Stochastic bounding models for performance analysis of clouds

Farah Aït-Salaht, Hind Castel-Taleb
INSTITUT TELECOM/Telecom SudParis
SAMOVAR, UMR 5157
9, rue Charles Fourier, 91011 Evry Cedex, France
Email: {farah.ait_salaht, hind.castel}@it-sudparis.eu

***Abstract*—We propose to evaluate the performance of a cloud node (data center) using hysteresis queueing systems and stochastic bound methods. We represent the dynamic behavior of the cloud node by a hysteresis queueing system with forward and backward threshold vectors. The client requests (or jobs) are represented by bulk arrivals entering the buffer, and executed by Virtual Machines (VMs) which are activated and deactivated according to the occupation of the queue, and the threshold sequences. As the system is quite difficult to analyze, we propose to define different bounding systems "less complex" and easier to study. Two approaches are used, one by aggregating the probability distribution of the batch arrivals and another by taking models with the same sequences of forward and backward thresholds. We show the relevance of the proposed bounding systems by presenting some numerical results for the performance measures of a data center.**

Keywords*—*Cloud performance; Stochastic bounds; Markov Chains.

I. Introduction

Cloud computing is a novel virtualized distributed technology in which different computing resources are made accessible over the internet to remote users in an on-demand fashion. Virtualization plays a key role in the success of cloud computing because it simplifies the delivery of the services by providing a platform for resources in a scalable manner. With virtualization, service providers can ensure isolation of multiple user workloads and provision resource in a cost-effective manner by consolidating Virtual Machines (VMs) onto fewer physical resources when system load is low, and quickly scale up workloads to more physical resources when system load is high. In [1], they study the right ratio of VM instances to physical processors that optimizes the workload's performance given a workload and a set of physical computing resources.

Performance evaluation of cloud centers is an important research task which becomes difficult due to the dynamic nature of cloud environments and diversity of user requests. Then, it is not surprising that in the recent area of cloud computing, only a portion of research results has been devoted to performance evaluation.

In [2], they propose an analytic approach based on multi-level interactive stochastic sub-models in order to evaluate the performance of a cloud system. In [3], they model the PM (Physical Machine) with m VMs using the $M^{[x]}/G/m/m+r$ queue. They derive exact formulas for performance measures as blocking probability and mean waiting time of tasks. In [4], they consider a cloud center with a number of physical machines that are allocated to users in the order of task arrivals. Physical Machines (PMs) are considered with high degree of virtualization, and are categorized into three server pools: hot, warm, and cold. Authors implement the sub-models using interactive Continuous Time Markov Chains (CTMCs). In [5], they use a multiserver queueing model with queue dependent heterogeneous servers in order to evaluate the performance of a cloud system.

We propose in this paper to generalize the model presented in [5], by considering a multi-server queueing model with threshold queues and hysteresis [6] in order to evaluate the performance of a cloud node (data center). We suppose that request arrivals are represented by a bulk arrivals process. In this model, virtual machines are activated and deactivated according to the intensity of user demand. Each server represents a VM, and the multi-server queueing model with hysteresis is governed by sequences of forward and reverse thresholds which are different. The forward (resp. the backward) thresholds represent the values of the number of customers from which an additional VM is activated (resp. deactivated). Obviously, the relevance of this model unlike the models proposed in [2]–[4] is to offer the flexibility of different thresholds for activating and deactivating VMs. Moreover, the hysteresis prevents the frequent activation or deactivation of VMs which could be costly in energy consumption.

As the system is difficult to analyze exactly, especially when the number of servers or the size of the batch arrivals distribution increases [7], we propose to use stochastic comparisons in order to compute performance measure bounds. We define bounding systems easier to solve in order to have a trade-off between accuracy and computation times which could be relevant for network dimensioning.

Several performance metrics as blocking probability, mean number of customers in the buffer and mean number of departures are evaluated according to different values of input parameters: buffer size, number of VMs (called also the degree of virtualization), and utilization rate. The paper is organized as follows: next we describe the cloud system, in Section III, we present the hysteresis queueing system used to model a data center. In section IV, we give a brief description of stochastic ordering theory and in Section V, the bounding models are presented with stochastic comparisons proofs. In the section VI, we give some numerical results for the performance measures of the cloud system. Finally, achieved results are discussed in the conclusion and comments about further research issues are given.

II. Cloud system description

We consider a data center in a cloud system composed of Physical Machines (PMs) with each physical machine hosting many Virtual Machines (VMs) [5], as illustrated in Figure 1. Incoming job requests are assumed to be a bulk arrival process and are enqueued in the queue. Such a queue has a finite size C; so, an arriving request can be rejected if it finds the buffer full. The system queue is managed according to a FIFO (First In First Out) scheduling policy. When a resource is available, a job is accepted and the corresponding VM is instantiated. We assume that the instantiation time is negligible.

In order to have a system able to handle the variability of the traffic intensity, we propose to activate and deactivate the VMs according to the system occupancy. In fact, the buffer management (the scheduler) governed by thresholds vectors and by the number of customer waiting in the queue, controls the operation of activating and deactivating the VMs. So, this system provides the dynamicity of the service according to the scalability of user requests. More formally, when the number of requests in the queue reaches a threshold, a new VM is activated, and in the same way, when it decreases below the threshold, a VM is deactivated. We detail in the next section, the considered queueing model used for the analysis of the performance of the cloud node.

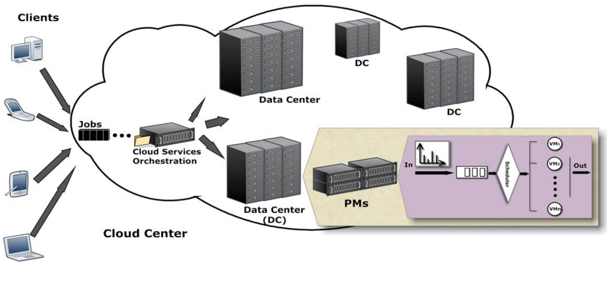

Fig. 1. Cloud center architecture.

