International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 3 (2015) pp. 6713-6729

http://www.ripublication.com

A Maximum Entropy Based Analytical Model to Evaluate Performances of Servers Management Schemes in a Cloud Computing Center

Mohamed Ben El Aattar[1], Abdellah Zaaloul[1], Abdelkrim Haqiq[1]

[1]Computer, Networks, Mobility and Modeling laboratory,
FST Hassan 1st University, Settat, Morocco .
benelaattar@gmail.com

Abstract

Cloud Computing intends to share resources for computation, storage, information and knowledge for users. Quantifying and characterizing performance measures in Cloud Computing require appropriate models. To model networks and estimate its QoS parameters, the queuing theory models are a powerful tool. In this paper, we suggest using the cost-effective server management system PSS scheme (Partial sharing scheme) to manage and differentiate services in Cloud Computing center. This scheme has proved its efficiency in the bandwidth management in WiMAX networks. We also suggest comparing this scheme with the classical CPS scheme (complete partition scheme). For the CPS scheme, the Cloud Computing center is modeled as a queuing network where the total capacity of the center is divided in order to serve three classes in a simple and separated manner. In the PSS scheme, the Cloud Computing center is modeled as queuing network where the servers are shared by three classes; the high priority class is independent from other traffic flows, the mid and the low priority tasks interact between them in order to share the available capacity of service. Both schemes take into consideration generalized exponential GE-type queuing systems which also take into consideration batch arrivals. The Maximum Entropy based analytical model is used for the performance evaluation. The key performance estimated indicators are: the blocking probability, utilization of servers, the average number of tasks and the response time in the system.

Keywords: Cloud Computing, Performance Analysis, Queuing Theory Model, Generalized Exponential, Maximum Entropy.

Introduction

Cloud Computing has been used to define applications delivered as services over the

Internet as well as the hardware and middleware which reside in data centers. The latter are used to provide those services (Armbrust et al., 2010) [1]. In this technology, three main services are provided by the Cloud Computing architecture, according to the needs of IT (Information Technology) customers (Saurabh et al., 2013) [2].

Firstly, Cusumano (2010) [3] has defined Software as a Service (SaaS) as a provider of access to complete applications as a service. Secondly, Ciurana (2009) [4] has also defined Platform as a Service (PaaS) as a provider of a platform for developing other applications on top of IT. Finally, Infrastructure as a Service (IaaS) provides an environment for deploying, running and managing virtual machines and storage. Technically, IaaS offers incremental scalability (scale up and down) of computing resources and on-demand storage (Buyya et al., 2009) [5].

Public cloud refers to situations where the cloud, and in particular infrastructure as-a-service, is made available publicly to individuals and organizations and is charged using metered billing (i.e. pay for what you use). Public cloud allows different end users to share hardware resources and network infrastructure and examples include Amazon and Rackspace.

The private cloud is targeted at large organizations, and generally provides more flexible billing models as well as the ability for these users to define secure zones within which only their company has access to the hardware and network (e.g. Rackspace private cloud, IBM).

Hybrid clouds often refer to situations where organizations are making use of both public and private cloud for their infrastructures. The concept of Cloud Computing also assumes that resources are theoretically available on demand, whereby users of the cloud can scale their cloud infrastructure immediately when the need arises, i.e. during a traffic surge.

There are a number of areas where results from performance engineering of software systems could benefit from the area of cloud computing. Examples include SaaS performance design,autonomics, performance monitoring ,resource utilization and data analysis.

Due to benefits offered by Cloud computing, Quality of Service (QoS) is a broad topic in this technology, and some important quality measures in cloud's users prospective have to be evaluated (Oumellal et al., 2014) [6]. Quantifying and characterizing such performance measures requires appropriate models. Kleinrock (1975) [7] has shown that to model networks and estimate its QoS parameters, the queuing theory models are a classic and powerful tool.

However, and due to the diversity of services in cloud computing, different tasks can be treated in a cloud center and queuing modeling of cloud services has to take into consideration the type of tasks that is studied and to adopt a differentiation policy between tasks which share the resources. On the other hand, an aspect that has not received enough interest in the research is the possibility of batched arrivals.

Thus, in this paper, we suggest using the PSS servers management schemes that proved its efficiency in the bandwidth management in WiMAX networks, compared to the classical CPS scheme (Yue et al., 2008) [8], and we study this mechanism using an analytical queuing model. This study meets the following requirements:

- Taking into account the priority tasks of the new arrivals in the Cloud Computing center;
- Taking into account the possibility of batch arrivals.
- Diversification of management schemes of servers in the Cloud Computing center.

In the proposed analytical model, an open queuing network model (QNM) is used, which consists of interacting multiclass generalized exponential (GE)-type queuing and delay systems with multiple servers and finite capacities. Maximum entropy (ME) analytic solutions are derived, subject to appropriate GE-type queuing and delay theoretic mean value constraints and expressions. The model states and the blocking probability distributions are determined by closed-form expressions.

These mathematical assumptions make the proposed model more convenient for the cloud environment nature, and it confers to the model the quality of being close to reality and the quality of scalability. Moreover, and due to the introduction of the finite capacity, the system may experience blocking of task requests.

The rest of the paper is organized as follows: in Section 2 we give a brief overview of related works on cloud performance evaluation and performance characterization of queuing systems. Management schemes of servers are introduced in Section 3, analytical model in detail in Section 4 and we proceed to performance measures in Section 5. The numerical results are presented in Section 6. Concluding remarks follow in Section 7.

