International Journal of Computer Applications Technology and Research
Volume 7–Issue 11, 407-411, 2018, ISSN:-2319–8656

# Performance Analysis of Load Balancing and Physical Machines in Cloud Computing Centers with General Service Time

Ashri Shabrina Afrah
Department of Electrical Engineering
University of Brawijaya
Malang, East Java, Indonesia

Moechammad Sarosa
Department of Electrical Engineering
State Polytechnic of Malang
Malang, East Java, Indonesia

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

**Abstract**: The number of Virtual Machines in the Cloud Data Center may affect the value of Cloud Data Center's utilization. When there are too many Virtual Machines dedicated in the Physical Machines, the utilization is tend to be low. As the result, some idle Virtual Machines will consume much electricity [12]. However, the adequate number of Virtual Machines should be used to maintain the data center's performance, which is indicated by some QoS parameters, such as queue length, system response time, and drop rate. The system model which is observed this research insists of a Load Balancing and some Physical Machines which contain a number of Virtual Machines in each. However, the commonly used Markovian Model of Queuing Theory will be replaced by The General Model, which is more suitable to the characteristics of the Cloud Data Center. This research is aimed to get the optimum number of Virtual Machines in a Cloud Data Center to improve its performance. The simulation results show that 25 Virtual Machines in each Physical Machine are needed to obtain the optimum level of utilization and other indicator parameters.

**Keywords**: Data Center; Cloud Computing; Queuing Theory; Utilization; Virtual Machine; General Service Time

## 1. INTRODUCTION

The developing of Cloud Computing has attracted more companies and individuals to use this service. This fact motivates a lot of researchers to find some new ways to improve the performance of Cloud Data Centers as the heart of Cloud Computing service. Every day, a Cloud Data Center receives so many tasks from the users. These tasks arrive in unpredictable time, in a huge amount, to be served in the data centers in a random service time. As the results, some tasks have to wait in a line before being served. According to this situation, Queuing Theory can be used to analyze the queue of user's tasks in the Cloud Data Center.

The number of Virtual Machines which are used in the Cloud Data Center may affect the value of Cloud Data Center's utilization. When there are too many Virtual Machines dedicated in the Physical Machines, the utilization is tend to be low. As the result, some idle Virtual Machines will consume much electricity. However, the adequate number of Virtual Machines should be used to maintain the data center's performance, which is indicated by some QoS parameters, such as queue length, system response time, and drop rate.

The system model which is observed in this research is according to the model used in the previous research which is conducted by Said El Kafhali and Khaled Salah, which insists of a Load balancing and some Physical Machines [2]. In each Physical Machine, there are some Virtual Machines dedicated to serve the user's requests. However, the commonly used Markovian Model of Queuing Theory will be replaced by the General Model, because it is more suitable to the characteristic of the Cloud Data Center. The dynamic characteristic of Cloud Data Center results in higher service time's variance coefficient, so the conventional Markovian Model isn't relevant [1][6]. In this research, M/M/1/C and M/G/m/K queue model will be used.

This research is aimed to get the optimum number of Virtual Machines in a Cloud Data Center to improve its performance. Some parameters which are used as the performance indicators are queue length, system response time, and drop rate are able to measure using Queuing Theory [3].

## 2. LITERATURE REVIEW

In other research, Queuing Theory is applied to observe the adequate number of Virtual Machines to meet the defined performance requirements of the Service Level Objective (SLO) [2]. The research modeled some cloud servers (Physical Machine) containing some Virtual Machines by using M/M/m/K (K>m) Queuing model, and a Load Balancing with the M/M/1/C Queuing model. With numerical examples, this research showed how the developed model was able to estimate the number of Virtual Machine needed to fulfill the set QoS parameters [2].

