Stochastic Model for Cloud Dada Center with M/G/c/c+r Queue

Assia Outamazirt
Research unit LaMOS University of Bejaia
Bejaia, Algeria
outamazirt.assia@gmail.com

Mohamed Escheikh
SYS'COM ENIT Tunis
Tunis, Tunisia
mohamed.escheikh@gmail.com

Djamil Aïssani
Research unit LaMOS, University of Bejaia
Bejaia, Algeria
lamos_bejaia@hotmail.com

Kamel Barkaoui
CEDRIC, CNAM
Paris, France
kamel.barkaoui@cnam.fr

Ouiza Lekadir
Research unit LaMOS University of Bejaia
Bejaia, Algeria
ouizalekadir@gmail.com

Analytical resolution of complex queuing systems remains nowadays an open and challenging issue and may be extensively used in modeling and representing dynamic behavior of sophisticated systems. This is particularly the case of $M/G/c/c+r$ queue where exact analytical solution is difficult to reach. In this paper, we propose a new approximate analytical model in order to evaluate the performance of cloud computing center using $M/G/c/c+r$ queuing system. The adopted modeling approach combines two models. The first one is a transform-based analytical model whereas the second relies on an approximate Markov chain. This combination enables to compute the one-step transition probabilities for the system $M/G/c/c+r$.

Cloud Computing, Performance Evaluation, $M/G/c/c+r$ queue, Transition Matrix, Embedded Markov Chain

1. INTRODUCTION

The US National Institute of Standards and Technology (NIST) has defined cloud computing as a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction (Mell et al. (2009)).

Cloud services are usually divided into main three categories. Software-as-a-Service (SaaS) is a software delivery model in which software and associated data are centrally hosted on the cloud. Platform-as-a Service (PaaS) provides a computing platform like execution runtime, database, web servers, development tools etc. In infrastructure-as-a-Service (IaaS), cloud providers offer computers, as physical or more often as virtual machines, servers, storage, load balancers, networks, etc. (Furht (2010)). Typically a service level agreement (SLA) gives all the aspects of usage of cloud service and the obligations of service providers and clients. It also includes various descriptors known as quality of service (QoS) which includes availability, reliability, throughput, security, performance indicators (Wang et al. (2010)).

Providing quality services require a solid model that provides detailed insights of computing centers. Assessing the QoS offered by cloud providers necessitates the accurate performance evaluation of the cloud center. Khazaei et al. (2012) adopted $M/G/c/c+r$ queuing system as the abstract model for performance evaluation due to the characteristics of cloud computing centers: having Poisson arrival of task requests, generally distributed service time, large number of servers and the finite capacity of system.

However analyzing of $M/G/c/c+r$ queuing system is not an easy task, an exact solution for this model is only possible for special cases, such as exponential service, a single server, or no waiting queue at all. Consequently, an extensive research on performance evaluation of $M/G/c$ queuing systems was described in the literature (Nozaki et al. (1978), Yao (1985)). According to Khazaei et al. (2012), the $M/G/c$ approximation approaches that were proposed in the literature (Kimura (1983), Tijms et al. (1981), Boxma et al. (1979)) are not suitable for performance evaluation of cloud center. They

cited three issues that make it difficult to apply these approximations:

- a cloud center can have a large number of facility (server) nodes (Amazon (2010));
- the complex task service times and higher coefficient of variation (CoV) of service time distribution (Zhang et al. (2011), Reiss et al. (2012));
- due to the dynamic nature of cloud environments, diversity of users requests and time dependency of load, cloud centers must provide expected quality of service at widely varying loads (Reiss et al. (2012), Xiong et al. (2009)).

A new approximate approach used the combination of a transform-based analytical model and an embedded Markov chain model for the steady state probabilities of the $M/G/c/c+r$ queueing system was proposed by Khazaei et al. (2012). This new approach is suitable for cases of large number of servers and when the distribution of service time has a coefficient of variation more than one. Improvements for this approach was proposed by Chang et al. (2014). In these approaches, the instants of regeneration of the embedded Markov chain were misinterpreted and the system state transition behavior was not describe accurately. This paper makes some progress towards a good approximation for the computation of the one-step transition probabilities of the system $M/G/c/c+r$. This enables to enhance previous approximations.

