# A simple analysis of heterogeneous multi-server threshold queues with hysteresis

L.M. Le Ny and B. Tuffin
IRISA
Campus universitaire de Beaulieu
35042 Rennes cedex, France
{leny,btuffin}@irisa.fr

## Abstract

We analyze a heterogeneous $K$-server queueing system with finite or infinite capacity and congestion control based on thresholds with hysteresis. A set of $K-1$ forward thresholds $(n_1, n_2, \cdots, n_{K-1})$ and a set of $K-1$ reverse thresholds $(m_1, m_2, \cdots, m_{K-1})$ are used as follows. For $i = 1, \cdots, K-1$, when the number of customers in the system goes beyond the forward threshold $n_i$, a server is immediately added to the system. Similarly, when the number of customers falls below the reverse threshold $m_i$, a server is removed from the system. We propose a direct method based on cuts which gives a closed-form for the stationary occupancy distribution. This closed-form was previously obtained for the homogeneous multi-server case and for the heterogeneous case with only two servers. Moreover, an interesting peculiarity of our method is that it uses much simpler mathematics than in the literature.

## 1 Introduction

We consider a heterogeneous multi-server queue with hysteresis. This model, at least in its homogeneous version, has already been described and studied by Ibe and Keilson in [1] and Lui and Golubchik in [2]. As in these two papers, we consider a queue with Poisson arrivals and exponential service times. the number of active servers is defined in connection with thresholds. More precisely, we have a sequence of forward thresholds $(n_1, \cdots, n_{K-1})$ and a sequence of reverse thresholds $(m_1, \cdots, m_{K-1})$ in order to reduce congestion. When an arriving customer finds the system empty, only one server is activated until the queue occupancy reaches the threshold $n_1$. If the number of customers goes beyond $n_1$, another server is activated until $n_2$ is hit, and so on up to $n_{K-1}$. The maximum number of active servers is then $K$. Similarly, the reverse thresholds are used to retire servers. The hysteresis property is linked to the fact that the two sequences $(n_1, \cdots, n_{K-1})$ and $(m_1, \cdots, m_{K-1})$ are different in order to reduce the cost due to frequent switching between backup servers.

The motivation for this kind of queueing model lies in its application to the dynamic routing problem in computer systems or communication networks (where for instance a server is a communication line and the service time is the transmission time) and also to the design of these systems and networks [3]. Let us now give a quick survey of the literature on threshold-based queueing systems. In [4], Larsen and Agrawala conjecture that a threshold-policy optimizes system performance. Lin and Kumar [5] prove this conjecture to be correct but their maximum number of servers is 2 and they do not give the steady state probabilities of the queue. In their conclusion, they suspect that it is difficult to generalize this result to the K-server queue with $K \geq 3$. Le Ny [6] obtains a closed form of the steady-state probabilities of a queue with two thresholds and hysteresis. Morrison [7] studies a two server queue with a threshold and focused on the equilibrium density of the waiting time. Finally, we quote two papers which are more closely related to our work on multi-server queues with hysteresis. In [1], Ibe and Keilson give a closed form for the occupancy distribution using the Green's function method when the servers are identical or when there are only two non-identical servers. Their analysis leads to a formula for the occupancy distribution. In [2], Lui and Golubchik consider also a $K$-server threshold-based queueing system with hysteresis. They use the so called method of stochastic complement analysis for uncoupling Markov chains. They give an explicit algorithm for computing the steady-state probabilities in the heterogeneous case but no compact closed-form.

In our work, we obtain a closed form for the steady-state probabilities of a heterogeneous $K$-server queueing system via the techniques of cuts. To our knowledge, this has never been done when $K > 2$. We then find again the closed-form of [1, 2] for the special case of identical servers. A very interesting point is that our results are obtained using much simpler mathematics than the ones used in [1, 2].

Let us now recall the content of the cuts method (see [8, page 8]), which is a generalization of the classical notion of flow balance equation. Consider an ergodic homogeneous Markov chain in continuous time whose state space is denoted by $\mathcal{S}$. Then, in equilibrium and for every partition $(A, \bar{A})$ of $\mathcal{S}$, the flow rate of probability from $A$ into $\bar{A}$ is equal to the flow rate of probability from $\bar{A}$ to $A$. Formally, if $\pi_i$ is the steady-state probability of being in state $i \in \mathcal{S}$ and $\alpha_{ij}$ is the transition rate from state $i$ to state $j$ we have

$$\sum_{j\in\bar{A}}\sum_{i\in A}\pi_i\alpha_{ij} = \sum_{j\in A}\sum_{i\in\bar{A}}\pi_i\alpha_{ij}.$$

The layout of this paper is the following. In Section 2 we describe the model and give the closed-form for the steady-state probabilities as well as their reduction in the homogeneous case. Section 3 is devoted to the proof of the result. In Section 3.1 we describe the cuts which give the relations between probabilities. The proof is given in Section 3.2.