III. Model description

We consider a queue with a finite buffer capacity and multi-homogeneous servers (VMs) [7]. We suppose a K multi-server thresholds-based queueing system with hysteresis for which a set of forward thresholds $(F_1, F_2, \ldots, F_{K-1})$ and a set of reverse thresholds $(R_1, R_2, \ldots, R_{K-1})$ are defined. We assume that $F_1 < F_2 < \ldots < F_{K-1}$, $R_1 < R_2 < \ldots < R_{K-1}$, and $R_i < F_i, \forall 1 \leq i \leq K-1$. The behavior of this system is as follows. We assume that the first VM is still active in the system. If a customer arrives in the system and finds F_i ($i = 1, \ldots, K-1$) customers, then an additional VM will be activated. When a customer leaves the system with R_i ($i = 1, \ldots, K-1$) customers, then a VM will be deactivated from the active VMs. We denote by $X(t)$ the model where each state is represented by $x = (x_1, x_2)$, with x_1 is the number of customers waiting in the system and x_2 is the number of active VMs. We suppose that job request arrivals follow a bulk-arrival process. So, we consider that requests are bulks (or batches), which arrive according to a Poisson process with rate λ, and length of bulks follow a probability distribution defined as follows: $p = (p_1, p_2, \ldots, p_k, \ldots, p_n)$, where $p_k = \Pr[\text{bulk length is } k,\ k \in E]$, $E \subset \mathbb{N}$, and we suppose that the size of E is n. Servers (or VMs) have an exponential service time distribution with mean rate $\mu_i = \mu$ ($i = 1, \ldots, K$). We suppose that the system has a finite capacity C. With these assumptions, we deduce that the system $X(t)$ is a Continuous-Time Markov Chain (CTMC) defined over the state space A such that:

$$\begin{aligned} A = \{(x_1,\, x_2) \quad | \quad & 0 \leq x_1 \leq F_1, \text{ if } x_2 = 1; \\ & R_{i-1} < x_1 \leq F_i, \text{ if } x_2 = i \text{ and } 1 < i < K; \\ & R_{K-1} < x_1 \leq C, \text{ if } x_2 = K\}. \end{aligned}$$

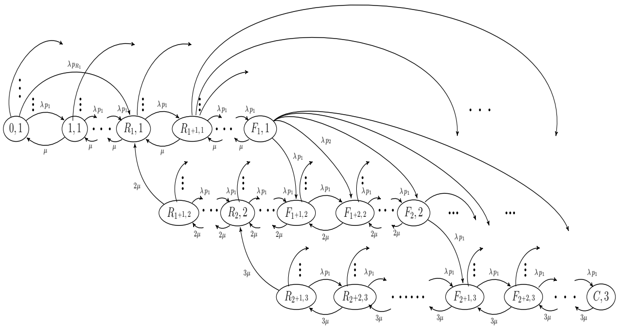

Fig. 2. State transition diagram for three-servers.

The evolution equations of $X(t)$ are defined as follows:

$$\begin{aligned} (x_1, x_2) \quad \rightarrow \quad & (\min\{C, x_1 + k\}, x_2), \\ & \text{with rate } \lambda p_k,\ \ \forall\, k \in E \\ & \text{if } (x_1 + k) \leq F_j, \text{ and } x_2 = j, \\ \rightarrow \quad & (\min\{C, x_1 + k\}, K), \\ & \text{with rate } \lambda p_k,\ \ \forall\, k \in E \\ & \text{if } x_2 = K \text{ or } (x_1 + k) > F_{K-1}, \\ \rightarrow \quad & (\min\{C, x_1 + k\}, l), \\ & \text{with rate } \lambda p_k,\ \ \forall\, k \in E \\ & \text{if } l = \min\{h | (x_1 + k \leq F_h) \text{ and } x_2\text{+1} \leq h \leq \text{K-1}\}, \\ \rightarrow \quad & (\max\{0, x_1 - 1\}, x_2), \\ & \text{with rate } x_2\mu, \\ & \text{if } (x_1 \neq R_i + 1 \text{ or } (x_1 = R_i + 1 \text{ and } x_2 \neq i + 1)) \\ \rightarrow \quad & (\max\{0, x_1 - 1\}, \max\{0, x_2 - 1\}), \\ & \text{with rate } x_2\mu, \\ & \text{if } x_1 = R_i + 1, \text{ and } x_2 = i + 1, \end{aligned}$$

where i, $j = 1, \ldots, K$-1. In the Figure 2, we illustrate the transition graph of such a Markov Chain.

Obviously, this system has been already studied in the literature [7]. In the paper, the authors use the concept of stochastic complementation to solve the system. They propose to partition the state space in disjoints sets in order to aggregate the Markov chain. The main advantage of this method is to obtain exact performance results, with reduced execution times. In this paper, we propose another approach by defining bounds rather than an exact resolution of the system which is often very cumbersome. Indeed, the relevance of using bounds is first to reduce dynamically the state space size of the system in order to obtain a trade-off between the accuracy of the results and the computation time. Next, we present briefly the stochastic ordering theory used to define our bounding systems.

IV. Stochastic ordering theory

We give some theorems and definitions about stochastic orderings [8] used in this paper. We consider a discrete, and countable state space A, endowed by the total order $\leq$ [8]. As an example, on the state space $A = \mathbb{R}$, "$\leq$" is the total order. We consider two independent random variables X and Y defined on A, with probability mass functions p and q ($p_i = \text{Prob}(X = i)$, and $q_i = \text{Prob}(Y = i)$, for $i = 1, 2, \ldots, |A|$). The most known stochastic ordering is the strong stochastic ordering, and it is denoted by $\leq_{st}$. It could be defined using increasing functions as follows [8].

