Bounding aggregations on bulk arrivals for performance analysis of clouds

Farah Aït-Salaht, Hind Castel
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*—Considering a cloud system, we propose in this paper to apply bounding aggregations for mathematical analysis of a data center. Modeled as a hysteresis queueing system, a data center is characterized by forward and backward thresholds which allow to represent its dynamic behavior. The client requests (or jobs) are represented by bulk arrivals which arrive into the buffers and are executed by Virtual Machines (VMs). According to the occupation of the queue and the thresholds, the VMs are activated and deactivated. The system is represented by a complex Markov chain which is difficult to analyze when the size of the system is huge. We propose to use in this case bounding aggregations on the batch arrivals, in order to compute performance measure bounds. We present some numerical results for the performance measures in order to compare the bounding values with the exact ones according to the different input parameters. The relevance of this paper is to propose a tradeoff between computational complexity and accuracy of the results, which provides very interesting solutions in networking dimensioning.**

I. Introduction

One of the most significant recent progresses in the field of information and communication technology is Cloud computing, which may change the way people do computing and manage information. In this environment, a pool of abstracted, virtualized, dynamically-scalable computing functions and services are made accessible over the internet to remote users in an on-demand fashion, without the need for infrastructure investments and maintenance.

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. One physical host can have more than one VM (Virtual Machine: it is a software that can run its own operating system and applications just like an operating system on a physical computer). With this flexibility, the cloud providers can rent the virtual machines depending on the demand and can gain more profit out of a single physical machine. With virtualization, service providers can ensure isolation of multiple user workloads, provide resources in a cost-effective manner by consolidating 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 [9], 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 because 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 [5], they develop an analytical model in order to evaluate the performance of cloud centers with a high degree of virtualization and Poisson batch arrivals. The model of the physical machine with m VMs is based on 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 [6], 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 a high degree of virtualization, and are categorized into three server pools: hot, warm, and cold. The authors implement the sub-models using interactive Continuous Time Markov Chain (CTMC). The sub-models are interactive such that the output of one sub-model is input to the other one.

In this paper, we propose to use a mathematical model in order to evaluate the performance of a cloud node, more precisely, a data center. We represent the system by a queueing model based on queue-dependent virtual machines in order to analyse quantitatively the dynamic behavior of the data center. The data center is represented by a set of PM (Physical Machines) hosting a set of VMs which are instanced according to user demand. In this paper, we represent the data center as a set of VMs which could be very large, especially if the user demand is high. With this model, virtual machines are activated and deactivated according to the intensity of user demand. The queueing model is a multi-server with threshold queues and hysteresis [4]. We suppose that customer request arrivals follow a bulk process. Each server represents a VM, and the multi-server queueing model with hysteresis is governed by a sequence of forward and reverse thresholds which are different. The forward (resp. the backward) thresholds represent the value of the number of customers from which an additional VM is activated (resp. deactivated). Obviously, the relevance of this model is to offer the flexibility of different thresholds for activating and removing VMs.

As the system is difficult to analyze exactly, especially when the number of VMs or the size of bulk arrivals is high, we propose to use stochastic comparisons in order to compute more easily, and so faster performance measure bounds.

The bounding models are obtained by the simplification of the hysteresis model in order to compute easily the performance measures. We propose to simplify the batch arrival process by generating aggregated bounding processes. So the

bounding systems are equivalent to the hysteresis system with aggregated bounding arrival process. We derive an upper bounding system (resp. a lower bounding system) from an upper bound batch arrival distribution (resp. a lower bound batch arrival distribution). We prove using stochastic comparisons that these processes provide really bounds for performance measures as blocking probabilities, expected buffer length and expected departure.

We give some numerical values according to different values of input parameters: arrival rate, size batches, and the number of VMs (called the degree of virtualization). The results show clearly the relevance of our approach to propose a tradeoff between computational complexity and accuracy of results. So it can efficiently solve the network dimensioning problem from QoS (Quality of Service) constraint requirements.

The paper is organized as follows: next, we describe the cloud system, and in section III, we present the queueing model for the analysis. In section IV, we give some theoretical notions of the stochastic ordering theory and in section V, we give the bounding models and we prove using the stochastic comparisons that they represent really bounds. In the section VI, we give numerical results of the performance measures. Finally, achieved results are discussed in the conclusion and comments about further research issues are given.

