# Performance analysis of clouds with phase-type arrivals

Farah Ait Salaht, Hind Castel-Taleb

HAL Id: hal-01310845

https://hal.science/hal-01310845

Submitted on 4 May 2016

# Performance analysis of clouds with phase-type arrivals

Farah Ait Salaht [1] and Hind Castel-Taleb [2] [†]

[1] *LIP6, Université Paris Ouest, Nanterre, France*
[2] *SAMOVAR, Télécom SudParis, CNRS, Université Paris Saclay, 9 rue Charles Fourier-91011 Evry Cedex*

We propose to evaluate the performance of a cloud system in terms of buffer occupancy, using mathematical analysis. The client requests (or jobs) arrive by batch, and follow a phase-type process in order to represent the variability of the traffic intensity. The PM (Physical Machine) is modeled by a hysteresis queueing system with phase-type and batch arrivals. To represent the dynamic allocation of the resources, the hysteresis queue activates and deactivates the Virtual Machines (VMs) according to the threshold values of the queue length. This system is represented by a complex Markov chain which is difficult to analyze especially when the size of the state space increases and the length of batch arrival distribution is large. We propose to use stochastic bounds in order to define bounding systems less complex. We derive performance measure bounds as mean buffer length, and blocking probabilities. The relevance of the results is to offer a trade-off between computational complexity and accuracy of the results, providing very interesting solutions in network dimensioning.

**Keywords:** cloud node, performance evaluation, hysteresis queue

## 1  Introduction

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). Performance evaluation of cloud centers is a very crucial research task which becomes difficult due to the dynamic behavior of cloud environments and variability of user demand. In [KMM13], the authors 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. To solve the model, they adopt the technique of embedded Markov chain, and derive exact formulas for different performance measures as blocking probability and mean waiting time of tasks. In this paper, we propose to use a mathematical model in order to evaluate the performance of a PM (Physical Machine) in a data center, in order to compute the mean number of customers in the system, and the blocking probabilities. The considered queueing model is a multi-server with hysteresis threshold queues [IK95]. In this model, each VM is represented by a server and the multi-server queueing model is governed by sequences of forward and reverse thresholds which are different (hysteresis). The forward (resp. the backward) thresholds represent the value of the number of customers from which an additional VM is activated (resp. deactivated). In order to represent the variability of customer requests, we suppose that arrivals are bulks and follow a phase-type process. Each phase can correspond to a traffic intensity, and each request arrival has a batch size defined by a probability distribution. Obviously, the relevance of this model is to represent both the variability of user demand, and dynamicity of resource provisioning. We obtain from the queueing system a multidimensional Markov chain which is difficult to solve, as the size of state space increases with the number of phases, the number of VMs, and the size of bulk arrivals. In order to reduce the complexity of this problem, we propose to derive bounding systems which are obtained by the simplification of the hysteresis model in order to compute bounds on the performance measures [AC15].

## 2   Model 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 jobs into the data centers 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). Different users may share a PM (Physical Machine) using virtualization technique which provides a well defined set of resources (as CPU, RAM, storage). In this paper, we represent the PM ( Physical machine) by a finite capacity system with multi-homogeneous servers (VMs). We denote by $C$ the capacity of the system and by $K$ the number of servers. We suppose a 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 $R_i < F_i, \forall 1 \leq i < K$. The behavior of this system is as follows. We assume that the first VM is always active. If a customer arrives in the system, and finds $F_i$ $(i = 1, \ldots, K-1)$ customers in the system, an additional VM will be activated, and the customer is served by one of the available VMs. When a customer leaves the system with $R_i$ $(i = 1, \ldots, K-1)$ customers, then one VM will be removed from the active VMs. We assume here that the arrival process is a phase-type process with batch arrivals. This process is defined by Poisson arrivals modulated by phases, and with batch size distribution.

