandwidth the service providers give various schemes of payment options. This brings in the difference in the speed and the amount of data you access. Though it has many demerits they seem to be meager with the benefits we get from the cloud. So no doubt that cloud computing will reach its heights in the future.

2. Cloud Computing and Queuing Theory

Quality of Service (QoS) is most important for the success of the business. Especially when many service providers are available for cloud there is a perfect competition among them in improving their quality of service. In order to improve QoS we must reduce the waiting time of the customer. Depending on the arrival rate the servers must be increased to provide service without any delay. Vilaplana [1] calculated the response time by modeling the cloud computing as an open network queuing model. Based on the parameters like arrival rate, service rate and the concept of bottle neck of the system they claim that their model can guarantee the QoS. Also they claim that their model is useful in creating a real cloud. Resource allocation plays another important role in the cloud because it is necessary to utilize the virtual machines available in the data centre in a best possible way. Since many of them access the data centre in a particular time, scheduling should be perfect to assure QoS. Vetha [2], compared cloud with E-M/M/1/K queue model and proposed scheduling procedure. Using properties of queuing model an algorithm for resource allocation is proposed. Comparing with other models they concluded as the presented queue model generated high throughput and minimized the waiting time. Many individuals and enterprises are attracted towards the cloud services due to its cost cutting measures, storage benefits etc... On the other hand the service providers compete in to reduce the costs but at the same time to increase the efficiency. Shaguna [3], explained about the merits and demerits of cloud in their work. A complete analysis of various cloud queuing models is done and results

are tabulated. They also added the missed out concepts in the papers and suggested the scope of improvement. In [4], the comparisons between M/M/1 and M/M/c queue models for cloud are discussed. The parameters are controlled to improve the performance of the system. For providing resources in cloud a scheme called Energy efficient for single server model was proposed by Yuxiang Shi et al [5]. The resource utilization is measured using the method of linear predicting. They could achieve comparatively a lesser response time and better energy consumption. We can extend this paper for multi servers which would actually suit the real could models. Santhosh Kumar [6] has presented an efficient allocation and scheduling technique in which a job is divided into tasks and allotted to various virtual machines simultaneously. This kind of parallel completion of tasks reduces the waiting time of jobs, thus improving QoS.

Multi server cloud model with many priority class of customers are discussed in [7]. The blocking probabilities for various classes of customers are calculated. Customer rejection probabilities for different classes are also calculated. Finite capacity cloud queuing models where the service rates of virtual machines are considered to be heterogeneous are designed in [8]. The cost is minimized using resource allocation and also the service response time becomes reduced in M/G/s model with infinite buffer capacity. The service time in the cloud is considered to be general and cloud network is taken as open. They observed that with increase in number of servers the waiting time decreases. Task response time and blocking probabilities are also evaluated. A cloud in which arrivals follow Poisson distribution and service follow Gamma distribution is presented in [9]. Here the authors considered variable arrival rates and the corresponding coefficient of variation in the numerical example taken. They used Artifex the discrete event simulation engine which works on the principal of object oriented petrinets to find the performance measures. The response time is very high and they didn't discuss as how to reduce it. Transient analysis of the cloud is done by P. Suresh Varma et al [10]. They concluded that it has tremendous changes in the performance measures.

Many authors have tried to fit in the available queue models with the cloud design. The high end concepts of Queue models like balking, impatience, breakdown, server vacation, priority etc... are not considered. In real time cloud all these factors play an important role in waiting time of the customer. Cloud performance measure will be realistic if the probabilities are calculated in transient

mode. Here our objective is to add the concept of breakdown and repair to a single server queue model and study the performance measures of the cloud. The number of customers getting lost due to breakdown or server maintenance is to be calculated. This would guarantee the QoS of the cloud.

3. The Mathematical Model

Consider a cloud operating with single server which will schedule and allot the user requests to the virtual machines in the system. As the cloud takes the responsibility of sharing and storing the data the maintenance becomes unavoidable. For this the system should be shut down for a shorter period of time. The working principle of cloud wholly depends on the internet. So whenever there is network failure the cloud system also under go random failures. During such failures all users are disconnected including the existing requests. Then the cloud undergoes a repair process which is also random. In the meantime the new requests that are trying to access the cloud become impatient after a time period of T and leave the system once for all. Thus the queues with catastrophes and impatient customers when system is down will fit aptly to the real cloud model. This model allows us to calculate QoS, rate of customers served, proportion of customers rejected due to disasters and finally abandonment rate that arises because of impatient customers when system is down. Assumptions and parameters of the model are as follows. λ represents rate with which the user requests to the cloud according to Poisson, μ denotes rate at which the requests is satisfied and follows exponential distribution, η denotes rate at which the server in the operating state undergoes breakdown, γ denotes rate at which the repair starts according to exponential distribution, ξ denotesrate at which the customers abandons the system once for all after the completion of the timer, T, J is state of the server (0 – idle server; 1 – Operating server), and ln denotes the number of customers arriving when n^{th} customer is getting service.