II. Cloud system description

The system under study is a cloud center, which contains several data centers. Customer requests arrive from different devices (mobile phones, laptops, computers) to the system, and the cloud service orchestrator dispatches the job into the data center in order to provide service. The data center is a set of resources or PM (Physical Machines) each of which can host a lot of VMs (Virtual Machines), as shown in Fig 1. Different users may share a PM (Physical Machine) using virtualization technique which provides a well defined set of resources (as CPU, RAM, storage). The VMs provide service for customer requests. We focus our study on one data center, and we propose to represent it by a stochastic model based on a queueing system which captures the dynamicity of the resource provisioning according to the current workload.

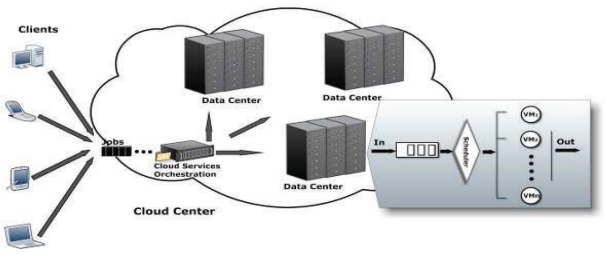

Fig. 1. Cloud center architecture

This system provides the dynamicity of the service according to the scalability of user requests. In order to have a system able to handle the variability of the traffic intensity, the VM are activated and deactivated according to the system occupancy. In fact, the buffer management is defined by thresholds for the number of customers waiting in the queue, which activate or deactivate the VMs. Clearly, when the number of customers in the queue reaches a threshold, then a new VM is activated, and when it decreases below the threshold, a VM is deactivated. In the next section, we present in details the queueing model used for the analysis of the performance of the data center.

III. Model description

We consider a finite buffer capacity with multi-homogeneous servers (VMs). 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 in the system, 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_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 client 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 size of bulks follow a probability distribution $p = (p_1, \ldots, p_k, \ldots, p_n)$, defined as follows:

$$p_k = \Pr[\text{bulk size is } k,\ k \in \mathcal{E}]$$

where $\mathcal{E} \subset \mathbb{N}$, and we suppose that the size of $\mathcal{E}$ 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 Chains (CTMCs) defined over the state space A such that:

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

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

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

where $i, j = 1, \ldots,$ K-1.

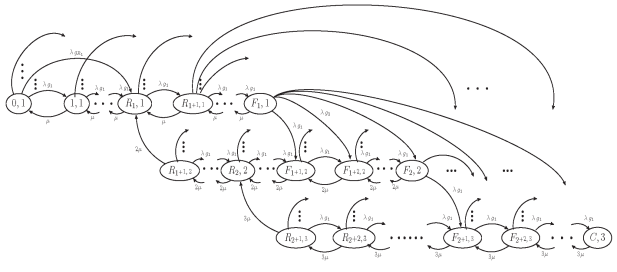

Fig. 2. Example of the state transition graph for a three-servers system.

Obviously, this system has been already studied in the literature by Lui and Golubchik in [7]. In [7], the authors use the concept of stochastic complementation to solve the system. They propose to partition the state space in disjoint sets in order to aggregate the Markov chain. We propose in this paper another approach which consists to define bounds rather than an exact resolution of the system. The relevance of using bounds is to offer a trade-off between the accuracy of the results and the computation time. Thus, in order to reduce the complexity of the Markov chain, we propose to apply the stochastic bounding approach to diminish the size of the batch probability distribution. The main advantage of this approach is the ability of computing bounds rather than approximations. Unlike approximation, the bounds allow us to check if QoS are satisfied or not. Next, we give some definitions and theorems about the stochastic ordering.

IV. Stochastic ordering theory

We refer to Stoyan's book [8] for theoretical issues of the stochastic comparison method. We consider state space $\mathcal{G} = \{1, 2, \ldots, n\}$ endowed with a total order denoted as $\leq$. Let X and Y be two discrete random variables taking values on $\mathcal{G}$, with cumulative probability distributions F_X and F_Y, and probability mass functions p and q ($p(i) = \text{Prob}(X = i)$, and $q(i) = \text{Prob}(Y = i)$, for $i = 1, 2, \ldots, n$). We give different manners to define the strong stochastic ordering $\leq_{st}$ for this case:

Definition 1: We can define the $\leq_{st}$ ordering as follows :

- **generic definition:** $X \leq_{st} Y \Longleftrightarrow \mathbb{E}f(X) \leq \mathbb{E}f(Y)$, for all non decreasing functions $f : \mathcal{G} \rightarrow \mathbb{R}^+$ whenever expectations exist.
- **cumulative probability distributions:**
$$X \leq_{st} Y \Leftrightarrow F_X(a) \geq F_Y(a),\ \forall a \in \mathcal{G}.$$
- **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 $p \leq_{st} q$.