The arrivals are Markovian modulated by a phase process with $\ell$ states. We denote by $L = \{1, 2, \ldots, \ell\}$ the set of phase values. The set of phases corresponds to the different traffic intensity (variability or burstiness). As an example, for $L = \{1, 2\}$, we can suppose that traffic intensity in phase 1 is low, and it is higher in phase 2. Let $\{X(t), t \geq 0\}$ be the stochastic process which models the behavior of the hysteresis system. Each state is represented by a 3-tuple $(x_1, x_2, \phi)$, where $x_1$ is the number of customers waiting in the system ($x_1 \in \{0, \ldots, C\}$), $x_2$ is the number of active VMs ($x_2 \in \{1, \ldots, K\}$), and $\phi$ expresses the arrival phase ($\phi \in L = \{1, 2, \ldots, \ell\}$). For the traffic arrivals, we suppose that at each phase $\phi$, the bulk requests arrive according to a Poisson process with rate $\lambda^{(\phi)}$, and bulk size follows a probability distribution $p_\phi = (p_\phi(1), \ldots, p_\phi(k), \ldots, p_\phi(n))$, defined as follows : $p_\phi(k) = \Pr[\text{In phase } \phi, \text{the bulk size is } k] \ \forall k \in E, \forall \phi \in L$, where $E \subset \mathbb{N}$. Servers (or VMs) have an exponential service time distribution with mean rate $\mu_i = \mu$ $(i = 1, \ldots, K)$.

With these assumptions, we deduce that the system $X(t)$ is a Continuous-Time Markov Chain (CTMC). Next, we derive the state evolution equations of $X(t)$. We suppose that the arrivals take place first in the system, and then we have a phase transition. So, from a state $x = (x_1, x_2, \phi)$, if an arrival occurs, we have a transition with the rate $\lambda_\phi p_\phi(k) \mathbf{M}(\phi, \phi')$. This rate is obtained by multiplying $\lambda_\phi p_\phi(k)$ by the probability $\mathbf{M}(\phi, \phi')$ of the phase transition (so, if $\phi' = \phi$, we stay in the same phase, otherwise we have a phase transition). The evolution equations of $X(t)$ are defined for $i, j = 1, \ldots, K-1$, as follows (where $\phi, \phi' \in L$) :

$$
\begin{aligned}
(x_1, x_2, \phi) \quad \rightarrow \quad & (\min\{C, x_1 + k\}, x_2, \phi'), \\
& \text{with rate } \lambda_\phi p_\phi(k) \mathbf{M}(\phi, \phi'), \text{if } (x_1 + k) \leq F_j, \text{ and } x_2 = j, \ j = 1, \ldots, K-1, \\
\rightarrow \quad & (\min\{C, x_1 + k\}, K, \phi'), \\
& \text{with rate } \lambda_\phi p_\phi(k) \mathbf{M}(\phi, \phi'), \text{ if } x_2 = K \text{ or } (x_1 + k) > F_{K-1}, \\
\rightarrow \quad & (\min\{C, x_1 + k\}, l, \phi'), \text{ with rate } \lambda_\phi p_\phi(k) \mathbf{M}(\phi, \phi'), \text{ if } l = \min\{h | (x_1 + k \leq F_h) \text{ and } x_2 + 1 \leq h \leq K-1\}, \\
\rightarrow \quad & (\max\{0, x_1 - 1\}, x_2, \phi), \text{ with rate } x_2\mu, \text{ if } (x_1 \neq R_i + 1 \text{ or } (x_1 = R_i + 1 \text{ and } x_2 \neq i + 1)) \\
\rightarrow \quad & (\max\{0, x_1 - 1\}, \max\{0, x_2 - 1\}, \phi), \text{ with rate } x_2\mu, \text{ if } x_1 = R_i + 1, \text{ and } x_2 = i + 1.
\end{aligned}
$$

We propose in this paper to define bounding models instead of performing an exact resolution of the system which can be often very difficult. The relevance and the key idea of using bounds is first to reduce the state space size from which we compute performance bounds in order to frame the exact values.