The rest of the paper is organized as follows. Section 2 presents a brief overview of related work on cloud performance evaluation and performance characterization of queuing systems. In Section 3, we present the analytical model in detail. In Section 4, we conclude by outlining same possible new directions for future work.

2. RELATED WORK

Xiong et al. (2009) modeled a cloud center as the classic open network, from which the distribution of response time was obtained by using Laplace transformation. Using the distribution of response time, the authors found the relationship between the maximum number of customers, the minimal service resources and the highest level of services

Yang et al. (2009) proposed the $M/M/c/c+r$ queuing system for modeling the cloud center, which indicates that both inter-arrival and service times are exponential, the system has a finite buffer of size r and its distribution of response time was obtained.

Ani Brown Mary et al. (2013) modeled a cloud center as an $[(M/G/1) : (\infty/GD\ model)]$ queuing system with a single task arrivals and a task request buffer of infinite capacity. They used analytical methods to evaluate the performance of queuing system and solve it to obtain important performance factors like mean number of tasks, blocking probability and probability of immediate service. Mean as well as standard deviation of the number of tasks is computed.

Analysis of queuing systems with multiple servers and general distributed service time is more complex. For these queuing systems the steady state probability, the distributions of response time and the queue length cannot be obtained in closed form. Consequently, several researchers have developed many methods for approximating its solution.

Kimura (1983) developed a diffusion approximation model for the queue $M/G/c$. The main idea is to approximate formulas for the distributions of the number of customers. Kimura (1996a) described an approximation for the steady state queue length distribution in $M/G/c$ queue with finite waiting spaces.
A similar approach in the context of $M/G/c$ queues was described by Kimura (1996b), but extended so as to approximate the blocking probability and, thus, to determine the smallest buffer capacity such that the rate of lost tasks remains under predefined level.

Nozaki et al. (1978) proposed an approximation for the average queuing delay in a $M/G/c/c+r$ queue based on the relationship of joint distribution of remaining service time to the equilibrium service distribution. Smith (2003) proposed a different approximation for the blocking probability based on the exact solution for finite capacity $M/M/c/c+r$ queues. The estimate of the blocking probability is used to guide the allocation of buffers so that the blocking probability remains below a specific threshold.

However, the most of the approximations mentioned above are not directly applicable to performance analysis of cloud computing center due these following limitations (Khazaei et al. (2012)):

- for example, approximations proposed by Kimura (1996a), Smith (2003) are reasonably accurate when the number of servers is small, below 10 or so. They are not being suitable for the cloud computing centers with more than 100 servers;
- approximations proposed by Nozaki et al. (1978), Yao (1985) are inaccurate when the

coefficient of variation of the service time, CoV, is above 1.0;

- approximation errors are particularly pronounced when the traffic intensity ρ is small, and/or when both the number of servers c and the CoV of the service time are large (Kimura (1983), Tijms et al. (1981), Boxma et al. (1979)).

As these approximations mentioned above are not directly applicable to performance analysis of cloud computing center, Khazaei et al. (2012) described a new approximate analytical model using $M/G/c/c+r$ queueing system. In order to evaluate its performances, they used a combination of a transform-based analytical model and an embedded Markov chain model, which obtained a complete probability distribution of response time and number of task in the system.

Chang et al. (2014) proposed an approximate analytical model using $M/G/c/c+r$ queueing system for performance evaluation of cloud center close to the proposed by Khazaei et al. (2012), both divided the transition probabilities matrix of this system into four regions. The difference between them lie in the approximation formula of the transition probabilities matrix in regions 3 and 4, because according to these authors, the approximate formula proposed by Khazaei et al. (2012) is an inaccurate approximation for the transition probabilities in these regions.

3. THE ANALYTICAL MODEL

3.1. Model description

We consider a c-server queueing system with Poisson arrivals, general service times, and a capacity limit of $c+r$ for the number of tasks in the system, for modeling a cloud data center. In this model we assume that:

- all c servers render service in order of task request arrivals (FCFS);
- each busy server is independent of the other busy serves;
- if the waiting queue is empty and there is no new task request arrival, the server enters in the idle state;
- if the task arrives while the system capacity has already been attained, this task will depart immediately without service;
- each task is serviced by a single server.

The $M/G/c/c+r$ queuing system is a non-markovian process and can be analyzing by using the embedded Markov chain (EMC) technique Kleinrock (1975). This will be discussed below.