Example 1: We consider $\mathcal{G} = \{1, 2, \ldots, 7\}$, and two discrete random variables with $d1 = [0.1,\ 0.2,\ 0.1,\ 0.2,\ 0.05,\ 0.1,\ 0.25]$, and $d2 = [0,\ 0.25,\ 0.05,\ 0.1,\ 0.15,\ 0.15,\ 0.3]$. We can easily verify that $d1 \leq_{st} d2$: the probability mass of $d2$ is concentrated in higher states such as the probability cumulative distribution of $d2$ is always below the cumulative distribution of $d1$ (see Fig. 3).

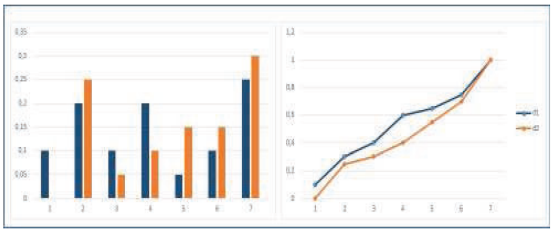

Fig. 3. $d1 \leq_{st} d2$: Their pmf (left) and their cumulative distribution functions (right).

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

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 states 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. $B \rightarrow S$). 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 [3]:

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 CTMCs. As presented in [3], it remains us to define two CTMCs: $\{\widehat{X}(t),\ t \geq 0\}$ and $\{\widehat{Y}(t),\ t \geq 0\}$ governed by the same infinitesimal generator matrix respectively as $\{X(t), t \geq 0\}$, and $\{Y(t), t \geq 0\}$, representing different realizations of these processes with different initial conditions. The following theorem establishes the $\leq_{st}$-comparison using the coupling [3]:

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. Hysteresis system with aggregated bounding arrival process

The bounding model is a hysteresis system equivalent to the exact system, except that the arrival process is defined as follows: the arrivals of bulks follow a Poisson process with the same rate λ, and batches follow a probability distribution

p^u (resp. p^l) for the upper bound (resp. the lower bound). The probability distributions of the batches for the bounds are obtained by aggregations, in order to reduce the size, with the following relation:

$$p \leq_{st} p^u$$

and

$$p^l \leq_{st} p.$$

If p is defined on a state space of size n, then p^u (resp. p^l) are defined on a state space of size m, and $m < n$. Moreover, p^u and p^l are obtained to be the closest distributions with m states, according to an increasing reward function [2]. Intuitively, the probability distribution p^u (resp. p^l) has been obtained by removing some states of p and by adding their probabilities into higher states (resp. lower states). The optimality of bounds proved in [1] helps to obtain the most accurate bounds according to an increasing reward function. We present thereafter a brief description of the bounding reduction algorithm developed in [1].

A. Bounding batch distribution reduction

For a given distribution p defined on $\mathcal{E}$ ($|\mathcal{E}| = n$), and for an increasing reward function r ($r : \mathcal{E} \to \mathbb{R}^+$), we compute bounding distributions p^u and p^l defined respectively on $\mathcal{E}^u$ and $\mathcal{E}^l$ ($|\mathcal{E}^u| = m$, $|\mathcal{E}^l| = m$) ($m < n$), such that:

1) $p^l \leq_{st} p \leq_{st} p^u$,
2) $\sum_{i\in\mathcal{E}} r(i)p(i) - \sum_{i\in\mathcal{E}^l} r(i)p^l(i)$ is minimal among the set of distributions on n states that are stochastically lower than p,
3) $\sum_{i\in\mathcal{E}^u} r(i)p^u(i) - \sum_{i\in\mathcal{E}} r(i)p(i)$ is minimal among the set of distributions on n states that are stochastically upper than p.

The distributions p^u and p^l are the closest bounding distributions defined on m states, according to the reward r. We note that the distributions p^u and p^l are derived from an algorithm based on dynamic programming [2], which guarantee the optimality of the bounds. The problem dealing with a discrete distribution is transformed into a graph theory problem. The set of vertices represents the states of the probability distributions, and the arcs have a weight which represent the error due to the suppression of states. So, the computation of the aggregated bounding distributions is equivalent to compute the path of length m with the minimum cost in the graph. Next, in order to be clearer, we give an example.