2. Related Works

Cloud Computing provides user with a complete software environment. It provides resources such as computing power, bandwidth and storage capacity. It has engrossed considerable investigating attention, but only a diminutive portion of the work done so far has addressed performance issues, and rigorous analytical approach has been adopted by only a handful among these, particularly that adopting queuing theory models.

Yang et al., (2009) [9] proposed a fault recovery system scheduling for cloud services and analyzed the system as an open queue problem using a M/M/m/m+r queuing system ; the results showed that addition of fault recovery increases average response time.

Xiong and Perros (2009) [10] modeled a cloud center as the classic open network from which the distribution of response time is obtained, assuming that both inter-arrival and service times are exponential. Using the distribution of response time, the relationship among the maximal number of tasks, the minimal service resources and the highest level of services was found.

The cases of queuing system with generally distributed service time were the subject of most theoretical analyses. However, in these cases the steady state probability, the distributions of response time and the queue length have yet to be solved exactly. Consequently, researchers have developed many methods for approximating its solution.

Yao (1985) [11] defined a diffusion approximation model for a M/G/m queue as solving the equations incorporating some known results of the queue into the model. Their numerical studies indicate that the refined model provides significantly improved performance.

However, in many practical cloud service situations, the time dependent arrivals of tasks are to be considered in order to have accurate prediction of the performance measure of the Cloud Computing (Satyanarayana et al., 2013)[12]. Besides, a number of measurement studies have revealed that the traffic generated by many real world applications exhibit a high degree of burstness and poses correlation in the number of request arrivals (Sai et al., 2011)[13]. Therefore, the traditional Poisson process models cannot capture the burst nature of request arrival process. Hence, it is necessary to develop queuing model taking into account the characteristics of the type of modeled traffic.

The generalization of the M/G/m/m+r model given by (Khazaei et al., 2012) [14], has been improved by (Oumellal et al., 2014) [6] using an MMPP model for the task of arriving in the center. Thus, the diversity and burstness of user requests was made in this paper.

Another aspect of arrival that can be studied in a Cloud Computing center is the possibility of batch arrivals. At the best of our knowledge, there is no study that uses a model taking into account this aspect. Modeling arrival process in a queue system using Generalized Exponential distribution (GE) is a useful tool, as it is a generalization that can take into consideration the batch arrivals, and several features of the models in which customers arrive singly are maintained in this generalization.

There are few works using GE distribution in network modeling. Kouvatsos, in a series of publications (Kouvatsos, 1988; Kouvatsos, 1988; Kouvatsos, 1986; Kouvatsos, 1986; Kouvatsos, 1994) [15, 16, 17, 18, 19], laid the foundation of a new analytical frame work which can be used to derive minimally biased approximations of performance distributions for queues, and queuing networks with General / Generalized exponential traffic with infinite buffers, subject to mean rate and squared coefficient of variance (SCV) of inter arrival and service time distribution. The minimal biased approximations of performance distributions for finite queue with General/Generalized exponential distribution with finite/no finite buffer derived are hypothetically deduced from queues with infinite buffer (Kouvatsos, 1986; Awan, 2006; Kouvatsos, 2003) [20, 18, 21].

Authors (Kouvatsos et al., 2003) [22] have extended the model proposed by Hu and Kleinrock (1997) [23], by relaxing the assumption of Poisson arrival process using a GE/G/1/K queue to model finite input buffer. This paper proposes an analytical model for predicting the average worm latency in the hypercube with deterministic routing, wormhole switching and finite size input buffers.

In this work, we model a Cloud Computing centre management scheme as a queuing network model (QNM) that consists of three interacting multiclass generalized exponential (GE)-type queuing and delay systems with multiple servers and finite capacities.

3. The Management Schemes of Servers

Management of servers in a Cloud Computing center is a complex issue due to a broad range of traffic characteristics and performance constraints. Continuous servers control ensures that differentiation between traffic classes is treated in the proper way. In literature, there are several servers management schemes that are used in telecommunication networks. In this section we present two among the best-known management schemes: CPS and PSS.

3.1 Complete partitioning scheme (CPS)

CPS divides the total cell capacity to simultaneously serve multimedia traffic. As a consequence, the transmission of tasks with three classes: high priority, mid priority and low priority, can be studied separately (Fig 1).