## 2 Model and result

We consider a heterogeneous $K$-server system as in [1, 2] where the number of servers employed at any time is governed by a forward threshold sequence $(n_1, \cdots, n_{K-1})$ and a reverse threshold sequence $(m_1, \cdots, m_{K-1})$ such that, without loss of generality, $n_k < n_{k+1}$ and $m_k < m_{k+1}$ for all $1 \le k \le K-2$. We also assume that $m_{k+1} < n_k\ \forall k$ even if our method extends to this case as well. We assume that all servers are non-identical and that the service times are exponentially distributed with distinct rates. Moreover we assume that inter-arrivals of customers are also exponentially distributed, with rate $\lambda$.

The system behaves as follows. When the system is empty and a customer arrives, then server one is activated. If the number of busy servers is $k$ ($1 \le k \le K-1$) and the number of customers in the system goes beyond the forward threshold $n_k$, server $k+1$ is immediately added to the system. In the same way, if the number of busy server is $k$ ($2 \le k \le K$) and the number of customers in the system reaches $m_{k-1}$, server $k$ is immediately removed from the system. Note that customer switching (when a server completes a service before server $k$) is not a problem for we have assumed that the activation and disactivation of server is an instantaneous operation. Returning to the notations, for $1 \le k \le K$, let $\mu_k$ be the service rate of the system when the number of active servers is $k$ (which basically represent the sum of service rates of active servers). Define then $\rho_k = \lambda/\mu_k$.

The system is modelled by a Markov chain defined over the state space $\mathcal{S} = \{(m,k) | m = 0 \text{ if } k = 0,\ m_{k-1}+1 \le m \le n_k \text{ if } 1 \le k \le K\}$ where we define $m_0 = 0$ and where $n_K$ is the capacity of the buffer (which may be infinite). $k$ is the number of busy servers and $m$ the number of customers in the queue. Figure 1 describes the state space and its associated transitions when $K = 3$. The generalized case can be easily derived.

The following theorem gives the steady state probabilities of the model.

**Theorem 2.1** *Let* $\alpha_1 = \dfrac{\rho_1^{n_1}(1-\rho_1)}{1-\rho_1^{n_1-m_1+1}}$ *and* $\alpha_k$ *for* $2 \le k \le K$ *be defined by*

$$\alpha_k = \alpha_1 \prod_{j=2}^{k} \rho_j^{n_j - n_{j-1} - 1} \frac{1-\rho_j^{n_{j-1}-m_{j-1}+1}}{1-\rho_j^{n_j - m_j + 1}}.$$

*Assuming that* $\rho_k \ne 1\ \forall 1 \le k \le K$*, we have, at level 1,*

- $\forall 0 \le m \le m_1,\ \pi_{m,1} = \rho_1^m \pi_{0,0}$
- $\forall m > m_1,\ \pi_{m,1} = \left(\rho_1^m - \frac{\rho_1^{n_1+1}(1-\rho_1^{m-m_1})}{1-\rho_1^{n_1-m_1+1}}\right)\pi_{0,0}.$

$\forall 2 \le k \le K-1,$

- $\pi_{m,k} = \alpha_{k-1}\left(\prod_{j=2}^{k}\rho_j\right)\pi_{0,0}$ *if* $m = m_{k-1}+1$
- $\pi_{m,k} = \alpha_{k-1}\dfrac{\left(1-\rho_k^{m-m_{k-1}}\right)}{(1-\rho_k)}\left(\prod_{j=2}^{k}\rho_j\right)\pi_{0,0}$ *if* $m_{k-1}+2 \le m \le m_k$
- $\pi_{m,k} = \left(\alpha_{k-1}\left(1-\rho_k^{m-m_{k-1}}\right) - \alpha_k\rho_k\left(1-\rho_k^{m-m_k}\right)\right) \times \frac{1}{1-\rho_k}\left(\prod_{j=2}^{k}\rho_j\right)\pi_{0,0}$ *if* $m_k+1 \le m \le n_{k-1}+1$
- $\pi_{m,k} = \dfrac{1}{1-\rho_k}\left(\prod_{j=2}^{k}\rho_j\right)\pi_{0,0}\left(\alpha_{k-1}\rho_k^{m-n_{k-1}-1} \times \left(1-\rho_k^{n_{k-1}-m_{k-1}+1}\right) - \alpha_k\rho_k\left(1-\rho_k^{m-m_k}\right)\right)$ *if* $n_{k-1}+2 \le m \le n_k$