Example 2: Let $\mathcal{A} = (\mathbf{A}, p(\mathbf{A}))$ be a discrete distribution with support $\mathbf{A} = \{0, 2, 3, 5, 7\}$ and probability vector $p(\mathbf{A}) = [0.05, 0.3, 0.15, 0.2, 0.3]$. For reward function r defined as follows: $\forall\, a_i \in \mathbf{A}$, $r(a_i) = a_i$, we aim to reduce the state space of $\mathcal{A}$ to $m = 3$ states. The expected reward function of the initial distribution is $R[\mathcal{A}] = \sum_{a_i\in\mathbf{A}} r(a_i)\ p_{\mathbf{A}}(i) = 4.15$. The computation of the optimal upper bound ($\overline{\mathcal{A}}$) corresponds to explore all 3-hops paths from the upper state 7 such that $R[\overline{\mathcal{A}}] - R[\mathcal{A}]$ is minimal (see Figure 4). This can be done by applying the algorithm presented in [1]. The optimal upper bound obtained is $\overline{\mathcal{A}} = (\overline{\mathbf{A}}, p(\overline{\mathbf{A}}))$ with $\overline{\mathbf{A}} = \{2, 5, 7\}$, $p(\overline{\mathbf{A}}) = [0.35, 0.35, 0.3]$ and $R[\overline{\mathcal{A}}] = 4.55$.

We propose now to define from the aggregated bounding batch probability distributions two Markov chains. Let $X^u(t)$ (resp. $X^l(t)$) be the hysteresis system built with the arrival bulk probability distribution p^u (resp. p^l). Next, we will prove that they represent really bounding systems for $X(t)$.

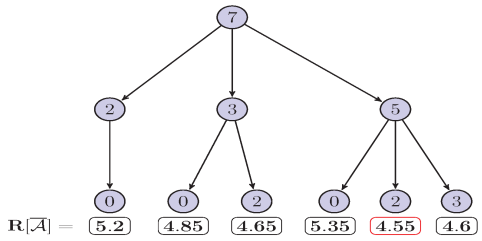

Fig. 4. The tree explored to define the optimal 3 single hops.

B. Stochastic comparison of the systems

We define the many to one mapping function $g : A \to S$, such that $g(x) = x_1$, where $x_1 \in S$ and $S = \{0, \ldots, C\}$. We note that in the state space S, we use the total order $\leq$. We have the following theorem:

Theorem 2: We have the following relations:

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

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, $g(X^u(t)) = y$, and $g(X(t)) = x$. The proof is by induction, so we suppose that the order is verified at time t ($x \leq y$) , and we prove that at time $t+dt$ the order is still verified. We denote by $g(X^u(t+dt)) = y'$, and $g(X(t+dt)) = x'$. 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 $t+dt$, $x' = x+k$, then we can have also a transition from y to y' such that $y' = y+l$, and $k \leq l$, as $p \leq_{st} p^u$. So $x' \leq y'$, and the order is still verified at time $t+dt$.
- services: if we have a service for $X^u(t)$ such that at time $t+dt$, $y' = y-1$, then we can have also a service in $X(t)$ such that at time $t+dt$, we have $x' = x-1$, as the transition rates are the same in the two systems.

■

For the lower bound $\{X^l(t), t \geq 0\}$, the proof is similar, as for the arrivals $p^l \leq_{st} p$, and the service rates are the same, then the second relation of theorem 2 is verified. Note that as the stochastic comparison of the processes is made by the mapping g, then it allows to compare the processes from the number of customers waiting in the system. So, this provides the comparison for performance measures as the expected number of customers waiting in the system, expected departure, blocking probabilities, etc..

Let Π_X be the steady state distribution of $\{X(t), t \geq 0\}$, and Π^u_X (resp. Π^l_X) be the steady state distribution of $\{X^u(t), t \geq 0\}$ (resp. $\{X^l(t), t \geq 0\}$), we have the following propositions.