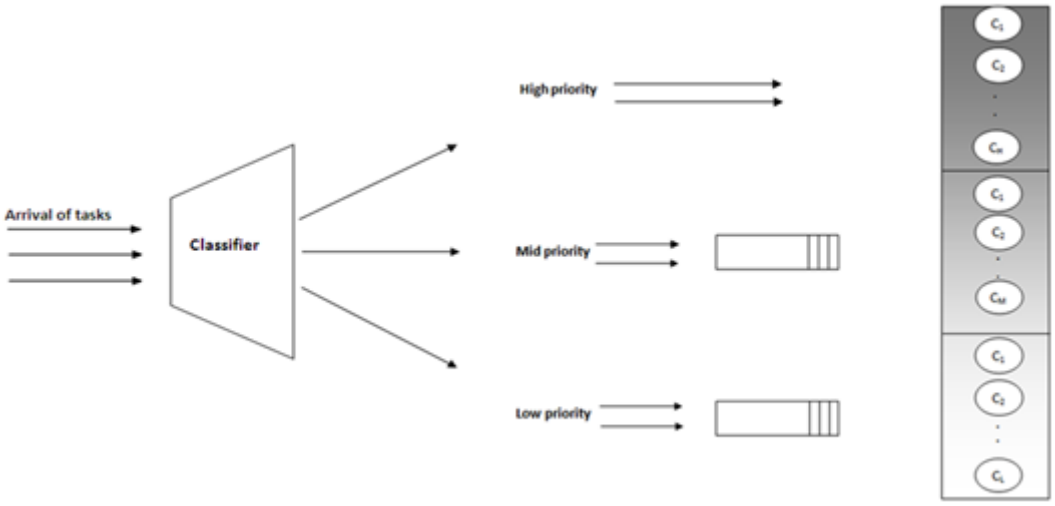

Fig. 1. An open QNM of a Cloud Computing system under CPS

3.2 Partial sharing scheme (PSS)

PSS is based on a partial sharing of service where the high priority class does not tolerate any delay, so it has the highest priority and it is independent of other traffic flows, followed by that of mid priority partition and low priority partition. Low priority tasks will get some available servers from the mid priority class, but it will give them back to the mid priority class when the new incoming mid priority tasks arrive (fig 2).

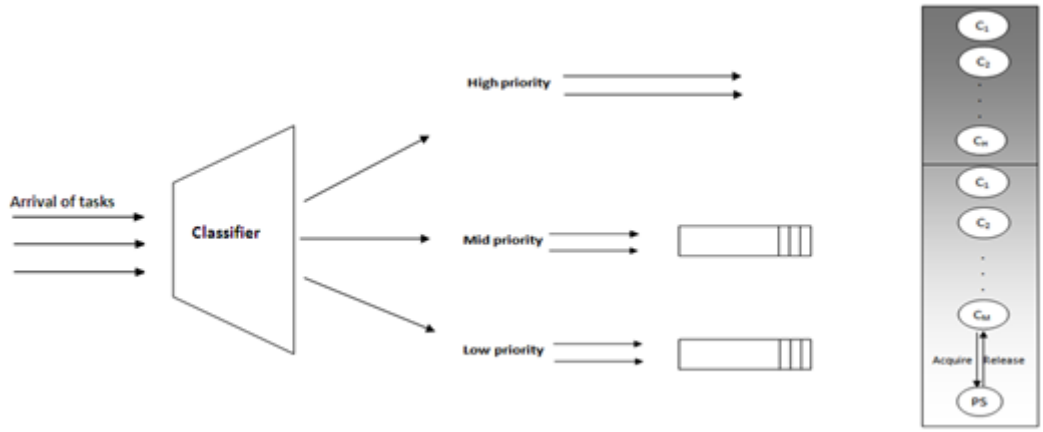

Fig. 2. An open QNM of a Cloud Computing system under PSS

4. Analytical Model for the Management Schemes of Servers

4.1 GE-Type Distribution

The case where simultaneous arrivals are allowed is a common case for traffic arriving in a Cloud Computing center. To give a model that has the quality of being close to reality, it is necessary to take into account this constraint.

The Generalized Exponential distribution is a suitable tool to model batch arrivals in a Cloud Computing center.

The GE-type distribution is of the form (Awan, 2006; Kouvatsos, 2003) [24]:

$$F(t) = P(x \le t) = 1 - \frac{2}{C^2+1}\exp\left[-\frac{2\nu t}{C^2+1}\right], \quad t \ge 0 \qquad (1)$$

where $1/\nu$ is the mean and C^2 is the squared coefficient of variation (SCV), used to approximate general distributions with known first two moments.

Although the GE distribution is improper for $C^2 < 1$, it is still a useful and versatile tool in the field of system modeling.

The GE distribution has a counting compound Poisson process (CPP) with geometrically distributed batch sizes with mean $\frac{C^2+1}{2}$. It may be meaningfully used to model the inter-arrival times of bursty multiple class.

4.2 Maximum Entropy Formalism

The ME formalism provides an analytical solution approximated to that of queuing system and networks with stochastic and operational analysis subject to well-defined constraints.

For each class $i(i = 1,2,...,R, R > 1)$, let $\left(1/\lambda_i, C_{ai}^2\right)$ and $\left(1/\mu_i, C_{si}^2\right)$ be the mean and SCV of the inter-arrival and service time distributions, respectively. Note that μ_i

$(i=1,2,...,R)$ under the PS (Process Sharing policy) rule is defined subject to discriminatory weights. Moreover, for either a GE/GE/c/N/FCFS/CBS (under Complete Buffer Sharing policy) queuing system or a GE/GE/1/N/PS delay system, let at any given time:

- **S**=(n_1,n_2,…, n_R,r) be a joint queue state , where $\sum_{i=1}^{R} n_i \le N$, n_i (i=*1,…,N*), be the number of class i task in the queue (waiting and/or receiving service) with r($1 \le r \le R$) denote the class of the current task in service (n.b., for an idle queue $\mathbf{S} \equiv \mathbf{0}$ with r=0).
- n_i(**S**)= the number of i tasks present in either the GE/GE/1/N/PS or the GE/GE/c/N/FCFS/CBS system.
- π_i be the blocking probability that an arrival of class i will find the system at full capacity.
- Q be the set of all feasible states of S.