Task request arrivals follow a Poisson process, which means that task inter-arrival time $A_X(x) \triangleq P[X \leq x]$ is exponentially distributed with a rate of λ, its probability density function (pdf) is $a(x) = \lambda e^{-\lambda x}$ and its Laplace transform is

$$A^*(s) = \int_0^\infty e^{-sx} a(x) dx = \frac{\lambda}{\lambda + s}.$$

Task service times are identically and independently distributed according to a general distribution $H_Y(y) \triangleq P[Y \leq y]$ with a mean service time equal to $\overline{h} = \frac{1}{\mu}$. its pdf is $h_Y(y)$ and its Laplace transform is

$$H^*(s) = \int_0^\infty e^{-sx} h(x) dy.$$

The traffic intensity is $\rho \triangleq \frac{\lambda}{c\mu}$. The Residual task service time, H_+, is the time interval from an arbitrary point (an arrival point) during a service time to the end of the service time. This time is necessary for our model since it represents time distribution between a task arrival and departure of the task which was in service when task arrival occurred. The Laplace transform of H_+ is calculated by Takagi (1991) as:

$$H_+^*(s) = \frac{1 - H^*(s)}{s\,\overline{h}}. \quad (1)$$

3.2. Model Analysis

We use the same EMC technique adopted by Khazaei et al. (2012) for analyzing the $M/G/c/c+r$ queuing system. This EMC technique consists of selecting the Markov renewal points at the instant of a new task arrival to the system. We choose two consecutive arrivals to be the observation interval for the EMC. The basic structure of the EMC is shown in Fig.1. Therefore, we model the number of the tasks in the system (both those in the service and those in the waiting queue) at the moments immediately before the new task arrival, if we enumerate these instances as $0, 1, ..., c+r$, we obtain a homogeneous Markov chain with state space $S = \{0, 1, 2, ..., c+r\}$. This Markov chain is ergodic because it is irreducible, recurrent non-null, and aperiodic (Khazaei et al. (2012)).

Let t_n and t_{n+1} the moments of n^{th} and $(n+1)^{\text{th}}$ arrivals to the system respectively, while X_n and X_{n+1} indicate the number of tasks found in the system immediately before these arrivals, T_n denotes the inter-arrival time between t_n and t_{n+1},

B_{n+1} indicates the number of tasks which depart from the system between t_n and t_{n+1}.
In the rest of this paper we use T to denote any inter-arrival time.

We must calculate the transition probabilities associated with EMC, and so we define

$$p_{ij} \triangleq P(X_{n+1} = j | X_n = i), \tag{2}$$

otherwise, we must calculate

$$P(B_{n+1} = i + 1 - j | X_n = i), \tag{3}$$

i.e., the probability that $i - j + 1$ tasks are serviced during T. It is clear that

$$p_{ij} = 0, \qquad \text{for} \qquad j > i + 1. \tag{4}$$

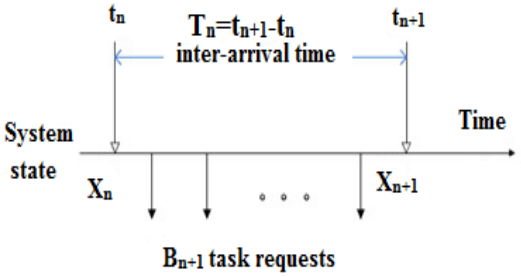

Figure 1: *Embedded Markov points.*

Since the EMC is ergodic, an equilibrium probability distribution $\pi = [\pi_0, \pi_1, \pi_2, ..., \pi_{m+r}]$ exists for the number of tasks present at the arrival instants with $\pi_i = \lim_{n\to\infty} P(X_n = i)$ for $0 \leq i \leq m + r$, and it is the solution of equation $\pi = \pi P$, where P is the matrix whose elements are the transition probabilities p_{ij}.