*At level* $K$*,*

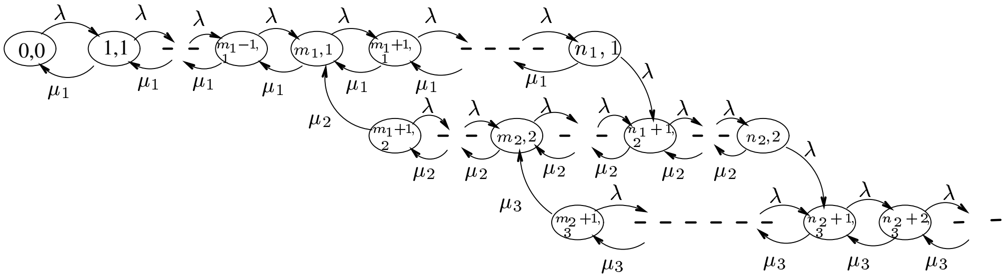

Figure 1: Model when K=3

- $\pi_{m_{K-1}+1,K} = \alpha_{K-1} \left( \prod_{j=2}^{K} \rho_j \right) \pi_{0,0}$

- $\pi_{m,K} = \alpha_{K-1} \frac{1-\rho_K^{m-m_{K-1}}}{1-\rho_K} \left( \prod_{j=2}^{K} \rho_j \right) \pi_{0,0}\ \forall m_K + 2 \leq m \leq n_{K-1}+1$

- $\pi_{m,K} = \alpha_{K-1} \frac{\rho_K^{m-n_{K-1}-1}\left(1-\rho_K^{n_{K-1}+1-m_{K-1}}\right)}{1-\rho_K} \times \left( \prod_{j=2}^{K} \rho_j \right) \pi_{0,0}\ \forall m \geq n_{K-1}+2.$

$\pi_{0,0}$ *is obtained by using the equation* $\sum_{i,j} \pi_{i,j} = 1$.

The proof of this theorem is given in Section 3. Of course, the case of infinite capacity queue is only valid if $\rho_K < 1$ (this is the ergodicity condition).

The case where $\rho_k = 1$ for some $k$ ($1 \leq k \leq K$) can be similarly derived using the same sketch of proof or from our result using L'Hôpital's rule. We omit it for clarity of presentation.

The homogeneous case ($K$ identical servers), is a special case of the model described above with $\mu_k = k\mu\ \forall k$ where $\mu$ is the service rate of each server (the reader can check that the result is the same as the one of [1, 2]. Note also that the different formulaes given in Theorem 2 of [1] following the values of $\rho = \lambda/\mu$ are actually identical after simplification).

# 3 Proof of Theorem 1 using cuts

## 3.1 The cuts and the equations

### 3.1.1 Cuts $\mathcal{A}_m$

If we cut as indicated in Figure 2, we obtain, denoting state $(0,0)$ by $(0,1)$, $\mu_1 \pi_{m+1,1} = \lambda \pi_{m,1}$, i.e.,

$$\pi_{m+1,1} = \rho_1 \pi_{m,1} \text{ for } 0 \leq m \leq m_1 - 1. \tag{1}$$

### 3.1.2 Cuts $\mathcal{B}_m$

If we cut as indicated in Figure 3, we obtain $\mu_K \pi_{m+1,K} = \lambda \pi_{m,K}$, i.e.,

$$\pi_{m+1,K} = \rho_K \pi_{m,K} \text{ for } m \geq n_{K-1} + 1. \tag{2}$$

### 3.1.3 Cuts $\mathcal{C}_{k,m}$

Let $2 \leq k \leq K-1$. If we cut as indicated in Figure 4, we obtain, if $m_{k-1}+1 \leq m \leq m_k - 1$,

$$\mu_k \pi_{m+1,k} = \mu_k \pi_{m_{k-1}+1,k} + \lambda \pi_{m,k},$$

$$\mu_k \pi_{m+1,k} + \mu_{k+1} \pi_{m_k+1,k+1} = \mu_k \pi_{m_{k-1}+1,k} + \lambda \pi_{m,k}$$