For each state S, S∈Q, and class $i,(i=1,2,...,R)$, the following auxiliary functions are defined:

$$s_i(\mathrm{S}) = \begin{cases} 1 \; if \;\; r = i. \\ 0, otherwise, \; GE/GE/1/N/PS \end{cases}, \forall i.$$

$$s_{ik}(\mathrm{S}) = \begin{cases} 1 \; if \;\; n_i \ge k \;\; and \;\; k \le c - \sum_{\substack{j=1 \\ j \ne i}}^{R} n_j . \\ 0, otherwise, \; GE/GE/c/N/FCFS/CBS \end{cases}$$

$$f_i(\mathrm{S}) = \begin{cases} 1, if \sum_{j=1}^{R} n_j(\mathrm{S}) = N, and \; s_i(\mathrm{S}) = 1 \\ 0, otherwise, \; GE/GE/1/N/PS \end{cases}$$

$$f_{ik}(\mathrm{S}) = \begin{cases} 1, if \sum_{j=1}^{R} n_j(\mathrm{S}) = N, and \; s_{ik}(\mathrm{S}) = 1 \\ 0, otherwise, \; GE/GE/c/N/FCFS/CBS \end{cases}$$

Note that the constraints of the server state are $s_i(\mathrm{S})$ or $s_{ik}(\mathrm{S})$, which is related to the service utilization. On the other hand, the constraints of the queue capacity are designed as $f_i(\mathrm{S})$ or $f_{ik}(\mathrm{S})$.

Suppose that the following mean value constraints about the state probability P(S) are known to exist:

- Normalization

$$\sum_{S\in Q} P(S) = 1 \qquad (2)$$

- Service utilization

$$\begin{cases} \sum_{S\in Q} s_i(\mathrm{S})P(S) = U_i, \quad 0 < U_i < 1, \\ \qquad i = 1,...,R \qquad GE/GE/1/N/PS \\ \sum_{S\in Q} s_{ik}(\mathrm{S})P(S) = U_{ik}, \quad 0 < U_{ik} < 1, \\ \qquad i = 1,...,R; k = 1,...,c \quad GE/GE/c/N/FCFS/CBS \end{cases} \qquad (3)$$

- Average number in the system{Li ,i=1, 2, ...,R },

$$\sum_{S\in Q} n_i(\mathrm{S})P(S) = L_i, i = 1,2,...,R \qquad (4)$$

$$with \begin{cases} U_i < L_i < N & GE/GE/1/N/PS \\ U_{ik} < L_i < N & GE/GE/c/N/FCFS/CBS \end{cases}$$

- Full buffer state probabilities $\{\phi_i, \phi_{ik}, i=1,...,R; k=1,...,c\}$

$$\begin{cases} \sum_{S\in Q} f_i(\mathrm{S})P(S) = \phi_i, \quad 0 < \phi_i < 1, \\ \qquad i = 1,...,R \qquad GE/GE/1/N/PS \\ \sum_{S\in Q} f_{ik}(\mathrm{S})P(S) = \phi_{ik}, \quad 0 < \phi_{ik} < 1, \\ \qquad i = 1,...,R; k = 1,...,c \quad GE/GE/c/N/FCFS/CBS \end{cases} \qquad (5)$$

satisfying the class flow balance equations, namely

$$\lambda_i(1-\pi_i) = \mu_i U_i, \ i = 1,2,...,R \qquad (6)$$

The form of the ME joint state probability distribution $\{P(S), S \in Q\}$ can be characterized by maximizing the entropy functional $H(P) = -\sum_{S\in Q} P(S)\log P(S)$ subject to prior information expressed by mean value constraints (2)-(5). By employing Lagrange's method of undetermined multipliers, the following solutions are obtained:

$$P(S) = \begin{cases} \frac{1}{Z}\prod_{i=1}^{R} g_i^{s_i(\mathrm{S})} x_i^{n_i(\mathrm{S})} y_i^{f_i(\mathrm{S})}, \forall S \in Q, & GE/GE/1/N/PS \\ \frac{1}{Z}\prod_{i=1}^{R}\left(\prod_{k=1}^{c} g_{ik}^{s_{ik}(\mathrm{S})} x_i^{n_i(\mathrm{S})} y_{ik}^{f_{ik}(\mathrm{S})}\right), \forall S \in Q, & GE/GE/c/N/FCFS/CBS \end{cases} \qquad (7)$$

where Z=1/P(0) is the normalizing constant, $\{g_i, x_i, y_i, i=1,...,R\}$ and $\{g_{ik}, x_i, y_{ik}, i=1,...,R, k=1,...,c\}$ are the Lagrangian coefficients corresponding to constraints (3)-(5) per class, respectively.

4.3 The Aggregate ME Probability Distribution

The aggregate state probabilities $\{P(n), n=1,...,N\}$, are given by:

$$P(n)=\begin{cases}\frac{1}{Z}\left[\sum_{i=1}^{R} g_i\, x_i\, y_i^{f_i(S)}\right] X^m & GE/GE/1/N/PS \\ \frac{1}{Z}\prod_{i=1}^{R}\left[\prod_{k=1}^{c} g_{ik}^{s_{ik}(S)} x_i\, y_{ik}^{f_{ik}(S)}\right] X^m & GE/GE/c/N/FCFS/CBS\end{cases} \quad (8)$$

Where $Z=1/P(0)$, $X=\sum_{i=1}^{R} x_i$ *and* $n=\sum_{i=1}^{R} n_i$, $m=\begin{cases} n-1 & if\ \ n\le N-1 \\ N & if\ \ n=N\end{cases}$,

4.4 The Lagrangian Coefficients

Kouvatsos and I.U. Awan (1998) have shown that the Lagrangian coefficients $\{x_i, g_i, g_{ik}, i=1,...,R; k=1,...,c\}$ can be approximately determined via closed form asymptotic expressions relating to the ME solution of the corresponding GE-type infinite capacity queues at equilibrium.