## 3   Bounding systems

Considering the same forward and the reverse thresholds 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 $(R_1, R_2, \ldots, R_{K-1})$

for the forward and the reverse thresholds. The main advantage of these systems is that they are easier to solve as states are represented by only two components : the number of customers in the system, and the phase. 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 (where $\phi, \phi' \in L$) :

$$(x, \phi) \quad \rightarrow \quad \min((C, x_1 + k), \phi'), \text{ with rate } \lambda_\phi p_\phi(k) \mathbf{M}(\phi, \phi'), \quad \forall k \in E \tag{1}$$
$$\rightarrow \quad \max((0, x_1 - 1), \phi), \text{ with rates :}$$
$$\bullet \; i\mu, \text{ if } F_{i-1} < x_1 \le F_i, \; \forall i = 1 \ldots K - 1 \tag{2}$$
$$\bullet \; K\mu, \text{ if } F_{K-1} < x_1 \le C \tag{3}$$

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 ((1)-(2)-(3)) are also available by changing the sequence $F_{i, i=1...K-1}$, by the sequence $R_{i, i=1...K-1}$. We can prove using the stochastic comparisons that $Y(t)$ (resp. $Z(t)$) represents an upper bound (resp. a lower bound) for $X(t)$. The proof is similar than in [AC15], except that we compare the processes from states in the same phase, and from the number of customers. We can see that the systems have the same arrival processes, but $Y(t)$ (resp. $Z(t)$) has lower (resp. upper) service rates than $X(t)$. The relevance of this result, is that we can derive performance measures bounds, with less computational times.

# 4 Numerical results

We consider a threshold-based queue with hysteresis, two phases input process and 50 homogeneous servers. For the input traffic, we consider a trace generated randomly with 1 million batch arrivals. We note that the minimum batch size in this trace is 64 and the maximum is 1518. From this trace, we propose to distinguish two phases, phase 1 corresponds to heavy traffic and phase 2 to a low traffic. So, we consider that for batch size less than 200 the arrival phase is 1 otherwise the arrival phase is 2. We note that the size of the support for the defined bulk phases distributions are : 136 states for phase 1 and 730 for phase 2. The probability transition matrix for phase modulation is defined as follows :

$$\mathbf{M}(i, j) = \frac{\text{number of transitions from phase } i \text{ to phase } j}{\text{number of packet in phase } i}, \tag{4}$$

for all $i, j \in L$. Therefore, the resulting transition matrix of phases $\mathbf{M}$ is :

$$\mathbf{M} = \begin{pmatrix} 0.8314 & 0.1686 \\ 0.4086 & 0.5914 \end{pmatrix},$$

In order to analyze this model, we consider a queue with service rate of 175 Kbatches/s, arrival rates $\lambda_1 = 200$ batches/sec and $\lambda_2 = 50$ batches/sec corresponding to an utilization of 94% of the system, and we vary the buffer size from 500 batches to 10 Kbatches. Depending on the value of the buffer ($C$), we present in the tables 1, 2 the blocking probabilities and the expected buffer length obtained for the different studied models. We report also in the Table 3 the computation times needed to resolve these models. Note that different bounding systems are presented : $Y(t)$, $Z(t)$, $Y^u(t)$, and $Z^l(t)$. $Y^u(t)$ (resp. $Z^l(t)$) is an upper (resp. lower) bound for $Y(t)$ (resp. $Z(t)$), it is generated from it by aggregating the batch probability distribution in order to obtain the closest probability distribution bound with a reduced size[ACFP13]. For this example, the size of bounding aggregated batch probability distributions is 10 for each phase.