Definition 1:

- *generic definition:* $X \leq_{st} Y \iff \mathbb{E}f(X) \leq \mathbb{E}f(Y)$, for all non decreasing functions $f : A \to \mathbb{R}^+$ whenever expectations exist.
- *probability mass functions:*

$$X \leq_{st} Y \Leftrightarrow \forall i,\, 1 \leq i \leq n,\; \sum_{k=i}^{n} p_k \leq \sum_{k=i}^{n} q_k. \quad (1)$$

Notice that we use interchangeably $X \leq_{st} Y$ and $p \leq_{st} q$.

We can also compare stochastic processes. Let $\{X(t), t \geq 0\}$ and $\{Y(t), t \geq 0\}$ be stochastic processes defined on A.

Definition 2: We say that $\{X(t), t \geq 0\} \leq_{st} \{Y(t), t \geq 0\}$, if $X(t) \leq_{st} Y(t), \forall t \geq 0$

When the processes are defined on different state spaces we can compare them on a common state space using mapping functions. Let $\{X(t), t \geq 0\}$ (resp. $\{Y(t), t \geq 0\}$) defined on A (resp. B), g (resp. h) be a many to one mapping from A to S, (resp. $B \to S$). Next, we compare the mapping of the process $\{X(t), t \geq 0\}$ (resp. $\{Y(t), t \geq 0\}$) by the mapping function g (resp. h), which means $g(X(t))$ (resp. $h(Y(t))$), on the common state space S.

The stochastic comparison of processes by mapping functions is defined as follows [9]:

Definition 3: We say that $\{g(X(t)), t \geq 0\} \leq_{st} \{h(Y(t)), t \geq 0\}$, if $g(X(t)) \leq_{st} h(Y(t)), \forall t \geq 0$

We can use the coupling method for the stochastic comparison of the processes. For the $\leq_{st}$ ordering, the coupling method can be used for the stochastic comparison of Continuous Time Markov Chains (CTMCs). As presented in [9], it remains us to define two CTMCs: $\{\widehat{X}(t),\ t \geq 0\}$ (resp. $\{\widehat{Y}(t),\ t \geq 0\}$) governed by the same infinitesimal generator matrix as $\{X(t), t \geq 0\}$ (resp. $\{Y(t), t \geq 0\}$), representing different realizations of these processes with different initial conditions. The following theorem establishes the $\leq_{st}$-comparison using the coupling [9]:

Theorem 1:

$$\{g(X(t)), t \geq 0\} \leq_{st} \{h(Y(t)), t \geq 0\} \quad (2)$$

if there exists the coupling $\{(\widehat{X}(t), \widehat{Y}(t)),\ t \geq 0\}$ such that:

$$g(\widehat{X}(0)) \leq h(\widehat{Y}(0)) \Rightarrow g(\widehat{X}(t)) \leq h(\widehat{Y}(t)),\ \forall t > 0 \quad (3)$$

V. Bounding systems

We propose to define different bounding systems which are easier to solve. Different ways to simplify the exact system are used. The first bounding systems are defined by reducing the size of the bulk arrivals and so by aggregating the probability distribution of bulk arrivals, in order to obtain aggregated bounding bulk arrivals. So, these first bounding systems are represented by a hysteresis system with aggregated bounding bulk arrivals. The second aggregated bounding processes are obtained by defining bounding processes with the same sequences of forward and reverse thresholds. Next, we describe the proposed bounding models.

A. Hysteresis system with aggregated bounding arrival process

We consider hysteresis systems equivalent to the exact system with arrival process defined by a Poisson process with the rate λ. The bulk (or batch) lengths follow an upper bound probability distribution p^u (resp. a lower bound probability distribution p^l) of the probability distribution p. The bounding probability distributions of the bulk lengths are obtained by aggregations, in order to reduce the size of p, with the following relation:

$$p \leq_{st} p^u \quad \text{and} \quad p^l \leq_{st} p.$$

If p is defined on a state space of size n (called also bins), then p^u (resp. p^l) are defined on a state space of size m, such that $m << n$. These bounding probability distributions can be obtained through the approach and the algorithms developed in [10]. Intuitively, p^u (resp. p^l) is obtained by removing some states of p and adding their probability mass on higher states (resp. on lower states). The distributions p^u and p^l are computed to be the closest distributions with m states according to an increasing reward function. The optimality of computed bounding distributions, proved in [10] helps to obtain tight bounds on the results. Let $X^u(t)$ (resp. $X^l(t)$) be the hysteresis system built with the bulk arrival probability distribution p^u (resp. p^l). Next, we prove that these Markov chains represent stochastic bounds for $X(t)$.

1) Stochastic comparison proofs: We define the many to one mapping function $g : A \to S$, such that $g(x) = x_1$, where $x_1 \in S = \{0, \ldots, C\}$. And in the state space S, we use the total order $\leq$. We apply the stochastic ordering theory presented before to derive the following theorem:

Theorem 2: We have the following relations:

1) $g(X(0)) \leq_{st} g(X^u(0)) \Rightarrow g(X(t)) \leq_{st} g(X^u(t)),\ \forall\, t > 0.$
2) $g(X^l(0)) \leq_{st} g(X(0)) \Rightarrow g(X^l(t)) \leq_{st} g(X(t)),\ \forall\, t > 0.$

Proof: We use theorem 1 based on the coupling of the processes. We begin with the first relation of theorem 2, in order to establish that $\{X^u(t), t \geq 0\}$ is really an upper bound. For the proof, we suppose that at time t, $X(t) = x$ and $X^u(t) = y$, and from the definition of the mapping function g, $g(X^u(t)) = y_1$, and $g(X(t)) = x_1$. The proof is by induction, so we suppose that the order is verified at time t ($x_1 \leq y_1$), and we prove that at time $t + dt$ the order is still verified. We denote by $X^u(t+dt) = y'$ and $X(t+dt) = x'$, and so $g(X^u(t+dt)) = y'_1$, and $g(X(t+dt)) = x'_1$. We consider the two kinds of events: arrivals and services.