Moreover, the Lagrangian coefficients $\{y_i, y_{ik},\ i=1,...,R; k=1,...,c\}$ can be determined via the flow balance equation (6) (Xing et al., 2006) [25].

$$x_i=\frac{\beta_i(1-\alpha_i)+\alpha_i\rho_i}{\beta_i(1-\alpha_i\rho_i)+\alpha_i\rho_i} \quad (9)$$

$$\begin{cases} g_i=\dfrac{\alpha_i\beta_i\rho_i}{\beta_i(1-\alpha_i)+\alpha_i\rho_i} & GE/GE/1/N/PS \\ g_{ik}=\dfrac{\alpha_i c\rho_i+(k-1)\beta_i(1-\alpha_i)}{k(\beta_i(1-\alpha_i)+\alpha_i)} & GE/GE/c/N/FCFS/CBS\end{cases} \quad (10)$$

$$y_i=y_{ik}=\frac{\alpha_i\rho_i+\beta_i(1-\alpha_i\rho_i)}{\beta_i(\beta_i(1-\alpha_i)+\alpha_i)} \quad (11)$$

where: $\alpha_i=\frac{2}{C_{a_i}^2+1}$ and $\beta_i=\frac{2}{C_{s_i}^2+1}$ be the GE-type interarrival and service time non-zero stage selection probabilities associated with the GE/GE/1/N/PS and GE/GE/c/N/FCFS/CBS systems.

$$\rho_i = \begin{cases} \lambda_i / \mu_i & GE/GE/1/N/PS \\ \lambda_i / c\mu_i & GE/GE/c/N/FCFS/CBS \end{cases}$$

5. Performance Measures

5.1 Blocking probability

A universal expression for the marginal blocking probabilities $\{\pi_i, i = 1,...,R\}$ of a stable multiple class GE/GE/1/N/PS delay system and GE/GE/c/N/FCFS/CBS queuing systems can be approximated by focusing on a tagged task within an arriving bulk and making use of GE-type probabilistic arguments.

On assumption that the system is in the state v=(n_1, n_2 ,…,n_R) with $\sum_{i=1}^{R} n_i = n$, the number of available buffer spaces is equal to N-n. By focusing on a tagged task within an arriving bulk of class $i\,(i = 1,2,...,R, R > 1)$, the following blocking probability can be clearly determined:

P(a class i tagged task is blocked and its bulk finds the queue in state 0 = (0, 0, …,

$$0))= \begin{cases} \theta_i(0)(1-\alpha_i)^N P(0) & GE/GE/1/N/PS \\ \theta_i^c(0)(1-\alpha_i)^N P(0) & GE/GE/c/N/FCFS/CBS \end{cases} \quad (12)$$

where $\theta_i(0) = \dfrac{\beta_i}{\beta_i(1-\alpha_i)+\alpha_i}$.

P(a class i tagged task is blocked and its bulk finds a queue in state

$$\text{v}=(n_1,n_2,\ldots,n_R), \sum_{i=1}^{R} n_i \le N \;)= \begin{cases} \theta_i(\text{v})(1-\alpha_i)^N P(\text{v}) & GE/GE/1/N/PS \\ \begin{cases} \theta_i^{c-n}(0)(1-\alpha_i)^{N-n} P(\text{v}), 0 < n < c \\ (1-\alpha_i)^{N-n} P(\text{v}), c \le n \le N \end{cases} & GE/GE/c/N/FCFS/CBS \end{cases} \quad (13)$$

where $\theta_i(\text{v}) = \begin{cases} \theta_i(0) & if\ \text{v} = (0,0,...,0) \\ 1 & otherwise \end{cases}$

Therefore, the blocking probabilities $\{\pi_i, i = 1,2,...,R\}$ can be expressed by:

$$\pi_i = \begin{cases} \sum_{n=0}^{N} \theta_i(\text{v})(1-\alpha_i)^{N-n} P(n) & GE/GE/1/N/PS \\ \sum_{n=0}^{N} \theta_i^{l}(0)(1-\alpha_i)^{N-n} P(n) & GE/GE/c/N/FCFS/CBS \end{cases} \quad (14)$$

where $l = \begin{cases} c & if\ n = 0 \\ \max(0, c-n) & if\ 0 < n < c \\ 0 & if\ c \le n \le N \end{cases}$

5.2 Utilization of servers

Low priority tasks will get some available servers from the mid priority class, but it will give them back to the mid priority class when the new incoming mid priority tasks arrive, so the utilization for low priority and mid priority will be: $U = \sum_{n=1}^{N} P(n)$

Therefore:

$$U = \begin{cases} U_{lp} = \dfrac{1}{Z}\sum_{n=1}^{N}\left[\sum_{i=1}^{R} g_i\, x_i\, y_i^{f_i(\mathrm{S})}\right] X^m & GE/GE/1/N/PS \\ U_{mp} = \dfrac{1}{Z}\sum_{n=1}^{N}\prod_{i=1}^{R}\left[\prod_{k=1}^{c} g_{ik}^{s_{ik}(\mathrm{S})} x_i\, y_{ik}^{f_{ik}(\mathrm{S})}\right] X^m & GE/GE/c/N/FCFS/CBS \end{cases} \tag{15}$$

where $m = \begin{cases} n-1 & if \quad n \le N-1 \\ N & if \quad n = N \end{cases}$