3.2.1. Transition Probability Matrix

To find the elements of the transition probability matrix P, we need to count the number of tasks departing from the system in an observation interval. Each server has zero or more departures in T. The transition probability matrix P has four regions as that defined by Khazaei et al. (2012) and Chang et al. (2014). Before presenting p_{ij} in each region, we first define the departure probabilities P_x, P_y and $P_{z,k}$ as follows:

$$P_x \triangleq P(A > H_+) = H_+^*(\lambda), \tag{5}$$

$$P_y \triangleq P(A > H) = H^*(\lambda), \tag{6}$$

$$P_{z,k} = \Big[\prod_{i=2}^{k} P(A > H | A > (k-i)H + H_+)\Big] \cdot P(A > H_+), \tag{7}$$

where:

- P_x is the probability of completing the service of a task, which has already been in service during the previous observation interval and is completed in the current interval.
- P_y is the probability of completing the service of a task, which begins to be serviced in the current interval and is finished within the same interval.
- If a server completes the service of a task which has already been begun during the previous observation interval, in the current interval, this server will be idle. If the waiting queue is nonempty, that server as well may complete a second service in the current interval, and if the waiting queue is still nonempty, a new service may be completed, and so on until the waiting queue gets empty. The probability of k services are completed by a single server is given by $P_{z,k}$.

After defining the departure probabilities P_x, P_y and $P_{z,k}$ in the embedded Markov chain, we describe the four different regions of the transition probability matrix P. These regions are schematically shown in Fig. 2, where the numbers on horizontal and vertical axes correspond to the number of tasks in the system immediately before a task request arrival (i) and immediately before the next task request arrival (j), respectively.

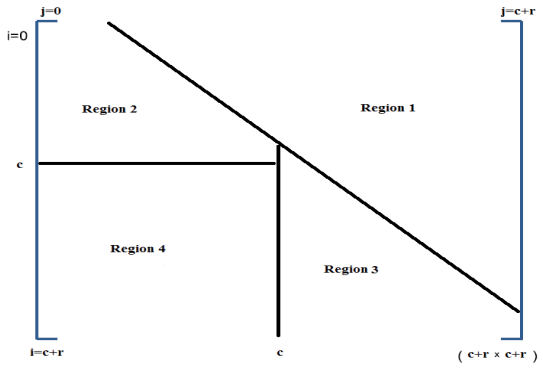

Figure 2: *Range of validity for* p_{ij}.

Region 1:

For $i + 1 < j$, $p_{ij} = 0$, since j cannot exceed $i + 1$ (Eq. 4).

Region 2:

For $i + 1 \leq c$, $0 \leq j \leq c$, and $i + 1 \geq j$, in this region, all tasks are in the service (no waiting). The

probability that $i-j+1$ tasks are served during T is:

$$P_{ij} = \begin{cases} C_i^{i-j} P_x^{i-j} (1-P_x)^j P_y + C_i^{i-j+1} P_x^{i-j+1} \\ (1-P_x)^{j-1}(1-P_y), \quad \text{if} \quad i+1 > j; \\ \\ (1-P_x)^j, \quad \text{if} \quad i+1 = j. \end{cases} \tag{8}$$

Region 3:

For $c \leq i \leq c+r$, $c \leq j \leq c+r$, and $i+1 \geq j$, all servers are busy during T. Let $\omega = i-j+1$ denotes the number of departures in the system, with $\omega \geq 0$ and, we assume each single server completes no more than three services of tasks in T. We define the transition probabilities for this region by:

$$p_{ij} = \begin{cases} M(i,j,w) \cdot \chi, & \text{if} \quad i < c+r; \\ \\ M(i,j,w-1) \cdot \chi, & \text{if} \quad i = c+r; \\ \\ 0, & \text{if} \quad i > c+r, \end{cases} \tag{9}$$

where:

$$\begin{aligned} M(i,j,w) \quad = \quad & \sum_{s_1=min(w,1)}^{min(w,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{m-s_1} \\ & \sum_{s_2=min(w-s_1,1)}^{min(w-s_1,s_1)} \Big[C_{s_1}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} \\ & C_{s_2}^{w-s_1-s_2} P_{z,3}^{w-s_1-s_2} \\ & (1-P_{z,3})^{s_2-(w-s_1-s_2)} \Big] \Big\}, \end{aligned} \tag{10}$$

and χ is the indicator function

$$\chi = \begin{cases} 1, & if \quad w-s_1-s_2 \leq s_2; \\ \\ 0, & if \quad w-s_1-s_2 > s_2. \end{cases} \tag{11}$$

Region 4:

