A Hybrid Queuing Model for Virtual Machine Placement in Cloud Data Center

Sourav Kanti Addya*, Ashok Kumar Turuk†
and Bibhudatta Sahoo‡
Department of Computer Science and Engineering
National Institute of Technology, Rourkela, India
Email: {*kanti.sourav, †akturuk,
‡bibhudatta.sahoo}@gmail.com

Mahasweta Sarkar
Department of Electrical and Computer Engineering
San Diego State University, CA, USA
Email: msarkar2@mail.sdsu.edu

***Abstract*—Virtual Machine (VM) placement is an important research area for power conservation in data centers. In this paper, we introduce a hybrid queuing model for VM placement for data centers to improve total placement time and earn more revenue. For the large data center management smaller placement times lead to greater quality of service (QoS) experienced by an user of the cloud infrastructure. Similarly, the larger the number of VMs that can be placed in a server racks, higher the justification of the placement cost. We thus propose a ILP, that provides maximum justification of the revenue earn along with minimizing placement time. In addition, we also study the rate of loss of VM placement requests and total utilization of the system as the arrival rate of VMs increase.**

Keywords*—Cloud computing, Virtual machine, Placement time, Revenue.

I. Introduction

Cloud computing is an emerging technology that uses the Internet and centrally located remote servers for maintenance of data and applications. Among all key technologies, virtualization enables dynamic sharing of physical resources in cloud environments [1]. Through virtualization, physical resources such as: CPU, memory, disk space are made available to applications on-demand [2].

The process of mapping VMs onto Physical Machines (PM) is known as VM placement. It is an important approach for improving energy efficiency and resource utilization in a cloud infrastructure.A Cloud Service Provider (CSP) will try to accommodate maximum number of VM requests onto data centers to earn his maximum revenue. Revenue can be defined as the cost incurred to map or place one VM on to a specific server. This cost is generally negotiated between the user and the Cloud Service Provider (CSP) during service level agreements (SLA). To maximally earn the revenue, the CSP aims to accommodate as many VMs as possible. However, placing VMs on a server takes time - the placement time. This metric is crucial to the success of a CSP in accruing subscribers. A longer placement time might lead to frustrated users. Additionally, longer placement times may not meet the client's technical requirements. The importance of appropriate placing of VMs on a data center has been addressed by several researchers [3], [4]. In this work, we analysis the total placement time for different sets of VM request arrival rates. We also consider the rate of arrival request loss during requesting processing in the system for placement. We proposed a queuing theory based mathematical model to address the previously mentioned issue. The proposed model was analyze and verified using simulation. Proposed model, simulation results and conclusion are describe in following sections.

II. Proposed Model

In this paper, we propose a hybrid multi-system VM placement model for cloud data center, shown in figure 1.In this model the VM placement requests will pass through two distinct system facilities namely, CSP system and Data Center (DC) system. CSP system is describe as a single server system associated with a Broker queue (BQ). VM placement requests will arrive at CSP system trough the BQ. After being processed from the CSP system, it will be assigned to a DC system. Each DC system has its own DC queue and have it's own a multi-server system. Note that the processing rate/departure rate at the CSP has a direct implication on the arrival rate of these requests to the DCs. Also note that there are multiple DCs (the number of DCs might vary) to which these requests will be distributed over.

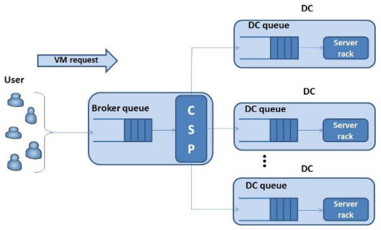


Fig. 1. Schematic Diagram of VM request flow for cloud model

Let λ_n be the arrival rate of n VM requests to the BQ in the CSP. Let μ_n be the service rate of the CSP. Let us also assume that the system functions under identical conditions.