5.3 The average number of tasks in the system

$$L = \sum_{n=0}^{N} nP(n)$$

Therefore:

$$L = \begin{cases} L_{lp} = \dfrac{1}{Z}\sum_{n=0}^{N} n\left[\sum_{i=1}^{R} g_i\, x_i\, y_i^{f_i(\mathrm{S})}\right] X^m & GE/GE/1/N/PS \\ L_{mp} = \dfrac{1}{Z}\sum_{n=0}^{N} n\prod_{i=1}^{R}\left[\prod_{k=1}^{c} g_{ik}^{s_{ik}(\mathrm{S})} x_i\, y_{ik}^{f_{ik}(\mathrm{S})}\right] X^m & GE/GE/c/N/FCFS/CBS \end{cases} \tag{16}$$

where $m = \begin{cases} n-1 & if \quad n \le N-1 \\ N & if \quad n = N \end{cases}$

5.4 The response time of tasks in the system

Using Little's Law, the average waiting time of tasks in the system for each class is given by:

$$W = \frac{L}{\lambda}$$

where $\lambda = \begin{cases} \lambda_{lp} = \lambda_{i_{lp}}(1-\pi_{i_{lp}}) \\ \lambda_{mp} = \lambda_{i_{mp}}(1-\pi_{i_{mp}}) \end{cases}$

Therefore:

$$W = \begin{cases} W_{lp} = \dfrac{L_{lp}}{\lambda_{i_{lp}}(1-\pi_{i_{lp}})} & GE/GE/1/N/PS \\ W_{mp} = \dfrac{L_{mp}}{\lambda_{i_{mp}}(1-\pi_{i_{mp}})} & GE/GE/c/N/FCFS/CBS \end{cases} \quad (17)$$

6. Numerical Results

In this section we give numerical results where the performances: utilization and response time are plotted under different scenarios.

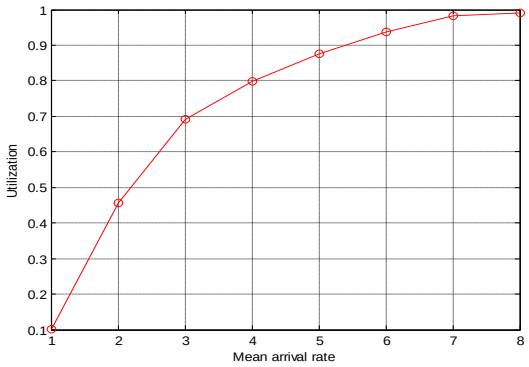

Fig. 3. Effect of traffic variability on the utilization of high priority tasks for a GE/GE/c/c queuing system

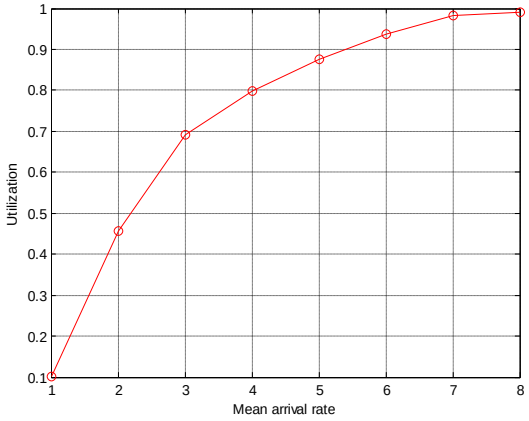

Fig. 4. Effect of traffic variability on the utilization of mid priority tasks for a GE/GE/c/N/FCFS/CBS queuing system

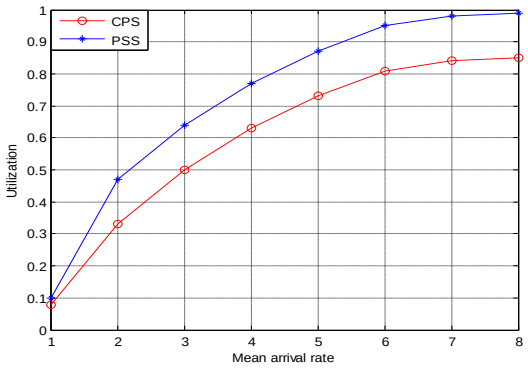

Fig. 5. Effect of traffic variability on the utilization of low priority tasks for a GE/GE/1/N/PS delay system under CPS and PSS

One of the most important goals in Cloud Computing should be the servers utilization that can be achieved. In Figure 3, Figure 4 and Figure 5 the utilizations are given for the three traffic classes as their arrival rates varied.

From Figure 3 and Figure 4 it can be seen that the utilization is an increasing function of the arrival rates of tasks. And when all servers are utilized, the traffic will be queued in the buffer and it will suffer through additional delay and reduced service.

Figure 5 illustrates the use according to the arrival rate of low priority. It can be seen that the PSS scheme improves utilization compared with CPS scheme. The improvement is very clear when the average arrival rate is becoming higher. This occurs since, in PSS scheme, some of low priority tasks can use some servers assigned to mid priority tasks. As a consequence, the response time and lost of packets will be improved for the low priority traffic.

