The Effect of Queuing System Capacity on the Blocking Probability Value Using M/G/1/C and G/G/m/K Model in Cloud Data Center

Yuniarmelinda Ikha Meru B.
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Sholeh Hadi Pramono
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Erni Yudaningtyas
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Abstract: Information technology is growing fast. The growth of information technology is caused by the increasing of human need for technology. The example of it is cloud computing. The more user who accesses the cloud service can cause the possibility of users being rejected or blocking. The more tasks that are not served so that the performance of cloud data centers becomes less effective. One of method to minimize the blocking probability by controlling the queue system capacity [3]. The system model used in this research is a queuing theory model, load balancing and several physical machines. The queuing model used to analyze is M/G/1/C on load balancing and G/G/m/K on physical machine. General distribution queuing model compatible with the dynamic characteristics of cloud computing. The purpose of this research is to see the effect of queuing system capacity on blocking probability and see the effect of load balancing in this system model. The results of the simulation are the capacity of the queue system which is getting bigger, reducing the possibility of blocking. From this simulation the effective blocking probability value with a queuing system capacity of 3000. The existence of load balancing in this model makes physical machine performance effective. Because load balancing divides the task load equally into each physical machine.

Keywords: Data center, Cloud computing, Blocking probability; Response system; Queuing theory

1. INTRODUCTION

Information technology is growing fast. The growth of information technology is caused by the increasing of human need for technology. The example of it is cloud computing. In recent years cloud computing has attracted attention in the industrial world. Definition of Cloud Computing is a computing model that allows ubiquitous (wherever and whenever), convenient, on-demand network access to computing resources that can be quickly released or added. The advantage of cloud computing is flexible and efficient access [1]. In general, cloud computing is divided into three types namely Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Platform as a Service (PaaS) [1].

Cloud computing technology consists of several components, one of which is the cloud data center. Cloud data centers are like a home for all data that is connected through a server. The Cloud Data Center receives every day from users who want to access services randomly and in large numbers. this makes users have to wait to get service from CDC. To analyze it, you can use queuing theory. Queue theory can be used to examine the activities of service facilities in a series of random conditions from a queue that happen [2].

The growth of cloud computing has resulted in a high increase in users to access cloud services. This can cause blocking probability. Blocking Probability is the possibility of blocking / rejection by the system of user requests for services to the Cloud Data Center. The higher the blocking probability, the more tasks that are not served so that the performance of the cloud data center becomes less effective. To minimize the occurrence of blocking probability, by controlling the capacity of the queuing system [3].

This research use queuing theory with the M/G/1/C queuing model on the load balancing unit and G/G/m/K on the physical machine. The function of load balancing in this research is to balance the workload on each physical machine. So the optimal results can be seen from the parameters of the average response time and the average queue length. Service rate on load balancing and physical machines uses general distribution, because the dynamic characteristics of cloud data centers produce high arrivals. Therefore, this research will increase the capacity of the queuing system to minimize the blocking probability value. This research also uses load balancing to support performance parameters of cloud data center, so it can obtain an effective value used queuing theory. The results of the implementation of this system were simulated using Java Modeling Tools V.1.0.4.

2. RELATED WORK

Although cloud computing has attracted research attention, but only a few of research that discusses this. Research [3], use queuing theory in cloud data centers. In this research, the queuing model on arrival and service time used general distribution with G/G/c model. In this research [2] used a general distribution because the distribution is more relevant to the characteristics of cloud data center. From the results of this study, the value of blocking probability is decreases equivalent with the increasing of system capacity.

Other research, queuing theory is used to analyze the performance of cloud data centers by Said, et al. On this research [5] used load balancing which is modeled by M/M/1/C and M/M/m/K (K>m) queuing models on physical machines. This study analyzes the performance of cloud data center performance with load balancing and makes a numerical model to find out the number of virtual. Therefore, in this research used a queuing model with general distribution service time and adding load balancing to the system model.