- Arrivals: if we have an arrival of size k in $X(t)$ such that at time $t+dt$, $x'_1 = x_1 + k$, then we can have also a transition in $X^u(t)$ from y to y' such that $y'_1 = y_1 + l$, and $k \leq l$, as $p \leq_{st} p^u$. So $x'_1 \leq y'_1$, and the order between the processes is still verified at time $t + dt$.
- Services: if we have a service for $X^u(t)$ such that at time $t + dt$, $y'_1 = y_1 - 1$, then we can have also a service in $X(t)$ such that at time $t + dt$, we have $x'_1 = x_1 - 1$, as the transition rates are the same in the two systems. So, $x'_1 \leq y'_1$, and the order between the processes is still verified at time $t + dt$.

■

For the lower bound $\{X^l(t), t \geq 0\}$, the proof is similar. As the bulk length probability distributions are such that $p^l \leq_{st} p$, and the service rates are the same, then the second relation of theorem 2 is verified.

The second kinds of bounding models are obtained by removing the notion of hysteresis in the system. Considering the same sequence for forward and reverse threshold sequences, these new models are easier to analyze and allow to derive also bounds for performance measures of a cloud node.

B. Bounding systems with equal forward and backward sequence

Considering the same forward and reverse threshold vectors, we derive upper and lower bounding models for the threshold queueing system with hysteresis. For the upper bound, we take the vector $(F_1, F_2, \ldots, F_{K-1})$ as a forward and reverse thresholds. And, for the lower bound, we take $(R_1, R_2, ...R_{K-1})$ for the forward and the reverse thresholds.

The behavior of each of these systems are represented by CTMCs defined on state space $S = \{0, \ldots, C\}$. We denote by $Y(t)$ the CTMC associated to the upper bounding model (the forward and the reverse thresholds are given by $(F_1, F_2, ...F_{K-1})$). The evolution equation of this model is given as follows:

$$\begin{aligned} x \quad &\rightarrow \quad \min(C,\, x+k),\ \text{with rate } \lambda p_k,\ \ \forall\, k \in E & (4)\\ &\rightarrow \quad \max(0,\, x-1), \text{with rates:} & (5)\\ &\qquad \bullet\ i\mu,\ \text{if } F_{i-1} < x \leq F_i,\ \forall i = 1 \ldots K-1 & (6)\\ &\qquad \bullet\ K\mu,\ \text{if } F_{K-1} < x \leq C & (7) \end{aligned}$$

where $F_0 = 0$.

In the same way, we define by $Z(t)$ the CTMC which represents the lower bound (with $(R_1, R_2, ...R_{K-1})$ for the forward and the reverse thresholds). In this case, the above equations ((4)-(7)) are also available by changing the sequence $F_{i,\, i=1\ldots K-1}$, by the sequence $R_{i,\, i=1\ldots K-1}$. We give in figures 3 and 4, the transition diagrams of upper bounding and lower bounding models for three-servers.

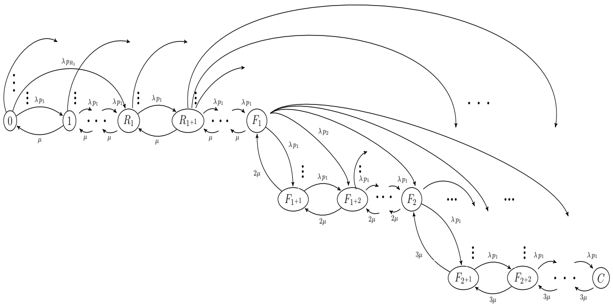

Fig. 3. State transition diagram for upper bounding model with three-servers.

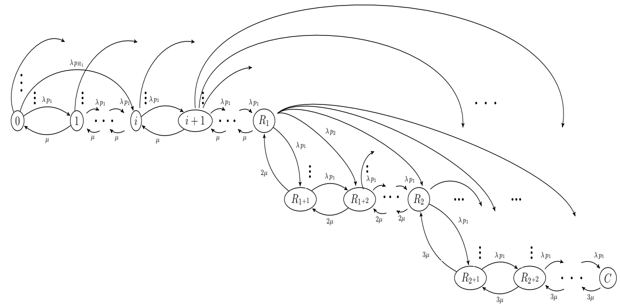

Fig. 4. State transition diagram for lower bounding model with three-servers.

Next, we will prove that $Y(t)$ (resp. $Z(t)$) is a stochastic upper bound (resp. lower bound) for $X(t)$.

1) Stochastic comparison proofs: We have the following theorem:

Theorem 3: If $X(t)$, $Y(t)$, and $Z(t)$ represent the systems defined previously, then we have:

1) $g(X(0)) \leq_{st} Y(0) \Rightarrow g(X(t)) \leq_{st} Y(t), \forall t > 0$
2) $Z(0) \leq_{st} g(X(0)) \Rightarrow Z(t) \leq_{st} g(X(t)), \forall t > 0$

Proof: We begin by the first equation of theorem 3. For the proof, we apply the theorem 1, in order to use the coupling by the mapping function for the stochastic comparison. In our case, we have $X(t)$ defined on A, g a mapping function $A \rightarrow S$, $Y(t)$ defined on S, and h is the identity function. The proof is done by induction: we consider $X(t) = x = (x_1, x_2)$ and $Y(t) = y$, such that $g(x) \leq y$ and we prove that for any event, at time $t + dt$, if $X(t+dt) = x' = (x'_1, x'_2)$, and $Y(t+dt) = y'$, then $g(x') \leq y'$. In fact, for states x and y such that $g(x) \leq y$, we consider the two events for the proof of the stochastic comparison by the mapping function g:

- Arrivals: if we have a transition from x to x' such that $x'_1 = x_1 + k$ (for $k > 0$) then we are sure that we can have also a transition from y to y' such that $y' = y + l$ (for $l > 0$, $k \leq l$), because the arrival processes are the same in the two systems. So, we deduce that $g(x') \leq y'$.
- Services: if we have a transition from y to y' such that $y' = y - 1$, then we are sure that we have also

a transition from x to x' such that $g(x') = x_1 - 1$, because the rate for decreasing in $Y(t)$ is lower than in $X(t)$ (as the threshold for deactivation is lower in $X(t)$ than in $Y(t)$). So, $g(x') \leq y'$.