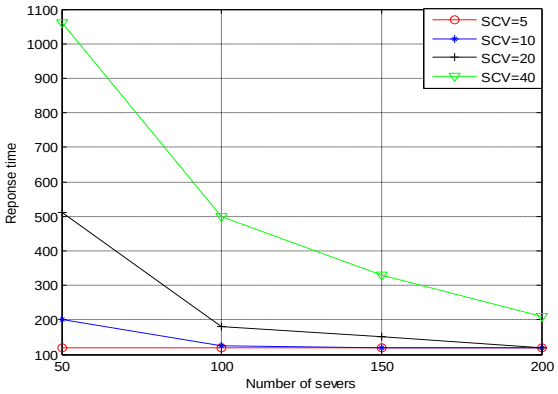

Fig. 6. Effect of number of servers variability on the response time of high priority tasks for a GE/GE/c/c queuing system

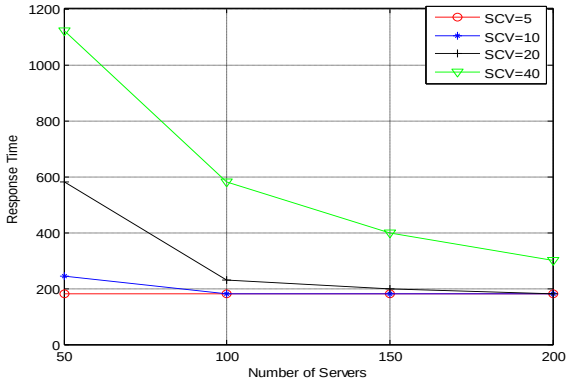

Fig. 7. Effect of number of servers variability on the response time of mid priority tasks for a GE/GE/c/N/FCFS/CBS queuing system

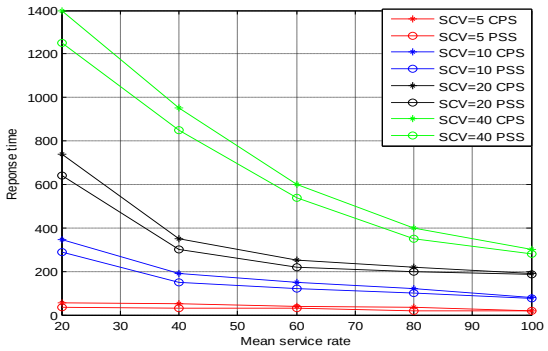

Fig. 8. Effect of number of servers variability on the response time of mid priority tasks for a GE/GE/1/N/PS delay system under CPS and PSS

Figure 6, Figure 7 and figure 8 show the evolution of response time when the capacity of service is varying and with different values of SCV.

The Fig 6 and 7 show the results when the number of servers increases. We can notice that the response time is a decreasing function, which is a logical result; we notice also that the higher the SCV, the greater the response time. Indeed, a high value of SCV is associated to a great dispersion of the service rate even if the mean service rate remains unchanged.

Figure 8 shows the same behavior of the response time versus mean service rate and SCV. Moreover, it can be seen that the PSS scheme provides more efficient results in comparison with CPS scheme for response time for different values of squared coefficient of variation (SCV = 5,10,20,40). This is due to the impact of PSS scheme which gives to low priority tasks more opportunities to receive service when servers dedicated to mid priority are available.

Conclusion

Servers in a Cloud Computing center can receive traffic from multiple sources with different QoS requirements. Hence, designing and understanding servers management schemes are important research issues. In this paper, we studied two servers management schemes that can be used in a Cloud Computing center.

To evaluate the performances of the two studied schemes, and due to the inefficiency and the difficulty when using the traditional approach (Markov chains), we proposed a Maximum Entropy based analytical model. In this model, a GE-type queuing and delay system with a large number of servers and finite capacities is used. Closed-form expressions for the system states and blocking probability distributions are obtained. Typical numerical results confirmed the performance superiority of PSS (compared to CPS).

The proposed analytical model has the property to be more convenient for the cloud environment nature, since a Cloud Computing center can receive high data rates, multiple classes and can have high number of servers.

References

[1] Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R. H., Konwinski, A., Lee, G., Patterson, D. A., Rabkin, A., Stoica, I., and Zaharia, M., 2010, "A view of cloud computing," Communications of the ACM, Vol. 53, No. 4, pp: 50-58. DOI:10.1145/1721654.1721672.

[2] Saurabh Kumar Garg, J., Versteeg, S., and Buyya, R., 2013, "A Framework for Ranking of Cloud Computing Services," ELSEVIER of Future Generation Computer Systems, Vol. 29, No. 4, pp: 1012- 1023. DOI: 10.1016/j.future.2012.06.006.

[3] Cusumano, M., 2010, "Cloud Computing and SaaS as new computing platforms, " Communications of the ACM, Vol. 53, No. 4, pp: 27-29. DOI:10.1145/1721654.1721667.

[4] Ciurana, E. 2009, "Developing with Google App Engine," Apress, Berkeley, CA, USA. ISBN13: 978-1-4302-1831-9, pp: 2.

[5] Buyya, R., Yeo, C., Venugopal, S., Broberg, J., and Brandic, I., 2009, "Cloud Computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems," Elsevier Science,Vol. 25, No. 6, pp: 599-616. DOI:10.1016/j.future.2008.12.001.