3. METHOD

3.1 Queuing Theory

Queuing is a situation that we see in our daily lives. Such as waiting in line in shopping malls or in buildings, vehicles waiting for traffic light, customers waiting in cashier at supermarkets and so on. According to Taha (2007), queuing theory is a theory that discusses mathematical learning from queues or waiting lines.

In general, customers come into a system with a random time, can't be arranged and can't be served immediately so they have to wait. Therefore, queuing theory is used to optimize services without having to wait too long. In addition, queuing theory can also be used to examine the activities of service facilities in a series of random conditions of a queuing system that happen [3].

The queuing factors are the distribution of arrivals, service distribution, service facilities, service discipline and queue length. The user arrival is usually calculated through time between arrivals. It is time between arrival of two consecutive customers in a service. Then, the service is determined by the service time. It is the time needed to serve users in a service. The next component is the system capacity. System capacity is the maximum number of customers, including those being served and those in the queue, which can be accommodated by service facilities at the same time. The last component is queuing discipline.

3.2 M/G/1/C

The queuing system model in this research used M/G/1/C queuing model on the load balancing unit. User arrival process used Markovian, arrival time used exponential distribution, service time used gamma distribution with mean 1/μ . The explanation of the method is:

M/G/1/C

with:

M : arrival rate used exponential distribution
G : service rate used gamma distribution
1 : number of unit is 1
C : queue system capacity is C

The average value of response time (R) and the average value of the number of tasks in the system (q) can be described as follows [4]:

$$R = \frac{\rho\ (1 + \mu^2\sigma_s^2)}{2\mu\ (1-\rho)}$$

$$\bar{n} = \bar{q} + \rho$$

$$\bar{q} = \rho^2\ \frac{(1 + \mu^2\sigma_s^2)}{2(1-\rho)}$$

with:

R	: Average response time
$\bar{n}$	: Average number task in systems
$\bar{q}$	: queue length
ρ	: average arrival rate divided average service rate
μ	: Service rate
σ_s	: Coefficient of Variation

3.3 G/G/m/K

The queuing system model on the physical machine used the G/G/m/K queuing model. User arrival process used general distribution and arrival time used gamma distribution, service time used gamma distribution with mean 1/μ. The explanation of the method is:

G/G/m/K

with:

G : arrival rate used general distribution
G : service rate used general distribution
m : number of unit is 1
K : queue system capacity is K

This physical machine considers the average queue length, the average number of tasks in the system, the throughput and the average system response time. To calculate the average number of tasks in the system (L) can be described as follow [3]

$$\mathrm{L} = \sum_{n=1}^{N} n\ x\ P_n$$

$$\theta = \sum_{n=1}^{N} U_n\ x\ P_n$$

$$\mathrm{R} = \frac{\mathrm{L}}{\theta}$$

with:

L	: Average number task in the system
N	: System Capacity
n	: Number task in the system
P_n	: Probability of n-task in the system
θ	: Throughput
U_n	: Total number of servers
R	: Average response time

3.4 Blocking Probability

Blocking Probability is the possibility of blocking / rejection by the system of user requests for services to the cloud data centre. The blocking probability value can be determined by queuing theory. The equation to find the value of blocking probability as follows [7]:

$$P_C = \frac{\rho\left(\sqrt{\rho}s^2 - \sqrt{\rho} + 2C\right)/\,(2 + \sqrt{\rho}s^2 - \sqrt{\rho})\,(\rho - 1)}{(\rho^2\,((1 + \sqrt{\rho}s^2 - \sqrt{\rho} + C)/(2 + \sqrt{\rho}s^2 - \sqrt{\rho})) - 1}$$

$$\rho = \frac{\lambda}{\mu}$$

with:

P_C	: Probability C- task in the system
λ	: Arrival rate
μ	: Service rate
s	: Coefficient of Variation
C	: System Capacity

3.5 Proposed Model

This research used a system model by IaaS (Infrastructure as a service). IaaS service providers provide IT infrastructure such as server, memory storage, virtual machines and operating system. Data center is a place that provides cloud computing services related to data and information communication [6]. Cloud data center also can be used as a group of servers used by IaaS service providers to meet user needs. Figure 1 below is a picture of a system model in the cloud data center.