From Table 1 and Table 2, we remark that the proposed models ($Z(t)$, $Z^l(t)$, $Y(t)$ and $Y^u(t)$) provide clearly bounds on performance measures of the original system $X(t)$ (lower bounds for the models $Z(t)$ and $Z^l(t)$ and upper bounds for the models $Y(t)$ and $Y^u(t)$). We can also see that the results provided by our models are very accurate (the bounds are close to the exact results). Regarding to the execution times (Table 3), we observe that the bounding models are less complex than the original one, particularly for the models with an aggregation of the batch arrival distributions. Indeed, we notice that the models $Z^l(t)$ and $Y^u(t)$ allows to reduce the computational time of the model $X(t)$ by 4, 6 and even by 15 depending on the

| $C$ | $X(t)$ | $Z(t)$ | $Z^l(t)$ | $Y(t)$ | $Y^u(t)$ |
|------|----------|----------|-----------|----------|-----------|
| 500 | 4.9712$e$-5 | 4.9546$e$-5 | 4.9529$e$-5 | 4.9716$e$-5 | 4.9737$e$-5 |
| 1000 | 4.7561$e$-5 | 4.7267$e$-5 | 4.7249$e$-5 | 4.7566$e$-5 | 4.7644$e$-5 |
| 2000 | 4.3482$e$-5 | 4.3207$e$-5 | 4.3115$e$-5 | 4.3492$e$-5 | 4.3603$e$-5 |
| 5000 | 2.00584$e$-5 | 2.0058$e$-5 | 2.0009$e$-5 | 2.0074$e$-5 | 2.0176$e$-5 |

**TABLE 1:** Blocking probabilities versus buffer size.

| $C$ | $X(t)$ | $Z(t)$ | $Z^l(t)$ | $Y(t)$ | $Y^u(t)$ |
|------|----------|----------|-----------|----------|-----------|
| 500 | 462.327 | 462.309 | 462.21 | 463.279 | 463.366 |
| 1000 | 924.361 | 924.305 | 923.962 | 926.542 | 926.699 |
| 2000 | 1854.78 | 1854.58 | 1853.67 | 1857.48 | 1858.37 |
| 5000 | 4009.08 | 4008.95 | 4001.55 | 4009.8 | 4017.87 |

**TABLE 2:** Expected buffer length versus buffer size.

| $C$ | $X(t)$ | $Z(t)$ | $Z^l(t)$ | $Y(t)$ | $Y^u(t)$ |
|------|--------|--------|----------|--------|----------|
| 500 | 439 | 327 | 89 | 322 | 91 |
| 1000 | 749 | 547 | 185 | 542 | 183 |
| 2000 | 1721 | 1431 | 229 | 1267 | 256 |
| 5000 | 7069 | 5611 | 447 | 5518 | 402 |

**TABLE 3:** Computation times (seconds) versus buffer size.

increasing size of the buffer. To conclude, we can say that the exact performance metrics of the original model can be bounded using the proposed models with relatively small computation complexity. And, we recall also that the purpose of using these bounding models consists to offer to the user an interesting trade-off between accuracy of the results and the computational complexity in order to satisfy the required QoS constraints.

## Références

[AC15]    F. Aït-Salaht and H. Castel-Taleb. Stochastic bounding models for performance analysis of clouds. In *The 15th IEEE International Conference on Computer and Information Technology (CIT-2015), 26-28 October 2015, Liverpool, UK*, pages 603–610. IEEE, 2015.

[ACFP13]  F. Aït-Salaht, H. Castel-Taleb, J.-M. Fourneau, and N. Pekergin. Stochastic bounds and histograms for network performance analysis. In *Computer Performance Engineering - 10th European Workshop, EPEW Italy*, volume 8168 of *Lecture Notes in Computer Science*, pages 13–27. Springer, 2013.

[IK95]    O. C. Ibe and J. Keilson. Multi-server threshold queues with hysteresis. *Perform. Eval.*, 21(3) :185–213, 1995.

[KMM13]   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.*, 24(12) :2429–2438, 2013.