Another research was about the application of Queuing Theory that modeled the Cloud Data Center by using M/G/m/m+r queue model [6]. Here the service time was modeled in general. As the service time's variance coefficient was extremely high, common Negative Exponential Distribution could not model it really well. Similar research was also carried out by Tulin Atmaca et al. [1]. In there, Cloud Data Center was modeled with the queue of G/G/c. Therefore, it can be concluded that General Model is more appropriate to use to analyze the service time in Cloud Data Center.

## 3. THE OBSERVED MODEL

### 3.1 Cloud Data Center Utilization

In order to operate a Cloud Data Center, a Cloud Computing service provider company needs to make a huge investment for the hardware, software, supporting operational components and also energy. For that reason, the service provider company has to do a thorough identification to find out whether they have set the Cloud Data Center in the

appropriate level, especially in terms of utilities (utilization) [5].

Utilities are strongly related to budgetary factor since the higher the utilization, the more efficient the performance of Cloud Data Center is. Low utilization, for example due to the idle status of a server, consumes a lot of electricity which affects the operational cost of the server. In addition, Cloud Data Center produces carbon dioxide emissions (CO2) so that the utilities are optimized and pollution can also be reduced due to the emission material [12].

Cloud Data Center's utilization is the level of CPU utilization in a set range of time, as seen in the formula [8]:

$$U = \frac{\sum_{n=1}^{T}(CPU\ Rate)}{T} \quad (1)$$

With:

U = Cloud Data Center utilization

CPU Rate = the level of CPU utilization

T = time range of utilization measurement

Thus, to sum up, utilities are closely related to how far CPU is actively working to process the tasks from the user. If in time range T the CPU is often in idle, the utilization will be down.

## 3.2 The Observed System's Model

This research discusses the system in a Cloud Data Center used by service provider IaaS (Infrastructure as A Service). In IaaS, the service provider provides IT infrastructure such as server user, storage memory, Virtual Machine and operating system based on users' needs (on demand). Cloud Data Center is a group of servers used by IaaS to fulfill consumers' needs. In this study, the system model consists of Load Balancing unit and some Physical Machines. Each server machine has a number of Virtual Machine as seen in the model used in the previous literature [2].

The Cloud Data Center's working principles is explained as follows:

1. For instance, an online shop uses the service of IaaS to host the website and store the order data of the consumers. The website and data basis of the online shop is store in Physical Machine in Cloud Data Center.
2. Some users use the service provided by Cloud Data Center from different places. When users access the website through browsers, the request from the users is sent from the device to Cloud Data Center. As Cloud Data Center consists of some Physical Machines, a Load Balancing is needed to balance the load of each Physical Machine. Generally, Load Balancing unit works based on a certain algorithm, for example Round Robind. The request from the users will get into the Load Balancing before being directed to the Physical Machines.
3. Then, a number of Physical Machines will do their function to serve the request from the users. There are some Virtual Machines in it whose working principle resembles parallel computing. After the request is processed, the data will be sent to the hardware so the users can use facilities managed by the website.
4. To ensure the service quality received by the user, performance measurement parameters are usually used. The parameters are responding time of the system, the number of tasks in the system, blocking probability, throughout, and so on.