Proposition 1: $\forall a \in \mathbb{N}^+$ and $\forall x_2 \in \{1, \ldots, K\}$, we have:

$$\sum_{x_1 \geq a} \Pi_X(x_1, x_2) \leq \sum_{x_1 \geq a} \Pi_{X^u}(x_1, x_2), \text{and}$$

$$\sum_{x_1 \geq a} \Pi_{X^l}(x_1, x_2) \leq \sum_{x_1 \geq a} \Pi_X(x_1, x_2).$$

This proposition is deduced from the stochastic comparison of the processes stated in Theorem 2. As the expected buffer length is expressed as follows: $\mathbb{E}[\Pi_X] = \sum_{x_1} \left(\sum_{x_2} \Pi_X(x_1, x_2)\right) \times x_1$, we deduce from the Proposition 1 the following result.

Proposition 2 (Expected buffer length): $\forall (x_1, x_2) \in \mathcal{E}$, we have:

$$\mathbb{E}[\Pi_X] \leq \mathbb{E}[\Pi_{X^u}], \text{ and } \mathbb{E}[\Pi_{X^l}] \leq \mathbb{E}[\Pi_X].$$

We note that from the comparison of the stationary probability distribution given in Proposition 1, and as $r(x_1, x_2) = x_1$ is an increasing function then Proposition 2 is verified.

In the same way, we can deduce the relation between the expected departure of the models. Let $\mathbb{E}[D_X]$ be the expected departure of $\{X(t), t \geq 0\}$ such that: $\forall (x_1, x_2) \in \mathcal{E}$, $\mathbb{E}[D_X] = \mu \sum_{x_1 \leq F_1} \pi(x_1, 1) + \sum_{i=2}^{K-1} i\, \mu \sum_{R_{i-1} < x_1 \leq F_i} \pi(x_1, i) + K\, \mu \sum_{R_{K-1} < x_1 \leq C} \pi(x_1, K)$. And let $\mathbb{E}[D_{X^u}]$ (resp. $\mathbb{E}[D_{X^l}]$) be the expected departure of $\{X^u(t), t \geq 0\}$ (resp. $\{X^l(t), t \geq 0\}$), we have from the Proposition 1 the following relations:

Proposition 3 (Expected departure):

$$\mathbb{E}[D_X] \leq \mathbb{E}[D_{X^u}], \text{ and } \mathbb{E}[D_{X^l}] \leq \mathbb{E}[D_X].$$

We derive also the blocking probability in the system. This metric is computed as follow: $Bp = \sum_{x_2 | x_1 = C} \Pi(x_1, x_2)$.

Proposition 4 (Blocking probabilities):

$$Bp \leq Bp^u, \text{ and } Bp^l \leq Bp.$$

This proposition is also deduced from Proposition 1 and the fact that the reward function r (defined as follows: $\forall x_1, x_2, \quad r(C, x_2) = C$, and $r(x_1, x_2) = 0$ otherwise) is an increasing function.

VI. Numerical examples

We consider a threshold-based queueing system with hysteresis and batch-arrival, such that the distribution of the batch arrivals is randomly generated on a support $\{1, 2, 3, \ldots, 500\}$. Depending on the input parameters, we propose to illustrate in this section some numerical examples which show the relevance and the accuracy of the bounding models presented in the paper.

We present below three examples through which we propose to vary some input parameters as buffer size, arrival rate, and degree of virtualization (number of servers). The studied models are:

- 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 ($X^u(t)$)

We note that to compute the steady state distribution probability vector of the considered models, we use the methodology proposed by Lui and Golubchik in [7] based on the stochastic complementation as it is proved to be less complex than the commonly used solution technique [10].

A. Some performance measures 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 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 size, 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 $C_i = i \times 1000$, the threshold vectors are $F_i = i \times F$ and $R_i = i \times R$. We note that the reductions considered for the batch-arrival distribution, are respectively $bins$=10 and $bins$=50 (*i.e.* the bounding distributions are defined on a support of length 10 and 50).

Depending on the values of the buffer size, the figures 5, 6 and 7 illustrate the expected buffer length, the expected departure and the blocking probabilities. We illustrate also in Figure 8 the computation times in seconds needed to solve the different models.

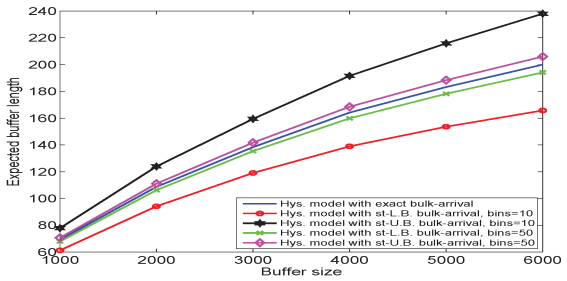

Fig. 5. Expected buffer lengths versus buffer size.