[6] Oumellal, F., Hanini, M., and Haqiq, A., 2014, "MMPP/G/m/m+r Queuing System Model to analytically evaluate Cloud Computing Center Performances," British Journal of mathematics and computer science. Vol. 4, Iss. 10, pp: 1301-1317. DOI: 10.9734/BJMCS/2014/8471.

[7] Kleinrock, L., 1975, "Queueing Systems: Theory," vol. 1, Wiley-Interscience. ISBN: 0471491101, pp: 370.

[8] Li, Y., He, J., and Xing, W., 2008, "Bandwidth Management of WiMAX Systems and Performance Modeling," KSII Transactions on Internet and Information Systems Vol. 2, No. 2, pp: 63-81. DOI: 10.3837/tiis.2008.02.001.

[9] Yang, B., Tan, F., Dai, Y., and Guo, S., 2009, "Performance evaluation of cloud service considering fault recovery," In First Int'l Conference on Cloud Computing (CloudCom), pp: 571–576. DOI: 10.5121/ijgca.2013.4101.

[10] Xiong, K., and Perros, H., 2009, "Service performance and analysis in cloud computing," In IEEE 2009 World Conference on Services, pp: 693–700. DOI:10.1109/SERVICES-I.2009.121.

[11] Yao, D., 1985, "Refining the diffusion approximation for the M/G/m queue," Operations Research, vol. 33, No. 6, pp: 1266-1277. DOI: 10.1287/opre.33.6.1266.

[12] Satyanarayana, A., Suresh, V. P., Rama Sundari, M.V., and Sarada, V. P., 2013, "Performance Analysis of Cloud Computing under Non Homogeneous Conditions," International Journal of Advanced Research in Computer Science and Software Engineering. Vol. 3, Iss. 5. pp: 969-974. http://www.ijarcsse.com/docs/papers/Volume_3/5_May2013/V3I5-0371.pdf.

[13] Sai Sowjanya, T., Praveen, D., Satish, K., and Rahmain, A., 2011, "The Queueing Theory in Cloud Computing to Reduce the waiting Time," in IJCSET , Vol 1, Issue 3, pp: 110-112. http://www.ijcset.net/docs/Volumes/volume1issue3/ijcset2011010304.pdf.

[14] Khazaei, H., Misic, J., and Misic, V. 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, No 5, pp: 936-943. DOI:10.1109/tpds.2011.199.

[15] Kouvatsos, D., 1988, "A maximum Entropy analysis of G/G/1 Queue at Equilibrium," Journal of Operational Research Society, Vol. 39, pp: 183-200. DOI: 10.1057/jors.1988.30.

[16] Kouvatsos, D., 1988, "Maximum Entropy Two-Station Cyclic Queues with Multiple General Servers," Acta Informatica 26, Vol. 26, Iss. 3, pp: 241-267. DOI:10.1007/bf00299634.

[17] Kouvatsos, D., 1986, "Maximum Entropy and G/G/1/N Queue," Acta Informatica 23, Vol. 23, Iss. 5, pp: 545-556. DOI:10.1007/bf00288469.

[18] Kouvatsos, D., 1986, "A Maximum Entropy Queue Length Distribution for the G/G/1 Finite Capacity Queue," Sigmetrics, joint International Conference on Computer Performance Modelling, Measurement and Evaluation, Vol. 14, Iss. 1, pp: 224-236. DOI:10.1145/317531.317555.

[19] Kouvatsos, D., 1994, "Entropy maximization and queuing network models," Annals of Operation Research, Vol. 48, Iss. 1, pp: 63-126. DOI: 10.1007/bf02023095.

[20] Awan, I., Yar, A., and Woodward, M. E., 2006, "Analysis of queueing networks with blocking under active queue management scheme," IEEE ICPADS'06, 12th International Conference. Minneapolis, MN, pp: 61-68. DOI: 10.1109/ICPADS.2006.25.

[21] Kouvatsos, D., Awan, I., and Al-Begain, K., 2003, "Performance Modelling of GPRS with bursty Multiclass Traffic Computer and Digital Techniques," IEEE, Vol. 150, Iss. 2, pp: 75-85. DOI: 10.1049/ip-cdt:20030278.

[22] Kouvatsos, D., Assi, S., and Ould-Khaoua, M., 2003, "Performance modelling of hypercubes with deterministic wormhole routing," Proc. 1st International Working Conference Performance Modelling and Evaluation of Heterogeneous Networks (HET-NETs '03), D. D. Kouvatsos (Ed.), pp: 77/1-77/10. ISBN: 0-9540151-3-4.

[23] HU, P., and Kleinrock, L., 1997, "An Analytical model for wormhole routing with finite size input Buffers," 15th International Telegraphic Congress, pp: 549-560. DOI: 10.1.1.48.315&rank=1.

[24] Kouvatsos, D. and Awan, I. U., 1998, "MEM for arbitrary closed queuing networks with RS-blocking and multiple job classes," Annals of Operations Research, Special Issue on Queuing Networks with Blocking, Vol. 79, pp: 231-269. DOI : 10.1023/A:1018922705462.

[25] Xing, W., Li, Y., and Kouvatsos, D., 2006, "An Efficient WiMAX System Design and Structure Protocol for the IRIS Project," The forth international conference on performance modeling and evaluation of Heterogeneous Networks (HET-NET'S 2006), Ilkley, West Yorkshire, Bradford, UK, pp: WP02/1-WP02/1-WP02/7. ISBN: 0-9550624-3-8.