3.1. **Balance Equations in Transient State.**

Let $P_{jn}(t)$ denote the system state probabilities in transient state. From the transition diagram the Kolmogorov differential difference equations are given by Yechiali are as follows.

$$\begin{aligned}
P'_{00}(t) &= -(\lambda+\gamma)P_{00}(t)+\xi P_{01}(t)+\eta\sum_{n=o}^{\infty}P_{1n}(t)\\
P'_{0n}(t) &= -(\lambda+\gamma+\eta\xi)P_{0n}(t)+\lambda P_{0,n-1}(t)+(n+1)\xi P_{0,n+1}(t), n\geq 1\\
P'_{10}(t) &= -(\lambda+\eta)P_{10}(t)+\gamma P_{00}(t)+\mu P_{11}(t)\\
P'_{1n}(t) &= -(\lambda+n\mu+\eta)P_{1n}(t)+\lambda P_{0,n-1}(t)+\mu P_{1,n+1}(t)+\gamma P_{0n}(t)
\end{aligned}$$

Also, assume that the system is working and there are no customers in the system at time t=0. (i.e.) $P_{10}(0)$=1 and $P_{1n}(0)$=0, n≥1. Clearly, for any t≥0, $\sum_{n=o}^{\infty}P_{0n}(t)+\sum_{n=o}^{\infty}P_{1n}(t)$=1. Under steady state conditions above equations become as follows. For J=0,

$$\begin{aligned}
(\lambda+\gamma)P_{00} &= \xi P_{01}+\eta\sum_{n=0}^{\infty}P_{1n}=\xi P_{01}+\eta P_{1*} \qquad (3.1)\\
(\lambda+\gamma+n\xi)P_{0n} &= \lambda P_{0,n-1}+(n+1)\xi P_{0,n+1}, n\geq 1
\end{aligned}$$

For J=1,

$$\begin{aligned}
(\lambda+\eta)P_{10} &= \gamma P_{00}+\mu P_{11}\\
(\lambda+n\mu+\eta)P_{1n} &= \lambda P_{1,n-1}+\mu P_{1,n+1}+\gamma P_{0n}, n\geq 1
\end{aligned}$$

Eq. 3.1 gives $\gamma P_{0*}=\eta P_{1*}$ and also $P_{0*}+P_{1*}=1$. Where, $P_{0*}=\frac{\eta}{\eta+\gamma}$, $P_{1*}=\frac{\gamma}{\eta+\gamma}$, and $P_{00}=\frac{\eta P_{1*}K}{\xi}$, in which, $K=\int(1-s)^{\frac{\gamma}{\xi}-1}e^{\frac{\lambda s}{\xi}}ds$, $P_{01}=\frac{\eta}{\xi}[\frac{\lambda+\gamma}{\xi}k-1]P_{0*}$, and $E[L_0]=\frac{\lambda}{\gamma+\xi}P_{0*}$.

The following equations are used to find P_{10}.

$$Z_0=\frac{(\lambda+\mu+\gamma)-\sqrt{(\lambda+\mu+\gamma)^2-4\lambda\mu}}{2\lambda}$$

$$G_0(Z)=P_{00}e^{\frac{\lambda z}{\xi}[1-\frac{\int(1-s)^{\frac{\gamma}{\xi}-1}e^{-\frac{\lambda s}{\xi}}ds}{k}]}(1-Z)^{-\frac{\gamma}{\xi}}$$

$$P_{10}=\frac{\gamma Z_0}{\mu(1-Z_0)}G_0(Z_0)$$

$$\eta E[L_1]=\frac{\lambda}{\gamma+\xi}(\xi P_{1*}+\gamma)-\mu(P_{1*}-P_{10})$$

Probability of customers being served as follows.

$$\text{Probability}=\frac{\mu}{\lambda}(P_{1*}-P_{10})$$

Rate of abandoned customer is as follows

$$\text{Rate} = \xi E[L_0] = \lambda - (\mu + \eta)E[L_1]$$

Expected time of a customer who completes service is as follows, in which, $E[L] = E[L_0] + E[L_1]$.

$$\text{Expected time} = \frac{E[L]}{\lambda}$$

4. Numerical Analysis and Discussion

Consider a cloud with a single server which operates in two states. Let requests arrive at a rate of 3/Hour and get service at a rate of 5/hour. Let the system breakdown at a rate of 0.4/hour. The system is repaired at a rate of 0.8/hour. The incoming requests abandon the system at a rate of 0.2/hour. Sensitivity analysis is performed by increasing the repair rate from 0.4 to 0.6 and by doubling the service rate. Various performance measures are computed and all three models are compared. By using the values from Table 1 - Table 3, other

Table 1. CASE-I

λ	μ	η	γ	ξ
3	5	0.4	0.8	0.2

Table 2. CASE-II

λ	μ	η	γ	ξ
3	5	0.6	0.8	0.2

Table 3. CASE-III

λ	μ	η	γ	ξ
4	10	0.4	0.8	0.2

parameters are calculated and presented in Table 4.