Through these figures, we remark that the aggregating models define a good coverage of the exact result and the accuracy of these bounds are very improved when we increase the number of bins ($bins = 50$). We remark also in Figure 8 that the time needed to compute bounds on performance measures is very short and our approach is largely faster than

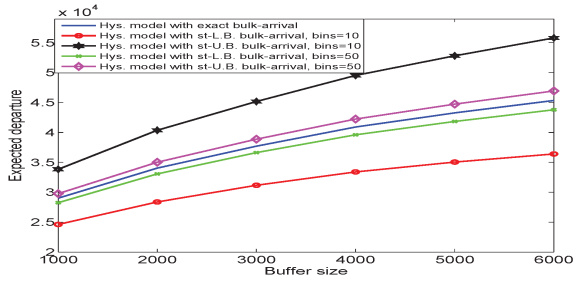

Fig. 6. Expected departures versus buffer size.

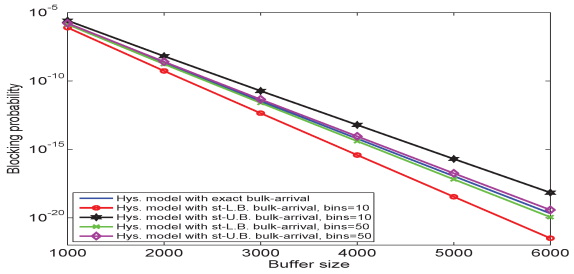

Fig. 7. Blocking probabilities versus buffer size.

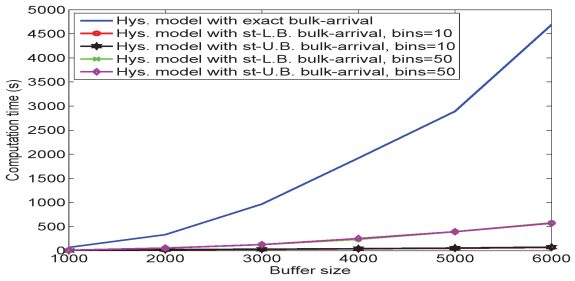

Fig. 8. Computation times (in seconds) versus buffer size.

the exact computation. Indeed, we recall that we derive a stochastic bounds on only 10 or 50 states knowing that the original distribution is defined on 500 states, so considering such reductions, we are forced to admit that the proposed bounding models are very relevant and close to the exact results with very low computational times.

B. Some performance measures versus arrival rate

We suppose here that the buffer size is $C = 1000$, the number of servers, K, is equal to 10, the threshold vectors are $F = (90, 140, 280, 400, 610, 690, 730, 840, 910)$ and $R = (30, 90, 190, 270, 410, 510, 620, 700, 800)$. We set the service rate μ to 1 and we propose to vary the arrival rate λ from $\lambda = 0.1$ to $\lambda = 2.5$. So, we vary the utilization rate of the system from the a lightly loaded system with $\lambda = 0.1$ to a highly loaded system with $\lambda = 2.5$. We note that the size of reduction considered for stochastic bounding distribution are also $bins = 10$ and $bins = 50$.

We depict in figures 9, 10 and 11 the expected buffer length, the expected departure and the blocking probabilities computed for different studied models. We illustrate in Figure 12 the computation times in seconds.

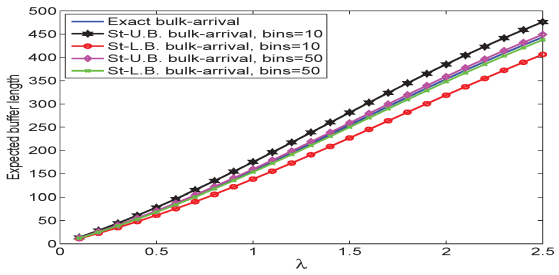

Fig. 9. Expected buffer lengths versus arrival rate.