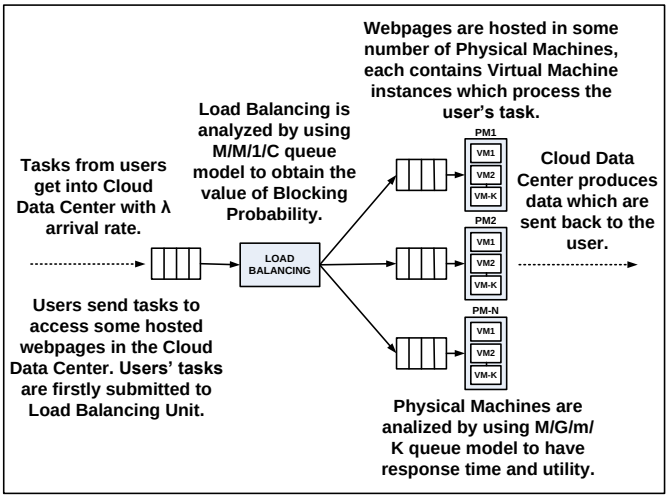


Figure. 1 The qqueuing model of Cloud Data Center

The model system will be analyzed by using the Queuing Theory and then tested through a simulation using Java Modeling Tools application. Tasks from users, for example the request to access a website hosted in Physical Machine, come to the Cloud Data Center with the average arrival of $\lambda$. $N$ symbol in the diagram is the number of Physical Machines in the Cloud Data Center, and $K$ is the number of Virtual Machines in every Physical Machine.

## 3.3 The Queuing Model M/M/1/C for Load Balancing's Analysis

M/M/1/C is a queue model with the arrival times which is distributed as Poisson, Negative Exponential distributed service times, and a server. The $C$ in the model is the total number of tasks loaded in the queue of Load Balancing:

$$C=N\times K \quad (2)$$

With:

C = total capacity of Load Balancing's queue system

N = the number of Physical Machine in Cloud Data Center

K = total capacity in each *Physical Machine*

### 3.4 The M/G/m/K Queue Model for Physical Machines

The calculation of the performance parameters values in this research is done according to the M/G/m/K queue model of Queuing Theory. M/G/m/K means that the queue have Poisson distributed arrival times, general service times, and *m* Virtual Machines in each Physical Machine. The total capacity of each Physical Machine is *K*.

Based on the previous research, the value of utilization in each server can be counted as follows [1]:

$$U = \sum_{n=1}^{C-1} P_n \times \frac{n}{m} + \sum_{n=c}^{N} P_n \quad (3)$$

With:

U = the utilization of each server in the queue system

Pn = steady state distribution

N = the number of tasks in the system

m = the number of server in the system

A blocking process occurs when the number of tasks in the Load Balancing is equal to the maximum capacity of the queuing system in the Load Balancing. Therefore, blocking probability can be calculated as [7]:

$$P_C = \frac{(1-\rho)}{(1-\rho)^{C+1}} \times (\rho)^C \quad (4)$$

With:

$P_C$ = blocking probability

ρ = arrival rate/service rate (λ/μ)

C = maximum capacity of Load Balancing's queuing system

The formula to obtain the average number of tasks in Cloud Data Center and the response time of the system can be found using M/G/m/K queuing model of Queuing Theory. In the research conducted by Hamzeh Khazaei et. al, average number of tasks in the system with general service time in the Cloud Data Center can be obtained from the derivation of the distribution function in every time unit [6]:

$$P(t) = \sum_{k=0}^{K} P_k \times t^K \quad (5)$$

$$\bar{n} = P'(1) \quad (6)$$

with:

P(t) = distribution function of the number of tasks in the system in every time unit

P(k) = steady state function

k = the number of tasks in the system

t = time unit

n = average number of tasks in the queuing system

The response time value can be calculated as [6]:

$$r = \frac{\bar{n}}{\lambda} \quad (7)$$

with:

r = average response time of the system

n = average number of tasks in the system

λ = *arrival rate*

## 4. SIMULATION RESULTS AND DISCUSSION

### 4.1 Testing Procedures

The testing procedures using Java Modeling Tools (JMT) simulation software can be explained as follows:

1. The first step is to determine the minimum utilization value, maximum queue length, expected drop rate, and maximum responding time. In this research, the minimum utilization value is 0.75, maximum queue length in Load Balancing is 5, expected drop rate is 0, and maximum responding time is 0.057 second.
2. The next, the initial amount of Virtual Machine in each Physical Machine is set, which is 30. The number of virtual machines will be subtracted by 10 in each simulation step, which is 1 in every Physical Machines.
3. In every step of simulation, the average of arrival rate is 100, 400, 700, 1000, 1300, 1600, 1900, 2200, and 2500. The number of simulated Virtual Machine is 275, 270, 265, 260, 255, 250, and 245. Another simulation is done to find out the effect of service rate in Load Balancing towards utilization and other performance parameters. The value of simulated service rate Load Balancing is 9000, 9200, 9300, 9400, 9500, 9600, 9700, 9800, 9900, and 10000.
4. An observation is done towards the result of the simulation. For the utilization parameter, the observation is done in Physical Machine 1.
5. Analyzing the result of simulation to find out:

a. The optimum number of virtual machines, in which the optimum utilization value and other performance parameters must fulfill the defined limit.