Then, we deduce that for any events, we have: $g(X(t + dt)) \leq Y(t + dt)$, and from theorem 1, we deduce that in theorem 3, equation 1 is verified. For the second equation in theorem 3, the proof is similar, and we deduce easily that $Z(t)$ is a lower bound.

■

Another simplification of these bounding models consists to consider aggregation on batch arrivals distribution for $Y(t)$ and $Z(t)$. We denote by $Y^u(t)$ (resp. $Z^l(t)$) the Markov chain with batch arrival distribution p^u (resp. p^l) (see Subsection V-A). In the sequel we give the following theorem.

Theorem 4: We have the following relations:

- $Y(0) \leq_{st} Y^u(0) \Rightarrow Y(t) \leq_{st} Y^u(t), \; \forall\, t > 0.$
- $Z^l(0) \leq_{st} Z(0) \Rightarrow Z^l(t) \leq_{st} Z(t), \; \forall\, t > 0.$

The proof is similar to Theorem 2. Indeed, $Y^u(t)$ is obtained by replacing the probability distribution of batch lengths p by p^u in $Y(t)$, and $p \leq_{st} p^u$. In the same way, $Z^l(t)$ is obtained by replacing p by p^l in $Z(t)$, and $p^l \leq_{st} p$.

From Theorem 3 and Theorem 4, we deduce that $Y^u(t)$ (resp. $Z^l(t)$) represents an upper bound (resp. lower bound) of $X(t)$.

The relevance of the definition of bounding systems is to generate bounds for performance measures as expected buffer length, expected departures, and blocking probabilities. Next, we will present some numerical results of performance measures.

VI. Numerical examples

In this section, we present some numerical examples which illustrate the relevance as well as the reduced complexity of the bounding models presented in the paper. We note that the bounding systems developed before represent bounds on the number of jobs in the system. These bounds provide a coverage on performance measures defined as an increasing function of the number of jobs in the system such as expected buffer length, expected departures and blocking probabilities. We also present the execution time (in seconds) needed to compute the performance metrics of the models. We note that to compute the steady state probability distribution vector of the considered models, we use the methodology proposed by Lui and Golubchik in [7] based on the stochastic complementation denoted here by "SCA" (Stochastic Complement Analysis). We note that this approach is proven to be less complex than the commonly used solution techniques [11].

Considering a threshold-based system with hysteresis, our goal through these examples consists to show the interest and the usefulness that may represent the stochastic bounding models proposed in the paper to obtain accurate results and guarantees on performance measures. We present below three examples through which we propose to vary some input parameters as buffer size, degree of virtualization (number of servers), and utilization rate of the system. We study here the quality of results and the computational times in order to offer to the user a range of models, which are "less complex" and very relevant for network dimensioning. For the three examples, we give the performance measures for these models:

- Hysteresis model with exact batch-arrival distribution ($X(t)$)
- Hysteresis model with stochastic lower bound of batch-arrival distribution ($X^l(t)$)
- Hysteresis model with stochastic upper bound of batch-arrival distribution ($X^u(t)$)
- Upper bounding model with exact batch-arrival distribution ($Y(t)$)
- Upper bounding model with stochastic upper bound of batch-arrival distribution ($Y^u(t)$)
- Lower bounding model with exact batch-arrival distribution ($Z(t)$)
- Lower bounding model with stochastic lower bound of batch-arrival distribution ($Z^l(t)$)

A. Some QoS parameters versus buffer size

As a first example, we consider a threshold-based queue with hysteresis and batch-arrival, such that: the number of servers is $K = 10$, the service rate is set to 100, the distribution of the batch arrivals is randomly generated on a support $\{1, 2, 3, \ldots, 500\}$ and the arrival rate is taken equal to 1. We propose to vary the buffer size from C=1000 to C=6000 and observe some performance measures for the studied models. According to the buffer length, the forward and the reverse threshold vectors are taken as follows: for C=C_1=1000, the threshold vectors are $F = [90, 140, 280, 400, 610, 690, 730, 840, 910]$ and $R = [30, 90, 190, 270, 410, 510, 620, 700, 800]$, for $C = C_i = i \times 1000$, the threshold vectors are $F_i = i \times F$ and $R_i = i \times R$.

Depending on the values of the buffer size, the figures 5, 6 and 7 (resp. figures 8, 9 and 10) illustrate the expected buffer length, the expected departures and the blocking probabilities for reduction of batch arrival distribution $bins$=10 (resp. $bins$=50). For $bins$=50, we illustrate in Figure 11 the computation times in seconds needed to solve the different models.

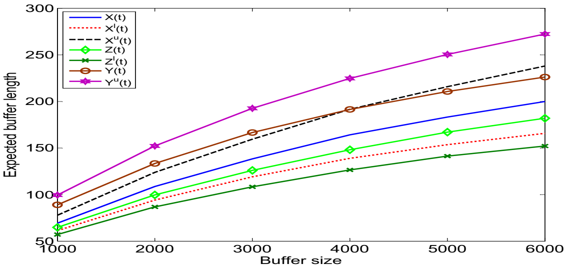

Fig. 5. Expected buffer lengths versus buffer size, for $bins = 10$.

Through these figures, we remark that the bounding systems define a good coverage of the exact result (hysteresis

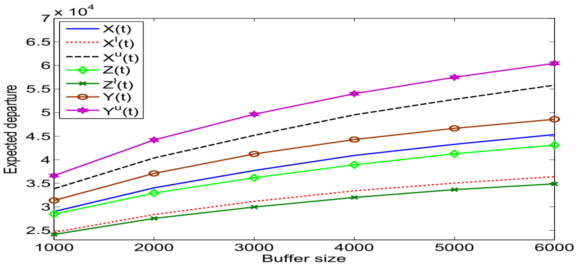

Fig. 6. Expected departures versus buffer size, for $bins = 10$.