For $c \leq i \leq c+r$, $0 \leq j < c$, and $i+1 \geq j$, all servers are busy at the beginning of T, and $c-j$ servers are idle at the end of T. In this region, the transition probabilities are defined by:

$$p_{ij} = \begin{cases} M(i,j,w) \cdot \chi, & \text{if} \quad i < c+r; \\ \\ M(i,j,w-1) \cdot \chi, & \text{if} \quad i = c+r; \\ \\ 0, & \text{if} \quad i > c+r, \end{cases} \tag{12}$$

where:

$$M(i,j,w) = \begin{cases} M_1(i,j,w), & if \quad s_1 \geq i-c+1; \\ \\ M_2(i,j,w), & if \quad n_1 < i-c+1, \end{cases} \tag{13}$$

$$\begin{aligned} M_1(i,j,w) \quad = \quad & \sum_{s_1=c-j}^{min(w,c)} \Big\{ C_m^{s_1} P_x^{s_1} (1-P_x)^{c-s_1} \\ & \sum_{s_2=min(w-s_1,c-j)}^{min(w-s_1,i-c+1)} \Big[C_{s_1}^{s_2} P_{z,2}^{s_2} \\ & (1-P_{z,2})^{i-c+1-s_2} \Big] \Big\}, \end{aligned} \tag{14}$$

$$\begin{aligned} M_2(i,j,w) \quad = \quad & \sum_{s_1=c-j}^{min(w,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{c-s_1} \\ & \sum_{s_2=min(w-s_1,c-j)}^{min(w-s_1,s_1)} \Big[C_{s_1}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} \\ & C_{s_2}^{w-s_1-s_2} P_{z,3}^{w-s_1-s_2} \\ & (1-P_{z,3})^{s_2-(w-s_1-s_2)} \Big] \Big\}. \end{aligned} \tag{15}$$

3.2.2. Discussion

Between two successive task arrivals (i.e. during T) there may be $\omega = i+1-j$ departures from the system. In the region 2 there will be at most one departure from each server of the system, however, in regions 3 and 4, there can be more than one departure from any given server.

Let us now examine in detail the equations of the transition probabilities for regions 2, 3 and 4 that we propose and we compare these equations with those proposed by Khazaei et al. (2012) and Chang et al. (2014).

- **Region 2** ($i+1 \leq c$, $0 \leq j \leq c$, and $i+1 \geq j$): In this region, the n^{th} arrival finds the waiting queue empty and i servers busy (all i tasks in the system are in service), then as $i \leq c-1 < c$, it will find an idle server and will immediately enter into service. However, the $(n+1)^{th}$ arrival finds exactly j on arrival, i.e. there are $i+1-j$ tasks leave the system between two successive arrivals. We distinguish two cases:
  - Case 1: If among these i busy servers, $i-j$ of them completed the services of tasks and, the service of the n^{th} arrival is also completed before the $(n+1)^{th}$ arrival, then the probability of having $i+1-j$ departures from the system in this case is equal to:

$$C_i^{i-j} P_x^{i-j} (1-P_x)^j P_y.$$

  - Case 2: If among these i busy servers, $i+1-j$ of them completed the services of tasks and, the service of the n^{th} arrival

is not completed, then the probability of having $i+1-j$ departures from the system in this case is equal to:

$$C_i^{i-j+1} P_x^{i-j+1} (1-P_x)^{j-1} (1-P_y).$$

In the two cases cited above, there are $i+1-j$ tasks that leave the system between two successive arrivals, thus the probability of having $i+1-j$ departures from the system during T is equal to:

$$C_i^{i-j} P_x^{i-j} (1-P_x)^j P_y + C_i^{i-j+1} P_x^{i-j+1} (1-P_x)^{j-1} (1-P_y). \quad (16)$$

This equation was proposed by Khazaei et al. (2012) and Chang et al. (2014), but when we have $i+1=j$, i.e. there is the same number of tasks in the system at the beginning and at the end of T, Eq. 16 can not be used to compute the conditional probability $P(X_{n+1} = i+1 | X_n = i)$. This probability can be obtained in the following manner:

$$(1-P_x)^j.$$

That is presents the probability of no having a departure between two successive arrivals. Consequently, the transition probabilities for the region 2 defined in this paper is given by Eq. 8.