Let P_n be defined as the steady state probability of n VM requests arriving at the CSP system at a given time t. From the generalized model of queuing theory, P_n is a function of λ_n and μ_n, which can be calculated as:

$$P_n = \left(\frac{\lambda_{n-1}...\lambda_0}{\mu_n...\mu_1} \right) P_0, \; n = 1, 2, ... \tag{1}$$

where, P_0 can be found from $\sum_{n=0}^{\infty} \mathrm{P_n} = 1$ i.e. the sum state probability of all states is equal to 1. Note that, $System =$

$queue + servicefacility$. Let R_s and R_q denote the expected number of VM requests in the system and the queue respectively. Thus, R_s = E[VM(s)] and R_q = E[VM(q)]. Obviously, $R_s > R_q$.

Additionally let T_s and T_q represent the expected waiting time of VM placement requests in the system and the queue respectively. Thus T_s = E[WT(s)], T_q = E[WT(q)], where WT represents WaitTime. In steady state, the probability of having n VM placement requests in the system, denoted by P_n can be used to compute R_s and R_q in the following manner:$R_s = \sum_{n-1}^{\infty} nP_n$ and $R_q = \sum_{n=c+1}^{\infty} (n-c)P_n$, where c denotes the number of parallel servers in the system. In our model c can take the value of 1 or m. if λ_{eff} denotes the effective arrival rate of VM placement requests in the system, then from Little's Theorem [5] we have $R_s = \lambda_{eff} T_s$ and $R_q = \lambda_{eff} T_q$. Where, λ_{eff} is the effective arrival rate in the system.

Note that λ_{eff} = normal λ when all placement requests arrive successfully to the system and no request is dropped by the network due to unfavorable system condition. Otherwise, $T_s = T_q + \frac{1}{\mu}$ and we can relate R_s to R_q by multiplying both side of the last formula by λ_{eff} , which together with the little formula gives, $R_s = R_q + \frac{\lambda_{eff}}{\mu}$. By the definition, the difference between the average number in the system R_s and the average number in the queue R_q, $\overline{c} = R_s - R_q = \frac{\lambda_{eff}}{\mu}$. Service facility utilization $= \frac{\overline{c}}{c}$.

A. Broker Queue System

We model the CSP system as the classic Broker Queue (BQ) from Queuing theory [5]. The BQ that we consider, has a single server model with finite system limit. The arrival rate of VM placement requests to this queue is denoted by λ. The requests are processed at a rate (service rate) μ per unit time. The system has an upper limit of accommodating upto (N-1) placement requests in the queue. Thus the N^th placement request and beyond is discarded.

Thus we have,

$$\begin{aligned} \lambda_n &= \begin{cases} \lambda, & n = 0, 1, ..., n-1 \\ 0, & n = N, N+1 \end{cases} \\ \mu_n &= \mu, \quad n = 0, 1, ... \end{aligned} \tag{2}$$

Using $\rho = \frac{\lambda}{\mu}$, the generalized model yields Equation 1.

$$P_n = \begin{cases} \rho^n P_0, & n \leq N \\ 0, & n > N \end{cases} \tag{3}$$

$$P_n = \begin{cases} \frac{(1-\rho)\rho^n}{1-\rho^{N+1}}, & \rho \neq 1 \\ \frac{1}{N+1}, & \rho = 1 \end{cases} \quad , \quad n = 0, 1, ..., N \tag{4}$$

The value $\rho = \frac{\lambda}{\mu}$, need not be less than 1 for this model, because the VM placement request arrival rate is bounded at N requests in the system at any pont in time t, i.e λ_{eff}. Because VM request will be lost there are N in the system.

We can write arrival loss rate as ,

$$\begin{aligned} \lambda_{lost} &= \lambda P_N \\ \lambda_{eff} &= \lambda - \lambda_{lost} = \lambda(1 - P_N) \end{aligned} \tag{5}$$

In this case, $\lambda_{eff} < \mu$. The expected number of VM request in the BQ system is computed as,

$$\begin{aligned} R_s &= \sum_{n=1}^{N} nP_n \\ &= \tfrac{\rho[1-(N+1)\rho^N + N\rho^{N+1}]}{(1-\rho)(1-\rho^{N+1})}; \;\; \rho \neq 1 \end{aligned} \tag{6}$$

or,

$$\begin{aligned} R_s &= \sum_{n=1}^{N} nP_n \\ &= \tfrac{N}{2}; \;\; \rho = 1 \end{aligned} \tag{7}$$