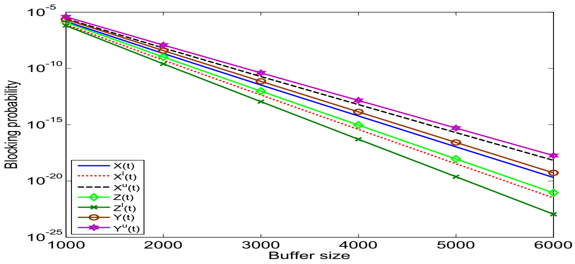

Fig. 7. Blocking probabilities versus buffer size, for $bins = 10$.

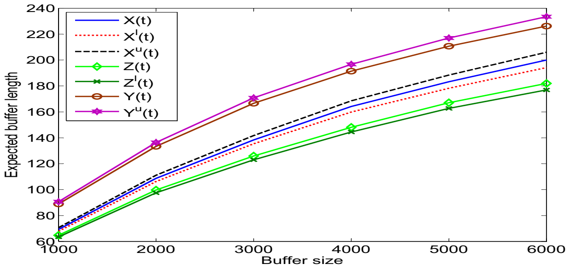

Fig. 8. Expected buffer lengths versus buffer size, for $bins = 50$.

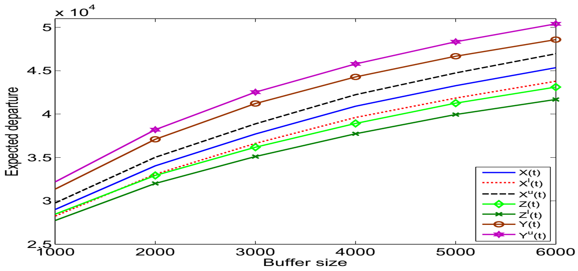

Fig. 9. Expected departures versus buffer size, for $bins = 50$.

model with exact batch-arrivals) and the accuracy of these bounds are very improved when we increase the number of bins ($bins = 50$). Regarding the computation times, we observe that the upper and the lower bounding models ($Y(t)$ and $Z(t)$) with exact arrival distribution allow to reduce slightly the computation times, and the bounding models with a reduction in the size of batch-arrival distribution ($X^u(t)$, $X^l(t)$, $Y^u(t)$ and $Z^l(t)$) allow for their part to reduce significantly the computational time. Indeed, even if a reduction of the batch-arrival distribution may seem important (from 500 states to only 10 states or 50 states), the results on performance metrics are, however, very relevant and close to the exact results with very low computational times. We remark that for $bins = 50$, $X^u(t)$ and $X^l(t)$ provide the most accurate bounds.

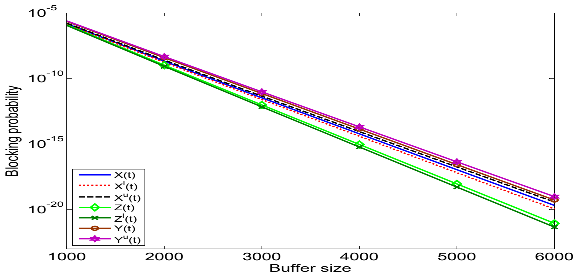

Fig. 10. Blocking probabilities versus buffer size, for $bins = 50$.

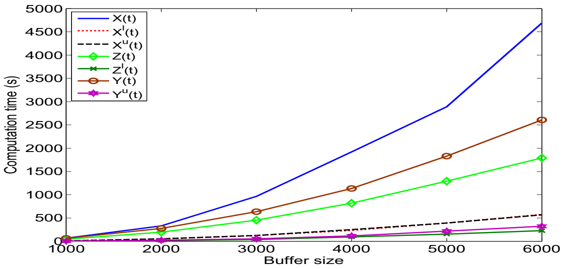

Fig. 11. Computation times (in second) versus buffer size.

B. Some QoS parameters versus number of servers

For the second example, we propose to vary the degree of virtualization of the servers in the threshold-based queue with hysteresis and observe the behavior of some performance measures. So, we consider a threshold-based queue with hysteresis and batch-arrival such that: the service rate is set to 100, the distribution of the batch arrivals is randomly generated on the support $\{1, 2, 3, \dots, 500\}$, the arrival rate is taken equal to 1, and the buffer size is set to $C = 2000$.

We are interested here in computing some performance measures by varying the number of servers from $K = 5$ to $K = 200$. For the different degree of virtualization considered, we use the following equation to define respectively the forward and the reverse threshold vectors: $F = (\lfloor \frac{C}{K} \rfloor, 2 \times \lfloor \frac{C}{K} \rfloor, \dots, (K-1) \times \lfloor \frac{C}{K} \rfloor)$ and $R_i = F_i - \lfloor \frac{C}{2\,K} \rfloor$, for $i = 1, \dots, K-1$.

Thus, depending on the degree of virtualization, the figures 12, 13 and 14 illustrate the expected buffer length, the expected departures and the blocking probabilities for the studied models. The reduction considered here is $bins = 50$. The Figure 15 presents the computation times in seconds for the different models.

From these curves, we see that the bounding results frame the exact results and are very accurate. We observe that the performance measures obtained by the lower bounding

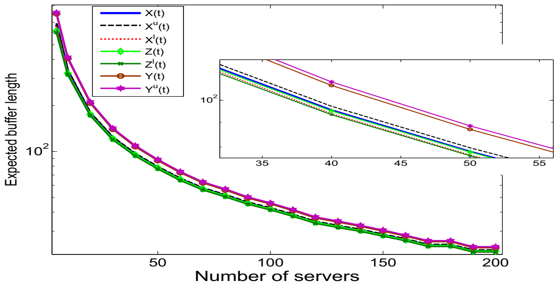

Fig. 12. Expected buffer lengths versus degree of virtualization.