- **Region 3** ($c \leq i \leq c+r$, $c \leq j \leq c+r$, and $i+1 \geq j$):
  The n^{th} arrival finds all c servers are busy and $i-c$ tasks in the waiting queue. If the number of tasks in the system is strictly less than $c+r$ (system capacity), then the n^{th} arrival will be allowed entry. Therefore, there should be $i-c+1$ tasks in the waiting queue at the beginning of T.

  In this region, there can be more than one departure from any given server. Among the c servers, s_1 of them complete at least one service during T. Note that all these completed services must have already been in service at the beginning of T. Among these s_1 servers, s_2 of them will complete a second service during T. As we assumed that each single server completes no more than three services of tasks in T, then the remaining $\omega - s_1 - s_2$ services must be completed in T. These services will complete by a subset of these s_2 servers. The number of servers in the subset is $\omega - s_1 - s_2$, that is, there are $\omega - s_1 - s_2$ servers, each of which completes exactly three services in T and, the number of servers that are still busy processing the third service should be set to $s_2 - (\omega - s_1 - s_2)$. This number is set to s_2 in the equation of the transition probabilities proposed by Khazaei et al. (2012), with the probability $(1-P_{z,3})^{s_2}$. According to this authors this equation is equal to:

$$\begin{aligned} p_{ij} = & \sum_{s_1=min(w,1)}^{min(w,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{m-s_1} \\ & \sum_{s_2=min(w-s_1,1)}^{min(w-s_1,s_1)} \Big[C_{s_1}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} \\ & C_{s_2}^{w-s_1-s_2} P_{z,3}^{w-s_1-s_2} \\ & (1-P_{z,3})^{s_2} \Big] \Big\}. \end{aligned} \quad (17)$$

However, it is impossible to find exactly j tasks in the system at the end of T in this equation because the authors considered the number of servers that are still busy processing the third service is set to s_2. Consequently, Chang et al. (2014) proposed a new equation for this region, defined as:

$$\begin{aligned} p_{ij} = & \sum_{s_1=min(w,1)}^{min(w,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{m-s_1} \\ & \sum_{s_2=min(w-s_1,1)}^{min(w-s_1,s_1)} \Big[C_{s_1}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} \\ & C_{s_2}^{w-s_1-s_2} P_{z,3}^{w-s_1-s_2} \\ & (1-P_{z,3})^{s_2-(w-s_1-s_2)} \Big] \Big\}. \end{aligned} \quad (18)$$

In this equation, the remaining $\omega - s_1 - s_2$ tasks that will must leave the system, can be serviced by the subset of s_2 servers that have completed two services. When the number of s_2 servers is small, then it can happen that the number of remaining tasks exceed this number, therefore, in order that the assumption to have "each single server completes no more than three services of tasks in T" can verify, an indicator function must be used and it is equal to 1 if the number of remaining tasks in the system is less than the number of s_2 servers, is equal to 0, otherwise (Eq. 11).

Also, in this region, if the n^{th} arrival finds the system full ($i = c+r$), then it will be lost. Therefore, there will be $i-j$ tasks that will leave the system between the n^{th} arrival and the $(n+1)^{th}$ arrival instead of $i+1-j$ tasks. Accounting for the above analysis we propose in this paper a new approximate formula for the computation of the element $p_{c+r,j}$, for all j. Consequently, a new accurate approximation of the transition probabilities for the region 3 is defined in this paper by Eq. 9. This enables to enhance previous approximations.

- **Region 4** ($c \leq i \leq c+r$, $0 \leq j < c$, and $i+1 \geq j$):

In this region, at the beginning of T there are $i-c+1$ tasks in the waiting queue and all c servers are busy, while at the end of T, the waiting queue is empty and there are $c-j$ servers are idle.

The equation of the transition probabilities proposed by Khazaei et al. (2012) for this region is given by:

$$p_{ij} = \sum_{s_1=c-j}^{min(\omega,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{c-s_1} \sum_{s_2=min(\omega-s_1,c-j)}^{min(w-s_1,s_1)} \Big[C_{s_1}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} C_{s_2}^{\omega-s_1-s_2} P_{z,3}^{w-s_1-s_2} (1-P_{z,3})^{s_2} \Big] \Big\}. \quad (19)$$

In this equation, Khazaei et al. (2012) had not taken into account the number of tasks in the waiting queue at the beginning of T. However, in this region we distinguish two cases:

- case 1: If $s_1 \geq i-c+1$, there are at most $i-c+1$ tasks that begin service at a subset of the s_1 servers that have completed one services in T and, there are $s_1-(i-c+1)$ servers remain idle because the waiting queue is empty.
- case 2: If $s_1 < i-c+1$, there are s_1 tasks begin service at a subset of the s_1 servers that have completed one services in T.