We can find $T_s = R_s/\lambda$; $T_q = T_s - 1/\mu$; $R_q = \lambda T_q$ and $\overline{c} = R_s - R_q$

B. Data Center Queue System

The Data Center Queue system comprises of m parallel number of service facilities. The DCQ system follows the BQ system. The departure rate from BQ will be the arrival rate of DCQ split over the m servicing facilities; where m is the total number of data center associated with CSP. The system limit is finite and equal to N'. The minimum queue size is $N' - c'$. VM placement request arrival rate at each of the individual m data centers is assumed to be λ' and the service rate to be μ'. The effective arrival rate $\lambda'_{eff} < \lambda'$ because system limit is N'. The system DC has following characteristic: (i.) The VM placement request arrivals occur in batches following the Poisson arrival model with mean rate λ'. (ii.) The requests are served in the DCQ system and leave the system following a Poisson model with mean rate μ'. (iii.) Number of parallel servers is c.

From the general queue model and equation 2 we can define λ'_n and μ'_n. Substituting λ'_n and μ'_n in the general expression of queuing theory and noting $\tau = \lambda'/\mu'$;

$$P'_n = \begin{cases} \frac{\tau^n}{n!} P'_0, & 0 \leq n \leq c' \\ \frac{\tau^n}{c'! c'^{n-c'}} P'_0, & c' \leq n \leq N' \end{cases} \tag{8}$$

Next, we compute R'_q for the case where $\frac{\tau}{c'} \neq 1$ as similar to the equation 6 and for $\tau/c' = 1$, then

$$R'_q = \frac{\tau^c (N' - c')(N' - c' + 1)}{2c'!} P_0, \quad \frac{\tau}{c'} = 1 \tag{9}$$

To determine T'_q and hence T'_s and R'_s, we compute the value of λ'_{eff}, similarly as in Equation 5

$$\begin{aligned} \lambda'_{lost} &= \lambda' P_N \\ \lambda'_{eff} &= \lambda' - \lambda'_{lost} = (1 - P_N)\lambda' \end{aligned} \tag{10}$$

Recall from Figure 1 that all VM placement requests undergoes service at two queue systems, namely the BQ followed by one of the m DCQs. The arrival rate of these requests at the BQ follows a Poisson process with mean rate λ and service/departure rate μ. The arrival rate (λ') of these requests to the DCQ is thereby a functionality of μ and the number of data center servers. Thus, for m data center servers, $\lambda' = \mu/m$. Total placement time for any VM request may be calculated from the following equation:

$$T = (T_s + T_q) + (T'_S + T'_q) \tag{11}$$

C. Deployment Cost Optimization

A VM which is placed in server rack, is defined by three tuples $VM_i.A$, $VM_i.C$ and $VM_i.B$. Where, $VM_i.A$ is deployment cost for i^{th} VM, which includes, hardware resource rent, placement charges and future maintenance cost. $VM_i.C$ denotes resource requirements of VM_i. $VM_i.B$ is an assignment variable and it is 1 if VM_i is placed in a server. With a higher arrival rate of VM requests and we intend to minimize the loss rate of requests (denoted by λ_{lost}) which will in turn lead to higher number of successful placements of VMs into servers thereby increasing the deployment cost. The objective function for maximizing the total deployment cost earned from active servers is outlined below (Equation 12).

$$Maximize \sum_{i=1}^{|VM|} VM_i.A \ \times \ VM_i.B \tag{12}$$

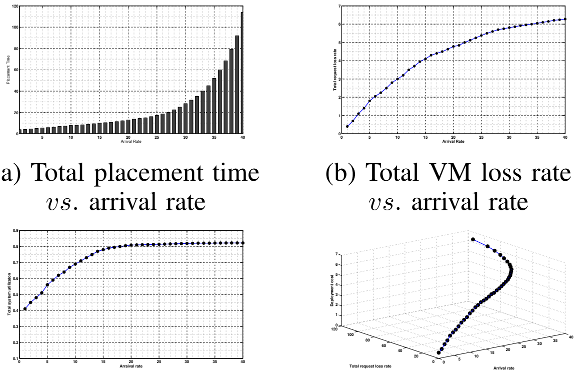
The deployment cost can be maximized by placing more number of VMs onto less number of active servers.