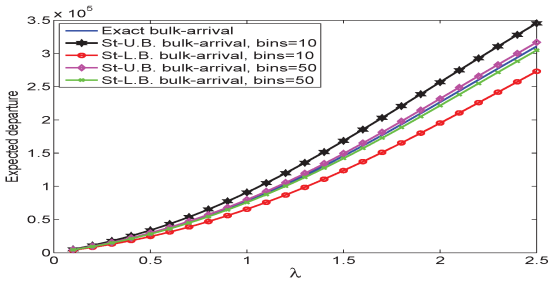

Fig. 10. Expected departures versus arrival rate.

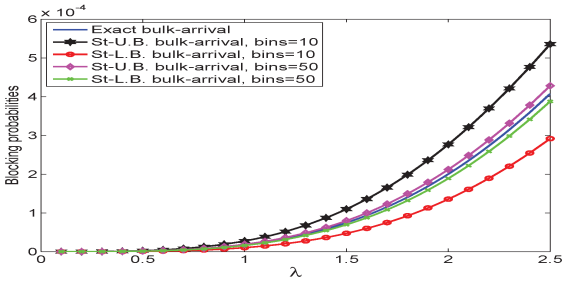

Fig. 11. Blocking probabilities versus arrival rate.

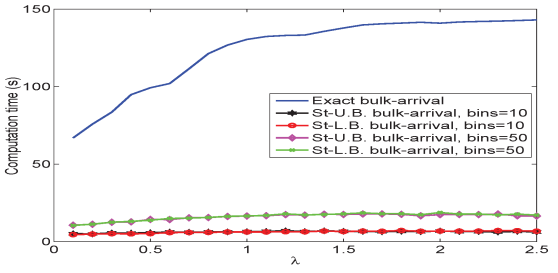

Fig. 12. Computation time (in seconds) versus arrival rate.

From these curves, we see that the bounding results frame the exact results and are very accurate. We observe that the

performance measures obtained form reduction $bins = 50$ are the closest results and are very relevant. We note also that the accuracy of bounds is not degraded when the arrival rate increase. Regarding the computation time (Fig 12), we observe that the fact that the size of the bulk-arrival distribution is aggregated represents an important contribution for decreasing the computation time and also the complexity of the studied model. Moreover the precision of the results is very relevant.

C. Some performance measures versus number of servers

For the third 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 arrival rate is taken equal to 1, and the buffer size is set to $C = 2000$.

We are interested 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: $F = (\lfloor \frac{C}{K} \rfloor, 2 \times \lfloor \frac{C}{K} \rfloor, \ldots, (K-1) \times \lfloor \frac{C}{K} \rfloor)$ and $R_i = F_i - \lfloor \frac{C}{2K} \rfloor$, for $i = 1, \ldots, K-1$.

Thus, depending on the degree of virtualization, the figures 13, 14, 15 and 16 illustrate the expected buffer length, the expected departure, the blocking probabilities and the computation time for the studied models. The reductions considered here are $bins = 10$ and $bins = 50$.

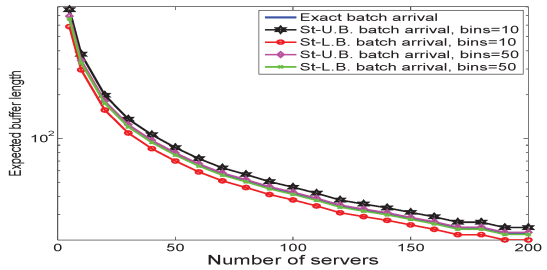

Fig. 13. Expected buffer lengths versus degree of virtualization.

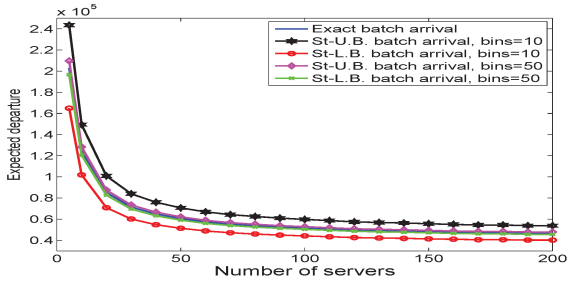

Fig. 14. Expected departures versus degree of virtualization.

From these figures, the observations and the conclusion made before are also ascertained in this example. So, we

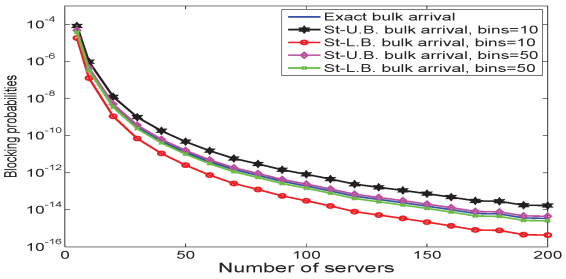

Fig. 15. Blocking probabilities versus degree of virtualization.