b. The effect of service rate from Load Balancing towards the performance of Cloud Data Center, measured from the utilization and other performance parameters.

### 4.2 The Load Balancing Analysis

With the Queuing Theory model M/M/1/C, the analysis of Load Balancing in Cloud Data Center can be done. The average of queue length goes down as the change of service rate value of Load Balancing, from 9000 to 9900. The average

responding time and Physical Machine utilization in PM1 tend to be stable (not affected by the change of service rate in Load Balancing). The reason is, the average of responding time and utilization are measured in Physical Machine. If the Physical Machine is overload, the length of the queue in Load Balancing is not affected because the data package from the Physical Machine queue will not return to the Load Balancing.

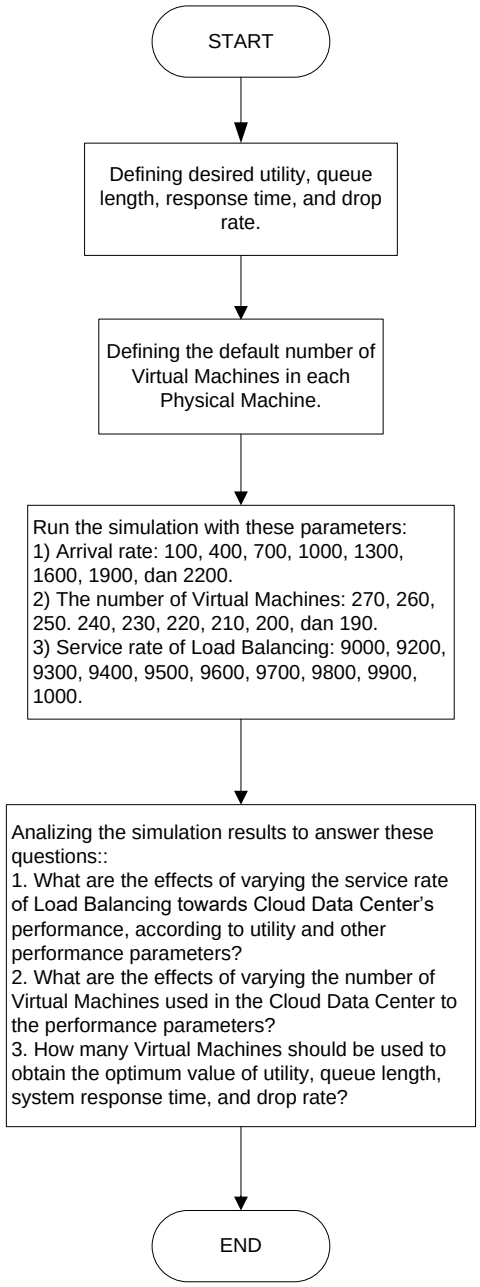


Figure. 2 Testing procedures for validating the model

## 4.3 The Analysis Of Physical Machines

In this research, the Physical Machine in Cloud Data Center is analyzed with the queue model M/G/m/K from the Queuing Theory. In the simulation, 10 Physical Machines were made. Each Physical Machine has the capacity system of 30 and some variation number of Virtual Machine from 19 to 27. The highest utilization based on the simulation happened when the number of Virtual Machine was 190 and the lowest is when it was 270. This result was showed in a lot of arrival rate variation from 100 to 2200.