Subject to:

$$\begin{aligned} &\sum_{i=1}^{|VM|} VM_i.B = 1 \\ &\sum_{i=1}^{|VM|} VM_i.C \times VM_i.B \leq S_j \ ; \ \forall j \end{aligned} \tag{13}$$

III. SIMULATION & RESULTS

The model has been simulated using MATLAB 7.0 on a workstation computer with Intel(R) core(TM)2 Duo cpu of 3.00 GHz and 4.00 GB memory. In our numerical study, we varied the arrival rate of the VM placement requests (λ) and studied its effect on different performance metrics namely the expected waiting time of these requests in the queue and the system in both BQ and DCQ, system utilization and lost requests.



(a) Total placement time $vs.$ arrival rate

(b) Total VM loss rate $vs.$ arrival rate

(c) Total system utilization $vs.$ arrival rate

(d) Deployment cost $vs.$ Deployment cost

Fig. 2. Analytical results

In Figure 2a,we vary the value of λ from 0 to 40 within a fixed interval of time and study its impact on the placement time in milliseconds. The graph shows a logarithmic increase in placement time with increasing λ values. However we see an exponentially increasing placement time when lambda values start growing beyond 35 arrivals/unit time. This is because with such high lambda values, the waiting time in queues also increase substantially. In Figure 2b, the ratio of λ_{loss} is decreases with increasing arrival rate. It is observed that for higher value of λ i.e. 35 and more, the λ_{loss} value are proportionally low. In Figure 2c, it is clearly observed from the graph that when the λ is increased i.e. the queue load is increased for system, the system-utilization is constant for higher VM placement request arrival rate. Figure 2d concludes that for higher VM request arrival rate (λ), the probability of more VM placed in serve racks also increases. Another reason is for higher value of λ the proportion of λ_{loss} is also decrease.

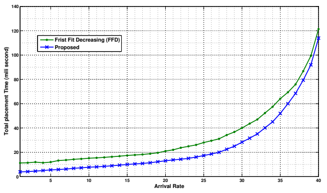


Fig. 3. Total placement time $vs.$ arrival rate

Figure 3 shown a comparative analysis with FFD [6]. The total placement time is better for proposed model than FFD.

IV. Conclusion

In this paper we addressed the issue of accommodating the maximum possible VMs on a server racks in a cloud infrastructure. We modeled the problem using a hybrid queuing system.Our aim was to study the effect of increasing placement requests on total placement time, deployment cost, request loss rate and total system utilization. We compared our result with FFD and the proposed technique is out perform.

References

[1] M. A. Vouk. Virtualization of information technology resources. In *Electronic Commerce: A Managerial Perspective 2008*. 5th Edition,Prentice-Hall Business Publishing.

[2] Dara Kusic, Jeffrey O Kephart, James E Hanson, Nagarajan Kandasamy, and Guofei Jiang. Power and performance management of virtualized computing environments via lookahead control. *Cluster computing*, 12(1):1–15, 2009.

[3] Michael Cardosa, Madhukar R. Korupolu, and Aameek Singh. Shares and utilities based power consolidation in virtualized server environments. In *Proceedings of the 11th IFIP/IEEE International Conference on Symposium on Integrated Network Management*, IM'09, pages 327–334, Piscataway, NJ, USA, 2009. IEEE Press.

[4] Yongqiang Gao, Haibing Guan, Zhengwei Qi, Yang Hou, and Liang Liu. A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. *Journal of Computer and System Sciences*, 79(8):1230 – 1242, 2013.

[5] János Sztrik. Basic queueing theory. *University of Debrecen: Faculty of Informatics*, 2011.

[6] Jzsef Bksi, Gbor Galambos, and Hans Kellerer. A 5/4 linear time bin packing algorithm. *Journal of Computer and System Sciences*, 60(1):145 – 160, 2000.