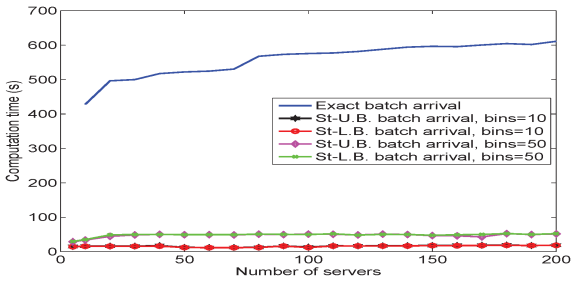

Fig. 16. Computation time (in seconds) versus degree of virtualization.

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 the exact results with considerably reduced computation times. Thus, from these observations, we can say that the approach proposed in this paper offer a very interesting trade-off between accuracy of the results and the computational complexity.

We emphasize however that the number of bins (the size of the reduction) in the bounding distributions is fixed in the algorithm. A good number of bins satisfying the required trade-off between the accuracy of the bounds and the computation time can be determined in an incremental manner: one begins with a reduced number of bins, if the accuracy of bounds is not satisfactory, the number of bins can be incremented. The iteration can be stopped, if the required accuracy is reached and/or the computation time of bounds exceeds a fixed threshold.

VII. CONCLUSION

We propose in this paper to model a data center in a cloud system by a hysteresis queueing system with bulk arrivals process and we derive bounds on performance measures. The interest of hysteresis model is to represent the dynamic behavior of a data center by activating/deactivating the servers (VMs) according to the queue occupancy. However, when the number of VMs and the size of the system increase, we remark that the resolution of this system becomes very cumbersome and difficult. So, to overcome this problem we propose to use the stochastic bounding technique to derive bounds which provide guarantees on performance measures of the system. Hence, through the paper we show clearly that defining bounds on the bulk arrival distribution with smaller size allows to manage the

computational complexity and so provides very relevant results for network dimensioning. Finally, it is important to emphasize that the methodology proposed here offers a good tradeoff between the accuracy of the results and the computational time.

As a future work, we expect to investigate systems with modulated traffic (bursty) as an input process of the queue and also systems with heterogeneous servers (VMs). We also plan to extend our analysis and define optimal thresholds vectors in order to optimize the performance of cloud systems.

REFERENCES

[1] F. Aït-Salaht, J. Cohen, H. Castel-Taleb, J.-M. Fourneau and N. Pekergin, "Accuracy vs. Complexity: the stochastic bound approach". In *11th International Workshop on Disrete Event Systems (WODES2012)*, 2012, pages 343-348.

[2] F. Aït-Salaht, H. Castel-Taleb, J.-M. Fourneau, and N. Pekergin, "Stochastic bounds and histograms for network performance analysis". In *10th European Workshop on Performance Engineering (EPEW'13)*, volume 8168, 2013, pages 13–27. Collection: Lecture Notes in Computer science.

[3] M. Doisy, "Coupling technique for comparison of functions of Markov processes". In *Applied Mathematic Decision Science*, 4, 2000, 131-154.

[4] O.C. Ibe, J. Keilson, "Multi-server threshold queues with hysteresis". In *Performance Evaluation* 21, 1995, 185-213.

[5] H. Khazei, J. Misic, V.B. Misic, "Performance of cloud centers with high degree of virtualization under batch task arrivals". In *IEEE Trans On Parallel and Distributed Systems*, 12,2012.

[6] H. Khazei, J. Misic, V.B. Misic, "A fine- grained performance model of cloud computing centers". In *IEEE Trans On Parallel and Distributed Systems*, Vol.24, n11, 2013.

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

[8] A. Muller, D. Stoyan, "Comparison methods for Stochastic Models and Risks". In *J. Wiley and son in Probability and Statistics*, 2002.

[9] P. Wang, W. Huang, C.A. Varela, "Impact of virtual machine granularity on cloud computing worloads performance". In *11th IEEE/ACM International Conference on Grid Computing*, 2010.

[10] W.J. Stewart, "Introduction to the Numerical Solution of Markov Chains". In *Princeton Press, Princeton, NJ*, 1994.