Chang et al. (2014) proposed a new equation for the transition probabilities for this region:

$$p_{ij} = \frac{1}{(1-P_{z,3})^{c-j}} \sum_{s_1=c-j}^{min(\omega,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{c-s_1} \sum_{s_2=min(\omega-s_1,c-j)}^{min(\omega-s_1,s_1)} \Big[C_{min(i-c+1,s_1)}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} C_{min(max(i-c+1-s_1,0),s_2)}^{\omega-s_1-s_2} P_{z,3}^{\omega-s_1-s_2} (1-P_{z,3})^{s_2-max(i-c+1-s_1-s_2,0)} \Big] \Big\}. \quad (20)$$

In this equation, the authors considered the number of servers that are still busy processing the third service is equal to $s_2 - max(i-c+1-s_1-s_2,0)$, because they assumed that all servers that enter in the idle state during T they do not complete the service, even though the servers are idle. In order to correctly account for the $c-j$ idle servers at the end of T, they multiplied the equation of the transition probabilities for this region by $\frac{1}{(1-P_{z,3})^{c-j}}$.

In the case where $s_1 \geq i-c+1$, the Eq. 20 will be equal to:

$$p_{ij} = \frac{1}{(1-P_{z,3})^{c-j}} \sum_{s_1=c-j}^{min(\omega,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{c-s_1} \sum_{s_2=min(\omega-s_1,c-j)}^{min(\omega-s_1,s_1)} \Big[C_{i-c+1}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} C_{min(max(i-c+1-s_1,0),s_2)}^{\omega-s_1-s_2} P_{z,3}^{\omega-s_1-s_2} (1-P_{z,3})^{s_2-max(i-c+1-s_1-s_2,0)} \Big] \Big\}, \quad (21)$$

therefore, the $max(i-c+1-s_1,0)=0$, $min(max(i-c+1-s_1,0),s_2)=0$ and $max(i-c+1-s_1-s_2,0)=0$. As the waiting queue is empty (all $i-c+1$ tasks in waiting queue enter service because $s_1 \geq i-c+1$), then the number of servers that will complete three services during T is equal to 0 (i.e. $\omega-s_1-s_2$=0). Therefore, the number of servers that will be idle at the end of T is equal to s_2, i.e. $c-j=s_2$. Thus, Eq. 21 will be equal to:

$$p_{ij} = \sum_{s_1=c-j}^{min(\omega,c)} \Big\{ C_c^{s_1} P_x^{s_1} (1-P_x)^{c-s_1} \sum_{s_2=min(\omega-s_1,c-j)}^{min(\omega-s_1,s_1)} \Big[C_{i-c+1}^{s_2} P_{z,2}^{s_2} (1-P_{z,2})^{s_1-s_2} \Big] \Big\}. \quad (22)$$

In this equation, if the number of servers that complete the second service is equal to $i-c+1$, then at the end of T, the number of servers remain busy processing the second service is equal to $i-c+1-s_2$, while in the Eq. 22, this number is equal to s_1-s_2, consequently, the number of tasks in the system at the end of T exceeds j. Therefore, the coefficient $\frac{1}{(1-P_{z,3})^{c-j}}$ is not valid when $s_1 \geq i-c+1$, but it is valid just when $s_1 < i-c+1$.

Accounting for the above analysis for this region we study the two cases cited above separately and we propose in this paper a new approximate formula for each case (Eq. 14, Eq. 15 resp.). Consequently, a new accurate approximation of the transition probabilities for the region 4 is defined in this paper by Eq. 12.

4. CONCLUSION

In this paper we used the same analytical model, $M/G/c/c+r$ queuing system, proposed by Khazaei et al. (2012) for modeling the cloud computing center. However, we focused on the one-step transition matrix of this model because we constated

that the analysis of this model differs from an author to another (see the analysis of Khazaei et al. (2012), Chang et al. (2014)). So, we proposed a new approximation formulas for the one-step state-transition probabilities. Compared to the existing approximation formulas for the one-step state-transition probabilities, our approximation formulas improved the computation of the transition probabilities of the $M/G/c/c+r$ transition matrix.