The alteration of Virtual Machine's number does not give any impact to the queue length of Load Balancing and the drop rate remains 0. This happened because of the observed queue was in Load Balancing, and based on the research problem which was explained previously, the queue length has no correlation with Physical Machine.

Besides affecting utilization value, the change of Virtual Machine number also affected the responding time of the system. The fewer the Virtual Machines, the higher the responding time. There are 2 magnitudes that affect the value of responding time: arrival rate and arrival service. Based on the simulation, the number alteration of the Virtual Machines gave a bigger impact compared to arrival rate to responding time. We can see this through figure. 3. As the number of Virtual Machines is increased, the response time decreases. However, when we increase the arrival rate, there is no clear effect to the response time. The alteration of Virtual Machine caused the big change of service rate, while the simulated variation arrival rate is not so great compared to the ability of Physical Machine in handling the request from the user.

## 4.4 Determining the Number of Virtual Machines

The number of Virtual Machine that produced the best utilization is 230-250. When the Virtual Machine is 230 – 250, the drop rate was 0 for every arrival rate, the average response time was below 57 ms, and the utilization is above 0,75. When the Virtual Machine was more than 250 or less than 230, there was reduction of utilization value or escalation of the average response time. Therefore, the best number of Virtual Machine based on the data is 250.

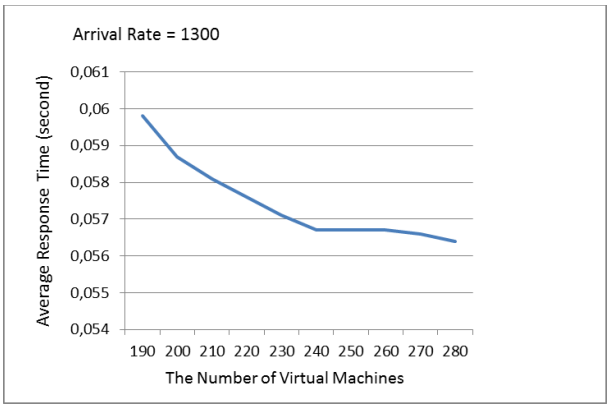


Figure. 3 The average response system time in every number of Virtual Machines for the number of arrival rate 1300

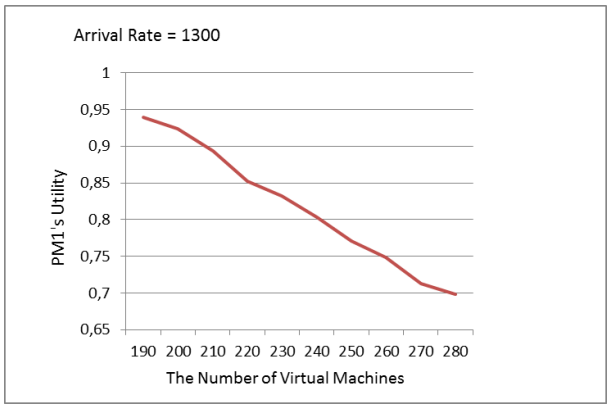


Figure. 4 The utilization of PM1 for every number of Virtual Machines for average arrival rate 1300

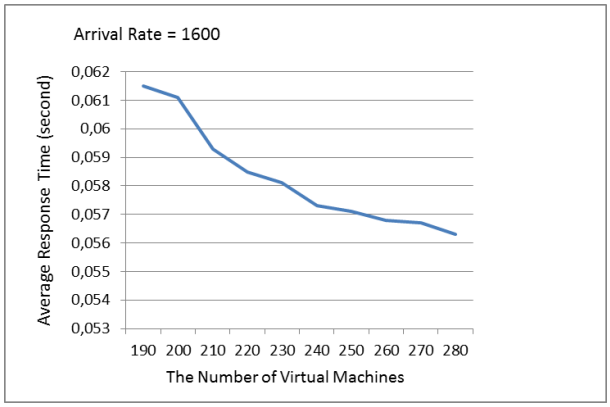


Figure. 5 The average response system time for every number of Virtual Machines for the arrival rate 1600

