
ISSN: 2321-2152
IJMECE
International Journal of modern
electronics and communication engineering
E-Mail
editor.ijmece@gmail.com
editor@ijmece.com
www.ijmece.com

ISSN2321-2152www.ijmece .com

Vol 8, Issuse.2 June 2020

Utilization of Virtual and Physical Machines in Cloud Computing Centers with General Service Time

[1]V.Sujatha, [2]A.Alekya ,[3]V.Muni Babu

ABSTRACT:

In certain cases, the Cloud Data Center's usage might be impacted by the number of virtual machines housed inside. When there are a lot of Virtual Machines committed to Physical Machines, the utilization tends to be poor. A few idle Virtual Machines will end up using a lot of power as a consequence of this. It is possible to gauge the efficiency of a data center's QoS measures, such as the length of the queue, response time, and drop-out rate. This sort of system is made up of a load balancer and a number of real computers, each of which has a virtual machine. As a result, Queuing Theory's Markovian Model will be replaced with the General Model, which better reflects the Cloud Data Center's unique characteristics. The study's objective is to discover the ideal number of virtual machines for a cloud data centre in order to boost performance. The simulation findings reveal that each Physical Machine requires 25 Virtual Machines in order to achieve the optimal level of utilization and other key metrics.

INTRODUCTION

Cloud Computing has grown in popularity as a result of the advancement of this technology. Cloud Data Centers are the core of Cloud Computing, and academics are looking for methods to increase their performance. A Cloud Data Center gets a plethora of requests from consumers every day. These jobs come at random times, in large quantities, and are serviced at random times in the data centres. As a consequence, there is a wait list for certain jobs to be completed. Based on the current scenario, Queuing Theory may be used in order to examine the Cloud Data Center's work queue.

"The Cloud Data Center's usage may be affected by the amount of Virtual Machines that are utilised in the Cloud Data Center. Utilization is reduced when there are too many Virtual Machines in a Physical Machine." As a consequence, certain Virtual Machines will use a lot of power while they are just sitting there. A sufficient number of virtual machines, as indicated by QoS characteristics like as queue length, system response time, and drop rate[11], should be employed in order to sustain the data center's performance.

Said El Kafhali and Khaled Salah's prior research, which relied on a Load balancing model and physical machines [3], seems to be the basis for the system model seen in this study. To suit the user's needs, For every physical machine, there are a slew of virtual counterparts. The General Model of Queuing Theory, which is better suited to the features of Cloud Data Centers, will eventually replace the Markovian Model. Because of Cloud Data Center's dynamic nature, the standard Markovian Model is no longer applicable [1]. It will be utilised in this study to study the queues of the M/1/1C and M/G/1/1C models.

[1,2,3]Assistant Professor
[1,2,3] Department of Computer Science & Engineering,
[1,2,3]Dr.K.V.Subba Reddy College Of Engineering For Women

Cloud Data Center performance may be improved by determining the optimal number of Virtual Machines in a Cloud Data Center. Queuing Theory may be used to gauge the performance metrics such as queue length,

response time, and drop rate. [5].

1. LITERATUREREVIEW

It has previously been shown that Queuing Theory may be used to determine the number of Virtual Machines

needed to achieve the Service Level Objective (SLO) [7]. With a queueing model of M1/M1/M/K (K>m) and a load

balancing of M1/M1/C, the researchers modelled some cloud servers (Physical Machines) including some Virtual

Machines. Using real-world data, this study demonstrated the model's ability to accurately predict the number of Virtual Machines required to meet the

specified QoScriteria[6].

The M/G/m/m+r queue model was used to represent the Cloud Data Center in another study based on Queuing Theory [9]. Service time was approximated broadly in this case study. Due to the large variation coefficient of service time, Negative Exponential Distribution was unable to accurately represent it. There was also a similar study done by TulinAtmaca and colleagues [10]. The G/G/c queue was used to mimic the Cloud Data Center in that environment. We may then deduce that a General Model is better suited for our purposes in assessing Cloud Data Center availability.