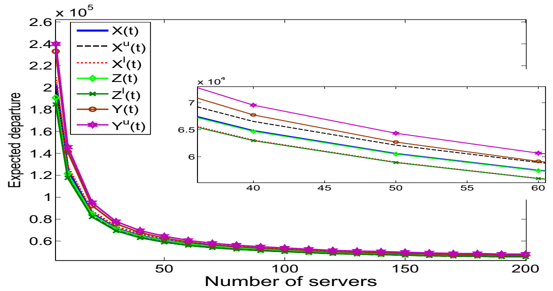

Fig. 13. Expected departures versus degree of virtualization.

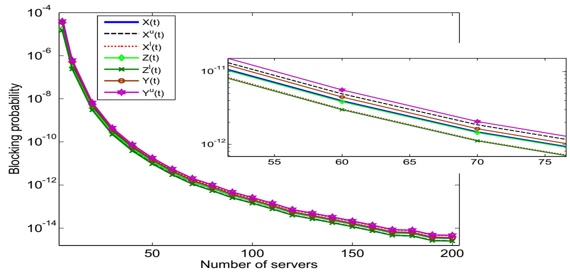

Fig. 14. Blocking probabilities versus degree of virtualization.

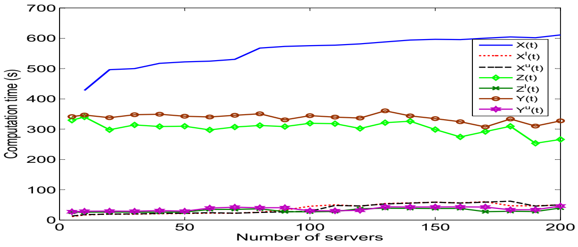

Fig. 15. Computation times (in second) versus degree of virtualization.

model with exact batch arrival distribution ($Z(t)$) are the closest results, but the other models remain very relevant. The times needed to make the computations for the models with smaller sizes of batch length arrival distribution are very short compared to the models with the original arrival distribution. Thus, from these observations, we can say that in order to accelerate the computation times and avoid a cumbersome resolution, the use of stochastic bounds on batch arrival distribution proposed here can be very interesting and represent a good trade-off between accuracy of the results and the computational complexity.

C. *Performance measures versus utilization rate and number of servers*

We consider here two cases where the number of servers, K, is equal to 10 and 50. We suppose that the buffer size is $C = 1500$, for $K = 10$, the threshold vectors are $F = (100, 200, 300, 400, 500, 600, 700, 800, 900)$ and $R = (80, 180, 280, 380, 480, 580, 680, 780, 880)$ and, for $K = 50$, the threshold vectors are $F = (20, 40, 60, \ldots, 980)$ and $R = (5, 25, 45, \ldots, 965)$. For all values of K, the service rate μ is set to 1. The distribution of the batch length arrivals is randomly generated with the support $\{1, 2, \ldots, 500\}$ and we propose to vary the arrival rate λ. So, we distinguish a lightly loaded system with $\lambda = 0.2$ for $K = 10$ and $\lambda = 0.5$ for $K = 50$ and a highly loaded system with $\lambda = 2$ for $K = 10$, and $\lambda = 6$ for $K = 50$. We note that the size of reduction considered for stochastic bounding distributions are respectively $bins = 10$ and $bins = 50$.

We depict in tables I, II, III and IV the expected buffer length (denoted $\mathbb{E}[\mathcal{Q}]$) and the expected departures (denoted $\mathbb{E}[\mathcal{D}]$) computed for different studied models and for two degree of virtualization of the server ($K = 10$ and $K = 50$) and utilization rate system (denoted "U"), $\mathrm{U} \simeq 0.20$ and $\mathrm{U} \simeq 0.96$. We present also the computation times for the different models. We propose here to compare the resolution method developed by Lui and Golubchik in [7] and used until now for our numerical resolution with a well known approach which consists to use the iterative power method on the matrix of the model as solution technique (see [11]) with precision $\epsilon = 10^{-10}$. We denote this solution technique by PM (Power Method).

			$\mathbb{E}[\mathcal{Q}]$	$\mathbb{E}[\mathcal{D}]$	time "PM"	time "SCA"
Hys. model	Exact		28.62	9653.2	107.89	49.32
	St-L.B.	bins=10	25.25	8256.54	7.15	4.44
		bins=50	28.07	9416.28	12.49	7.461
	St-U.B.	bins=10	32.28	11190.6	7.15	3.19
		bins=50	29.16	9890.27	12.51	7.07
U.B. model	Exact		35.77	10007	98.13	32.00
	St-U.B.	bins=10	39.67	1159.4	3.64	2.44
		bins=50	36.40	10254.3	10.79	3.27
L.B. model	Exact		27.96	9620.31	75.22	33.72
	St-L.B.	bins=10	24.78	8235.69	3.3102	2.55
		bins=50	27.44	9385.26	8.57	3.59

TABLE I. SOME QOS PARAMETERS FOR $K = 10$ AND $\mathrm{U} \simeq 0.20$.

From these tables, the observations made previously are also seen in this example. So, we show clearly that the results provided after using the stochastic bounds on the batch-arrival distribution, are very accurate and gives a good coverage of

			$\mathbb{E}[\mathcal{Q}]$	$\mathbb{E}[\mathcal{D}]$	time "PM"	time "SCA"
Hys. model	Exact		411.25	278610	219.11	31.77
	St- L.B.	bins=10	363.11	230805	18.69	4.78
		bins=50	403.43	270649	34.80	7.39
	St- U.B.	bins=10	459.56	329040	17.95	4.16
		bins=50	419.27	286798	41.54	6.81
U.B. model	Exact		455.63	299189	140.27	21.97
	St- U.B.	bins=10	503.34	350453	14.55	3.74
		bins=50	463.60	307568	20.89	4.39
L.B. model	Exact		407.07	276685	148.20	27.45
	St- L.B.	bins=10	358.92	229006	16.15	3.88
		bins=50	399.23	268734	24.03	5.10

TABLE II. SOME QOS PARAMETERS FOR $K = 10$ AND U $\simeq 0.96$.