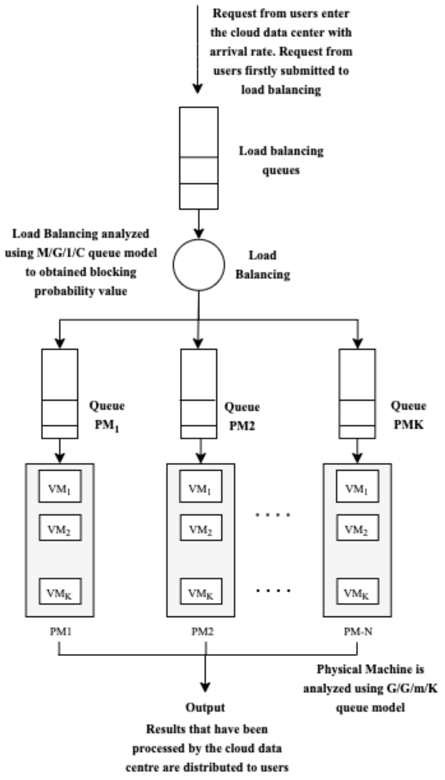
Request from users enter the cloud data center with arrival rate. Request from users firstly submitted to load balancing
Load balancing queues
Load Balancing analyzed using M/G/1/C queue model to obtained blocking probability value
Load Balancing
Queue PM$_1$
Queue PM2
Queue PMK
VM$_1$
VM$_2$
VM$_K$
. . . .
PM1
PM2
PM-N
Physical Machine is analyzed using G/G/m/K queue model
Output
Results that have been processed by the cloud data centre are distributed to users

Figure 1. System Model Cloud Data Center

Cloud data center system model in figure 1 was analyzed using queuing theory, which was tested using Java Modeling Tools simulator. Requests from users sent to the cloud data center, for example requests come to the cloud data center to access websites hosted on a physical machine with arrival rate λ. Requests from the user will enter Load Balancing before being directed to the Physical Machine. The function of load balancing is to balance the load on each physical machine. The N symbol in Figure 1 is the number of Physical Machines in the cloud data center and K is the number of Virtual Machines in each Physical Machine. To analyze load balancing used M/G/1/C queuing model. The symbol C in the queue model is the total number of tasks that can be accommodated in the load balancing queue. To analyze number of Physical Machines used G/G/m/K queuing model, where K is the total system capacity on each physical machine.

4. RESULT AND DISCUSSION

4.1 Blocking Probability Result

User requests that came to access cloud services are received and serve by load balancing with service time of 0.001. Queue system capacity value is changed from 400, 800, 1000, 2000, 3000, 3500. With the changing capacity of the queuing system, a blocking probability value is obtained. the blocking probability value is found in load balancing. The results are as in figure 2.

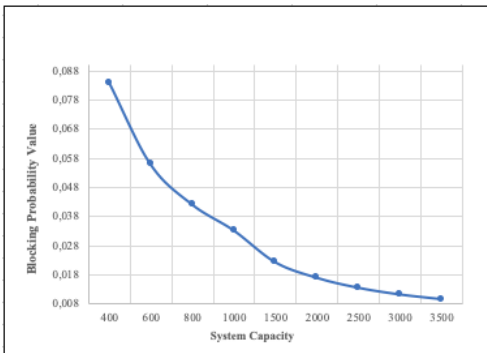

Figure 2. System Capacity vs Blocking Probability

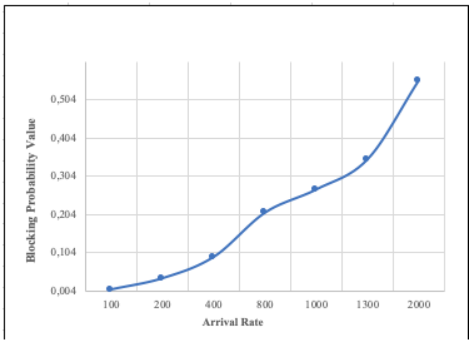

Figure 3. Arrival Rate vs Blocking Probability

From these results it can be analyzed that the blocking probability value decreases with the change in the queuing system capacity. The optimal blocking probability value when the system capacity is 3500. However, the higher arrival rate so the blocking probability value also higher.