2. THE OBSERVED MODEL

"Cloud Data Center Utilization"

For a Cloud Data Center to function, a service provider organization must spend a significant amount of money on hardware and software as well as other operational components. Therefore, the service provider must do a comprehensive investigation to determine if the Cloud Data Center has been configured at the proper level, particularly with regard to utility use [5].

The greater the usage of Cloud Data Center's resources, the more cost-effective it will be in terms of its performance. Because servers sit inactive for long periods of time, they use a lot of power, which has a direct impact on their operating costs. Carbon dioxide emissions (CO2) are also produced by the Cloud Data Center, allowing the utilities to be optimized and pollution to be decreased [12].As seen in the formula [8]: Cloud Data Center utilization is the level of CPU consumption over a certain period of time]:

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

TheObservedSystem'sModel

A Cloud Data Center utilized by service provider IaaS is the subject of this study (Infrastructure as A Service). IT infrastructure such as server users, storage memory, virtual machines and operating systems are all provided by the service provider under IaaS. (on demand). IaaS's Cloud Data Center is a collection of servers utilized to meet the demands of end users. A load balancing unit and a few physical machines make up the system model in this research. The paradigm employed in the prior literature [2] shows that each server machine has a certain number of Virtual Machines.

The functioning concepts of the Cloud Data Center are outlined as follows:

1. For example, an online retailer may utilizeIaaS to host its website and keep customer purchase information. The online store's website and data are housed on a physical machine in a cloud data center's data centre.
2. Cloud Data Center's service may be accessed from all over the world by certain of its customers. The user's browser sends a request to the Cloud Data Center when they visit the page. The load on each Physical Machine in a Cloud Data Center must be balanced, since the Cloud Data Center is made up of many Physical Machines. In general, a Load Balancing unit uses an algorithm like Round Robind to balance load. When a user requests a service, it will first be routed via the Load Balancing system.
3. It will then be up to a variety of physical machines to fulfil the user's request. Among the Virtual Machines, there are those that function like parallel processors. Requests are processed and then data is transmitted to the hardware so that users may use the services that are provided by this site.
4. Performance measurement metrics are often used to guarantee the quality of service received by the user. The parameters include the system's response time, the amount of jobs it has, the likelihood of a block, and so on.

The queueing theory will be used to analyse the model system, and the Java Modeling Tools programme will be used to simulate and test it. The Cloud Data Center receives requests from users, such as a request to view a website housed on a Physical Machine, on an average of. The Cloud Data Center has N physical machines, and each of them has K virtual machines. This is represented by the N and K symbols in the figure.

The Queuing Model M/M/1/C for LoadBalancing'sAnalysis

Poisson-distributed arrival times, Negative Exponentially-distributed service times, and a server comprise the M/M/1/C queue model. For Load Balancing, the C parameter represents the overall number of jobs in the queue:

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

With:

C=total capacityofLoad Balancing's queuesystem

N = the number of Physical Machine in Cloud Data CenterK= total capacityineach*Physical Machine*

"The M/G/m/K Queue Model forPhysicalMachines"

The M/G/m/K queue model of Queuing Theory is used to calculate the performance parameter values in this study. Mean arrival times, general service times, and m

Virtual Machines per Physical Machine in queues with M/G/m/K distributions. All of a Physical Machine's capabilities are K.Based on the previous research, the value of utilizatio ecountedasfollows [1]: $U = \sum_{n=1}^{c-1} P_n \times \frac{n}{m} + \sum_{n=c}^{N} P_n$

With:

4. SIMULATION RESULTS ANDDISCUSSION

TestingProcedures

ThetestingproceduresusingJavaModelingTools(JMT)
U=theutilizationofeachserverin thequeue systemPn=steadystate distribution
N= the number of tasks in the systemm=thenumber ofserver in thesystem