Our future work, will be consecrate to the presentation of the numerical results. We also plan to extend the model $M/G/c/c+r$ queuing system with one-step transition matrix by considering the services with different priorities and batch-task arrivals.

REFERENCES

Mell P., and Grance T. (2009) *The NIST Definition of Cloud Computing*. Available from http://www.cloudbook.net/resources/stories/the-nist-definition-of-cloud-computing

Furht B. (2010) *Cloud Computing Fundamentals*. In: Furht B. and Escalante A. (Eds.). *Handbook of Cloud Computing*. Springer. 3-19.

Wang L., Von Laszewski G., Younge A. , He X., Kunze M., Tao J., and Fu C. (2010) *Cloud Computing: A Perspective Study*. New Generation Computing, vol. 28. 137-146.

Khazaei H., Misic J. , and Vojislav B. M. (2012) *Performance analysis of cloud computing centers using* $M/G/m/m+r$ *queueing systems*. IEEE Transactions on Parallel and Distributed Systems, vol. 23. 936-943.

Nozaki S.A. and Ross S.M. (1978) *Approximations in Finite-Capacity Multi-Server Queues with Poisson Arrivals*. Applied Probability, vol. 15, 826-834.

Yao D.D. (1985) *Refining the diffusion approximation for the M/G/m queue*. Operations Research, vol. 33, 1266-1277.

Kimura T. (1983) *Diffusion Approximation for an* $M/G/m$ *Queue*. Operations Research, vol. 31, 304-321.

Tijms H.C., Hoorn M.H.V. and, Federgru A. (1981) *Approximations for the Steady-State Probabilities in the* $M/G/c$ *Queue*. Advances in Applied Probability, vol. 13, 186-206.

Boxma O.J., Cohen J.W. , and, Huffel N. (1979) *Approximations of the Mean Waiting Time in an M/G/s Queueing System*. Operations Research, vol. 27, 1115-1127.

Amazon (2010) *Amazon Elastic Compute Cloud, User Guide, API Version ed., Amazon Web Service LLC or Its Affiliate*. Available from http://aws.amazon.com/documentation/ec2.

Zhang Q., Hellerstein J. and Boutaba R. (2011) *Characterizing task usage shapes in Google's compute clusters*, LADIS Workshop.

Reiss C., Tumanov A., Ganger G. R., Katz R., and Kozuch M. (2012) *Heterogeneity and Dynamicity of Clouds at Scale: Google Trace Analysis*. Proc. ACM Symposium on Cloud Computing.

Xiong K. and Perros H. (2009) *Service Performance and Analysis in Cloud Computing*. proceedings of the 2009 Congress on Services. Los Angeles, CA, 6-10 July 2009. IEEE. 693-700.

Chang X., Wang B., Muppala J. K., Liu J. (2014) *Modeling active virtual machines on IaaS clouds using an M/G/m/m+K queue*. IEEE Transactions on Services Computing, vol. PP. 1-1.

Yang B., Tan F., Dai Y. and Guo S. (2009) *Performance Evaluation of Cloud Service Considering Fault Recovery*. proceedings First International Conference, CloudCom 2009. Beijing, China, 1-4 December 2009. Springer Berlin Heidelberg. 571-576

Ani Brown Mary N. and Saravanan K. (2013) *Performance Factors of Cloud Computing Data Centers Using* $[(M/G/1) : (\infty/GD\ MODEL)]$ *Queuing Systems*. International Journal of Grid Computing $\&$ Applications IJGCA, Vol 4, 1-9.

Kimura T. (1996a) *A Transform-Free Approximation for the Finite Capacity* $M/G/s$ *Queue*. Operations Research, vol. 44, 984-988.

Kimura T. (1996b) *Optimal Buffer Design of an* $M/G/s$ *Queue with Finite Capacity*. Comm. in Statistics Stochastic Models, vol. 12, 165-180.

Smith J.M. (2003) *M/G/c/K Blocking Probability Models and System Performance*. Performance Evaluation, vol. 52, 237-267.

Kleinrock L. (1975) *Queueing Systems* vol. 1, Theory. Wiley-Interscience.

Takagi H. (1991) *Queueing Analysis*. vol. 1, Vacation and Priority Systems. North-Holland.