Estimation of performance measures was carried out in M.S. Excel. This sheet can be used as a back end for C programming. The input to the problem is as per the data given as three cases. The output results are tabulated for various performance parameters. Whenever the breakdown rate increases the waiting time

TABLE 4. Various Parameters

Performance Measure	Case-I	Case-II	Case-III
P_{0*}	0.3333	0.4285	0.3333
P_{1*}	0.6666	0.5714	0.6666
K	0.1283	0.1283	0.1259
P_{00}	0.1711	0.2200	0.1679
P_{01}	0.12	0.7	0.2
No. of customers in state J=0	1	1.2857	1.3333
Z_0	0.8583	0.8106	0.9397
G_0	0.2	0.18	0.25
P_{10}	0.1939	0.1232	0.3120
No. of customers in state J=1	0.4365	0.5020	0.1871
P [No. of customers served]	0.7878	0.7469	0.8865
Rate of abundance in state J=0	0.2	0.2571	0.2666
Rate of abundance in state J=1	0.6424	0.1884	2.0538
Total number of customers in the system	1.4365	1.7877	1.5204
W_s	0.4788	0.5959	0.3801

of the customer also increases which is clear from the first two cases. Also when the traffic intensity decreases the waiting time of the requests also decreases and it is clear from performance measures of cases 1 and 3.

5. CONCLUSION

Performance calculation of small cloud center is presented with the theory and equations governing the model based on queuing concept. Single server model along with server breakdown, repair rate and rate of abundance of the incoming request is taken along with their formulations for performance parameters. M.S. Excel is used to calculate the performance of cloud data server. Comparisons among different values of the parameters are tried and relevant interpretations are listed. Thus all essential performance measures of cloud model were found using the suggested queue model. The model can be extended to transient solution which will give more appropriate values for the performance measures.

REFERENCES

[1] J. VILAPLANA, F. SOLSONA, I. TEIXIDÓ: *A Queuing Theory Model for Cloud Computing*, J Super Comput., **69** (2014), 492–507.

[2] S. VETHA, K. VIMALA DEVI: *Dynamic Resource Allocation in Cloud using Queuing Model*, J Indus Poll Cont., **33**(2) (2017), 1547–1554.

[3] G. SHAGUNA, A. SAKSHI: *Queuing Systems in Cloud Services Management: A Survey*, Inter J Pure App Mathe., **119**(12) (2018), 12741–12753.

[4] K. RUTH EVANGELIN, V. VIDHYA: *Performance Measures of Queuing Models Using Cloud Computing*, Asian J Engi App Tech., **4**(1) (2015), 15–24.

[5] Y. SHI, X. JIANG, K. YE: *An Energy-Efficient Scheme for Cloud Resource Provisioning Based on Cloud Sim*, In: Proc. of IEEE International Conference on Cluster Computing, Austin, TX, (2011), 595–599.

[6] B. SANTHOSH KUMAR: *Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud*, Inter J Comp Sci Mobile Comp., **4**(11) (2015), 108–113.

[7] W. ELLENS, M. IVKOVIC, J. AKKERBOOM, R. LITJENS, H. VAN DEN BERG: *Performance of Cloud Computing Centers with Multiple Priority Classes*, In: Proc. of IEEE Fifth International Conference on Cloud Computing, Honolulu, HI, (2012), 245–252.

[8] R. MURUGSAN, C. ELANGO, S. KANNAN: *Resource Allocation in Cloud Computing with M/G/s Queueing System*, Inter J Adv Resear Comp Sci Soft Engi., **3**(10) (2014), 118–124.

[9] H. KHAZAEI, J. MISIC, V.B. MISIC: *Modelling of Cloud Computing Centers Using M/G/m Queues*, In: Proc. of International Conference on Distributed Computing Systems Workshops, Minneapolis, MN, (2011), 87–92.

[10] P. SURESH VARMA, A. SATYANARAYANA, M.V. RAMA SUNDARI: *Performance analysis of cloud computing using Queuing Models*, In: Proc. of nternational Conference on Cloud Computing Technologies, Applications and Management (ICCCTAM), Dubai, (2012), 12–15.

DEPARTMENT OF SCIENCE AND HUMANITIES
SAVEETHA SCHOOL OF ENGINEERING
SIMATS, CHENNAI, TAMIL NADU, INDIA
Email address: arulfreedav@gmail.com