When the number of jobs in the Load Balancing equals the maximum capacity of the queuing system, a halting process begins. In this case, blocking probability may be estimated as [7]:

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

With:
P_C=blockingprobability
ρ =arrivalrate/servicerate(λ/μ)
C=maximumcapacityofLoad Balancing'squeuingsystem

For Cloud Data Center, Queuing Theory's M/G/m/K queuing model may be used to get an average number of jobs and system response times. HamzehKhazaei et al. found that the average number of jobs in the system with general service time in the Cloud Data Center can be determined from the derivation of the distribution function for each time unit [6]:

$$P(t) = \sum_{k=0}^{K} P_k \times t^K \quad (6)$$

with: $\bar{n} = P'(1)$
"P(t)=distribution functionofthenumberoftasksinthesystemineverytimeuni t
P(k)=steadystatefunction
k =thenumberoftasksinthesystemt =timeunit
n =averagenumberoftasksinthequeuingsystem"
Theresponsetimevaluecan becalculatedas[6]:

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

with:
r =averageresponsetimeofthesystem
"n = average number of tasks in the systemλ =*arrival rate"*

5. SIMULATION RESULTS ANDDISCUSSION

TestingProcedures

ThetestingproceduresusingJavaModelingTools(JMT) is as follows.

1. The first stage is to calculate the minimum utilisation value, maximum queue length, predicted drop rate, and maximum response time. Minimum utilisation is 0.75, the maximum queue length in Load Balancing is 5, and the predicted drop rate is 0%. The maximum response time is 0.057 seconds in this study.
2. Each Physical Machine has 30 virtual machines installed at the beginning of the process. Each simulation step will remove 10 virtual machines, or one for every physical machine.
3. When calculating arrival rates in each phase of the simulation, the average number of arrivals is between 100 and 2500. More than two hundred and seventy-five virtual machines have been created. Using yet another simulation, we examine the impact of the service rate on several performance metrics, including system usage. 9000, 9200, 9300, 9400, 9500, 9600, 9700, 9800, 9900, and 10000 are the simulated service rates for Load Balancing.
4. As a consequence of the simulation, an observation is made. Physical Machine 1 is used for the monitoring of the utilisation parameter.
5. Finding out what happened in the simulation:

a. Maximum number of virtual machines required to achieve optimal usage and other performance goals.
b. The impact of Load Balancing service rate on Cloud Data Center performance as evaluated by utilisation and other performance metrics.

TheAnalysisOf PhysicalMachines

An analysis of the Physical Machine in a Cloud Data Center is conducted using the M/G/m/K Queuing Theory model. Ten real-world machines were created in the computer simulation. Capacity for each Physical Machine varies between 30 and 19 to 27 virtual machines. Based on the simulation, the maximum usage occurred when the number of Virtual Machines was 190, and the lowest occurred when it was 270. An abundance of arrival rate fluctuation was seen between 100 and 2200, indicating this outcome.

When changing the number of virtual machines, there is no change in queue length or drop rate. Load balancing caused this, and as previously established in respect to queue lengths, there is no link between queue length and physical machine.Not only did this alter usage, but it also slowed down system response time. As the number of Virtual Machines decreases, the response time increases proportionately. Arriving service and arrival rate are two factors that influence the value of response time. According to the results of the experiment, changing the number of Virtual Machines had a greater influence on response time than changing the arrival rate.. For example, as you increase the number of Virtual Machines, response times drop. There is, however, no noticeable change in reaction time when the arrival rate is increased. The change in Virtual Machine generated a significant change in service rate, but the simulated variance arrival rate is not as substantial as the physical machine's capacity to handle user requests.

Determining the Number of Virtual Machines

It took 230-250 virtual machines to get the highest level of virtual machine efficiency. Zero percent of the arrivals were dropped, the reaction time was less than 60

milliseconds and utilisation was over 75% when the virtual machine is in the 220–250 range. A decrease in utilisation or an increase in average response time occurred when the Virtual Machine was more than 250 or less than 230 In light of the facts, the optimal number of virtual machines is 250.