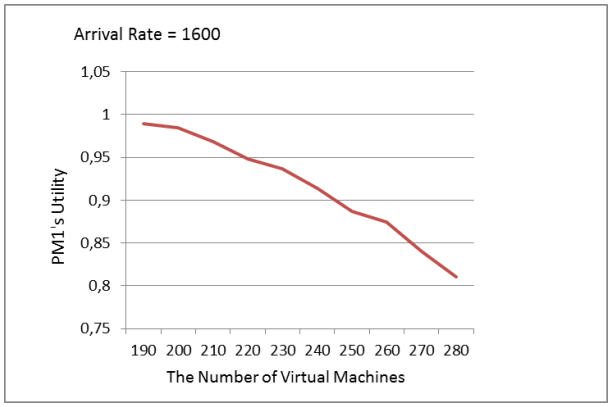


Figure. 6 The utilization of PM1 in every number of Virtual Machines for the arrival rate 1600

## 5. CONCLUSION

This research observes the effect of the number alteration of Virtual Machine in a Cloud Data Center. The architecture used here consisted of a Load Balancing and several Physical Machines. The method used was the queue model M/M/1/C and M/G/m/K from the Queuing Theory. Based on the simulation, it is concluded that the optimum number of Virtual Machine to increase the utilization was 23 – 25 in each Physical Machine. The number alteration of Virtual Machine affected the system response time, in which the fewer number of the Virtual Machine used, the higher the response system time is. Beside that, it can be concluded as well that the number of Virtual Machine did not give any impact to the length of the queue in Load Balancing.

## 6. REFERENCES

[1] Atmaca, T., Begin, T., Brandwajn, A., & Castel-Taleb, H. 2016. Performance Evaluation of Cloud Computing Centers with General Arrival and Service. IEEE

[2] El Kafhali, Said. & Salah, Khaled. 2017. Stochastic Modeling and Analysis of Cloud Computing Data Center. IEEE Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.

[3] Guo, L., Yan, T., Zhao, S., & Jiang, C. 2014. Dynamic Performance Optimization for Cloud Computing Using M/M/m Queuing System. Journal of Applied Mathematics Vol. 2014, Hindawi Publishing Corporation

[4] Hurwitz, J., Bloor, R., Kaufman, M. & Halper, F. 2010. Cloud Computing for Dummies. *Wiley Publishing Inc*.

*[5]* Hwang, K., Fox, Geoffrey C., & Dongarra, Jack C. 2012. Distributed and Cloud Computing. *Elsevier (Singapore) Inc.*

*[6]* Khazaei, H., Misic, J., & Misic, Vojislav B. 2012. Performance Analysis of Cloud Computing Centers Using M/G/m/m+r Queuing Systems. *IEEE Transactions On Parallel and Distributed Systems Vol. 23*

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

[8] Pawlish, M., Varde, Aparna S. & Robila, Stefan A. 2012. Analyzing Utilization Rates in Data Centers for Optimizing Energy Management. *IEEE International Green Computing Conference*

[9] Rittinghouse, John W. & Ransome, James F. 2010. Cloud Computing Implementation, Management, and Security. *CRC Press, Taylor & Francis Group*

*[10]* Shahin, Ashraf A. 2017. Enhancing Elasticity of SaaS Application Using Queuing Theory, *International Journal of Advanced Computer Science and Application Vol. 8*

*[11]* Velde, V. & Rama, B. 2017. Simulation of Optimized Load Balancing and User Job Scheduling Using CloudSim. *2nd International Conference On Recent Trends In Electronics Information & Communication Technology*

*[12]* Vondra, T. & Sedivy, J. 2017. Cloud Autoscaling Simulation Based On Queuing Network Model. *Simulation Modeling Practice and Theory Vol. 70, Science Direct*