			$\mathbb{E}[\mathcal{Q}]$	$\mathbb{E}[\mathcal{D}]$	time "PM"	time "SCA"
Hys. model	Exact		19.95	23060.4	170.54	106.69
	St- L.B.	bins=10	17.94	19821.4	6.46	5.31
		bins=50	19.63	22511.2	18.64	17.12
	St- U.B.	bins=10	21.99	26585.2	6.69	5.01
		bins=50	20.29	23610.6	19.17	17.38
U.B. model	Exact		22.33	23354.8	114.63	74.59
	St- U.B.	bins=10	24.53	26929.1	5.45	5.32
		bins=50	22.69	23913.5	13.85	11.51
L.B. model	Exact		19.90	23053.6	108.85	68.25
	St- L.B.	bins=10	17.91	19818.1	3.99	3.55
		bins=50	19.57	22504.8	12.01	9.18

TABLE III. SOME QOS PARAMETERS FOR $K = 50$ AND U $\simeq 0.2$.

			$\mathbb{E}[\mathcal{Q}]$	$\mathbb{E}[\mathcal{D}]$	time "PM"	time "SCA"
Hys. model	Exact		275.16	652925	247.96	162.96
	St- L.B.	bins=10	247.21	545394	9.01	7.77
		bins=50	270.63	634823	26.19	19.03
	St- U.B.	bins=10	303.18	769411	9.27	4.24
		bins=50	279.81	671516	27.81	18.69
U. B. model	Exact		288.93	673130	151.78	112.07
	St- U.B.	bins=10	317.07	791503	6.33	5.31
		bins=50	293.60	692051	16.36	16.01
L. B. model	Exact		274.81	652407	152.39	100.44
	St- L.B.	bins=10	246.89	544972	5.65	4.89
		bins=50	270.27	634317	17.04	16.98

TABLE IV. SOME QOS PARAMETERS FOR $K = 50$ AND U $\simeq 0.96$.

the exact results with considerably reduced computation times. Moreover, for computation times, we can observe easily that the time needed to compute the bounds even for $bins = 50$ are significantly smaller than those used to obtain exact results (PM and SCA methods).

To summarize, when we have a network dimensioning problem with QoS constraints, we propose first to build upper and lower bounding models with stochastic bounds on the bulk-arrival distribution as they represent the smallest models in terms of size of the state space and the fastest in term of resolution time. If the computed performance bounds do not verify the QoS constraints (even when we increase the number of bins), we can subsequently use the hysteresis models with stochastic bounds on the bulk-arrival distribution. Because these models can get closer to the exact values when we increase the number of bins.

VII. CONCLUSION

We propose in this paper a queue-dependent multiserver VMs with hysteresis in order to model the behavior of a data center in a cloud system. The relevance of this model is to represent the dynamicity of the resource according to the queue occupation, and with the hysteresis to reduce the cost due to activation/deactivation of the VMs. However, this system could be difficult to analyze when the number of VMs increases, so we propose to use bounding techniques in order to derive bounds for the performance measures. Through this paper we show clearly the accuracy and the relevance of the proposed stochastic bounding models. As a future work, we expect to investigate a more general system by considering heterogeneous servers (VMs) and defining optimal thresholds vectors in order to optimize the performance of cloud systems.

ACKNOWLEDGMENT

This work was supported by grant ANR MARMOTE (ANR-12-MONU-0019).

REFERENCES

[1] P. Wang, W. Huang, and C. A. Varela, "Impact of virtual machine granularity on cloud computing workloads performance," in *Proceedings of the 2010 11th IEEE/ACM International Conference on Grid Computing*, 2010, pp. 393–400.

[2] R. Ghosh, F. Longo, V. K. Naik, and K. S. Trivedi, "Modeling and performance analysis of large scale iaas clouds," *Future Generation Comp. Syst.*, vol. 29, no. 5, pp. 1216–1234, 2013.

[3] H. Khazaei, J. V. Misic, and V. B. Misic, "Performance of cloud centers with high degree of virtualization under batch task arrivals," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 12, pp. 2429–2438, 2013.

[4] H. Khazaei and J. V. Misic and V. B. Misic, "A fine-grained performance model of cloud computing centers," *IEEE Trans. Parallel Distrib. Syst.*, vol. 24, no. 11, pp. 2138–2147, 2013.

[5] V. Goswami, S. S. Patra, and G. B. Mund, "Performance analysis of cloud with queue-dependent virtual machines," in *1st International Conference on Recent Advances in Information Technology, RAIT 2012, Dhanbad, India, March 15-17, 2012*, 2012, pp. 357–362.

[6] O. C. Ibe and J. Keilson, "Multi-server threshold queues with hysteresis," *Perform. Eval.*, vol. 21, no. 3, pp. 185–213, 1995.

[7] J. C. S. Lui and L. Golubchik, "Stochastic complement analysis of multi-server threshold queues with hysteresis," *Performance Evaluation*, vol. 35, pp. 19–48, 1999.

[8] A. Müller and D. Stoyan, *Comparison methods for stochastic models and risks*, ser. Wiley series in probability and statistics. J. Wiley & sons, 2002.

[9] M. Doisy, "A coupling technique for stochastic comparison of functions of markov processes," *JAMDS*, vol. 4, no. 1, pp. 39–64, 2000.

[10] F. Aït-Salaht, H. Castel-Taleb, J. Fourneau, and N. Pekergin, "Stochastic bounds and histograms for network performance analysis," in *Computer Performance Engineering - 10th European Workshop, EPEW Italy*, ser. Lecture Notes in Computer Science, vol. 8168. Springer, 2013, pp. 13–27.

[11] W. Stewart, *Introduction to the numerical Solution of Markov Chains*. New Jersey: Princeton University Press, 1995.