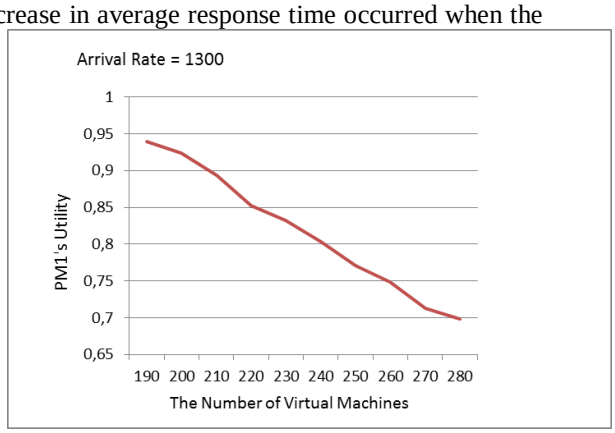

Figure. 4 The utilization of PM1 for every number of VirtualMachinesforaveragearrivalrate1300

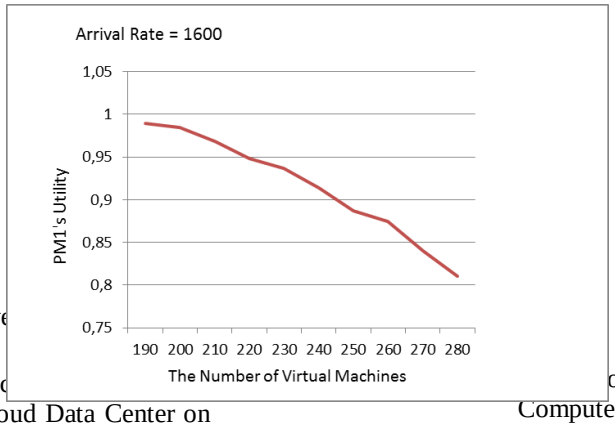

Figure. 6 The utilization of PM1 in eve

6. CONCLUSION

This study examines the impac number of Virtual Machines in a Cloud Data Center on the overall performance of the system. Several Physical Machines were employed in conjunction with a Load Balancing in order to achieve this architectural design. It was based on the Queuing Theory's queue models M/1C and M/G/m/K. A virtual machine with 23 – 25 virtual machines in each physical machine was shown to be the most effective in terms of increasing utilization. Changing the number of Virtual Machines utilized had an impact on the system's reaction time, which increased as the number of Virtual Machines decreased. It may also be argued that the number of virtual machines does not affect the duration of queues in load balancing, as well.

REFERENCES

1. Murdoch, J. 1978. Queuing Theory Worked Examples and Problems. The Macmillan Press. Ltd
2. Pawlish, M., Varde, Aparna S. &Robila, Stefan A. 2012. Analyzing Utilization Rates in Data Centers for Optimizing Energy Management. IEEE International Green Computing Conference
3. Rittinghouse, John W. &Ransome, James F. 2010. Cloud Computing Implementation, Management, and Security. CRC Press, Taylor & Francis Group
4. Shahin, Ashraf A. 2017. Enhancing Elasticity of SaaS Application Queuing Theory, onal Journal of Advanced Computer Science and Application Vol. 8
5. 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
6. Atmaca, T., Begin, T., Brandwajn, A., & Castel-Taleb,
7. H. 2016. Performance Evaluation of Cloud Computing Centers with General Arrival and Service. IEEE
8. 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.
9. 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
10. Hurwitz, J., Bloor, R., Kaufman, M. &Halper, F. 2010. Cloud Computing for Dummies. Wiley Publishing Inc.
11. Hwang, K., Fox, Geoffrey C.,

&Dongarra, Jack C. 2012. Distributed and Cloud Computing. Elsevier (Singapore) Inc.
12. 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