4.2 Effect Load Balancing in Physical Machine

The service rate on load balancing used as arrival rate on physical machine. However, to be arrival rate, the amount of service rate is divided by the number of physical machines. Service rate on physical machines is 0.015 with a coefficient of variance 1.4. The number of virtual machines is changed from 22-30 units. The following results are obtained in Figure 4.

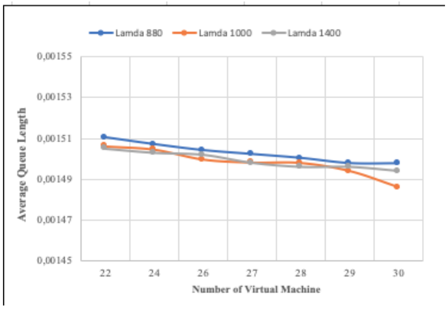

Figure 4. Average Queue Length in Physical Machine

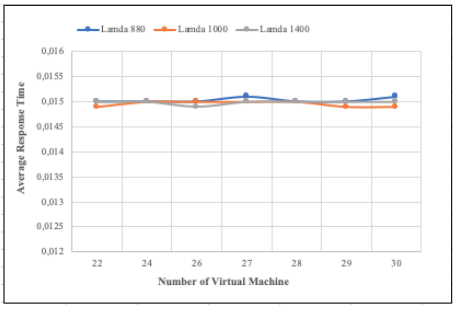

Figure 5. Average Response Time in Physical Machine

Using load balancing in the cloud data center system model, made physical machine work optimally. It can be seen as average queue length and average response time on each physical machine is constant. It is because load balancing divided the same load to each physical machine. The average queue length and response time decreases if the number of virtual machines being used is increasing.

5. CONCLUSION

This research can analyze the effect of system capacity on blocking probability values and load balancing on cloud data center performance. The method used is the M/G/1/C and G/G/m/K queuing model. Service time distribution on load balancing and physical machines used general distribution. The results of this research are to minimize the occurrence of blocking probability by increasing the capacity of the queuing system and with the load balancing unit, the average value of response time and the average length of the queue on the physical machine are more constant, because load balancing can balance the tasks on a physical machine.

6. REFERENCES

[1] Mell, Peter & Grance, Timothy. 2011. The NIST Definition of Cloud Computing. National Institute of Standards and Technology

[2] Kakiay, Thomas J. 2004. Dasar Teori Antrian untuk Kehidupan Nyata. Yogyakarta: Penerbit Andi

[3] Atmaca, T., Begin, T., Brandwajn, A., & Castel Taleb, H. 2016. Performance Evalution Centers with General Arrival and Service. IEEE

[4] Murdoch, J. 1978. Queuing Theory Worked Examples and Problems. The Macmillan Press. Ltd.

[5] El Kafhali, Said & Salah, Khaled. 2017. Stochastic Modelling and Analysis of Cloud Computing Data Center. IEEE

[6] Geng, Hwaiyu. 2015. Data Center Handbook. John Wiley & Sons, Inc., Hoboken, New Jersey

[7] MacGregor Smith, J. 2004. Optimal Design and Performance Modelling of M/G/1/K Queueing Systems. Elsevier

[8] Jagerman, D. L., Balcioglu, B., Altiok, T. & Melamed, B. 2004. Mean Waiting Approximations in the G/G/1 Queue. Kluwer Academic Publisher