if $m_k \leq m \leq n_{k-1}$ and $\mu_k \pi_{m+1,k} + \mu_{k+1} \pi_{m_k+1,k+1} + \lambda \pi_{n_{k-1},k-1} = \mu_k \pi_{m_{k-1}+1,k} + \lambda \pi_{m,k}$ if $n_{k-1}+1 \leq m \leq n_k - 1$, respectively for $m_{k-1}+1 \leq m \leq m_k - 1$, then $m_k \leq m \leq n_{k-1}$, and finally $n_{k-1}+1 \leq m \leq n_k - 1$, i.e.,

$$\pi_{m+1,k} = \pi_{m_{k-1}+1,k} + \rho_k \pi_{m,k} \tag{3}$$

$$\pi_{m+1,k} = \pi_{m_{k-1}+1,k} + \rho_k \pi_{m,k} - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1} \tag{4}$$

$$\pi_{m+1,k} = \pi_{m_{k-1}+1,k} + \rho_k \pi_{m,k} - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1} - \rho_k \pi_{n_{k-1},k-1}. \tag{5}$$

### 3.1.4 Cuts $\mathcal{D}_k$

If we cut as indicated in Figure 5, we obtain $\mu_{k+1} \pi_{m_k+1,k+1} = \lambda \pi_{n_k,k}$, i.e.,

$$\pi_{n_k,k} = \frac{1}{\rho_{k+1}} \pi_{m_k+1,k+1} \text{ for } 1 \leq k \leq K-1. \tag{6}$$

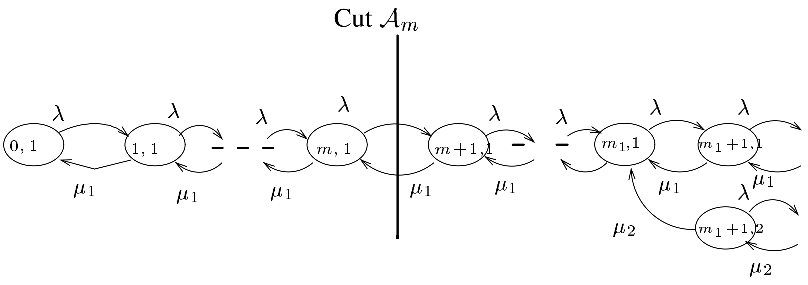


Figure 2: Cut $\mathcal{A}_m$

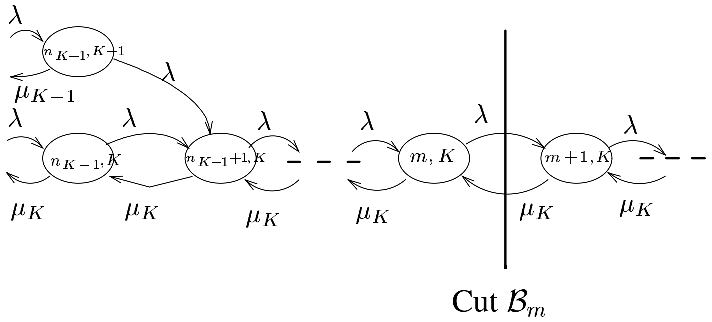


Figure 3: Cut $\mathcal{B}_m$

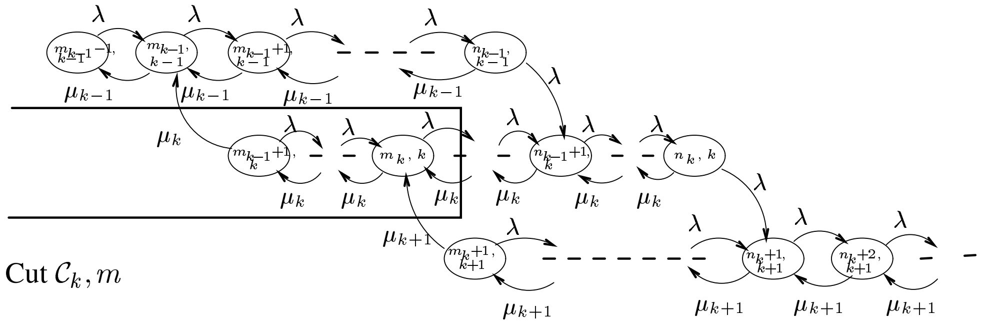


Figure 4: Cut $\mathcal{C}_{k,m}$

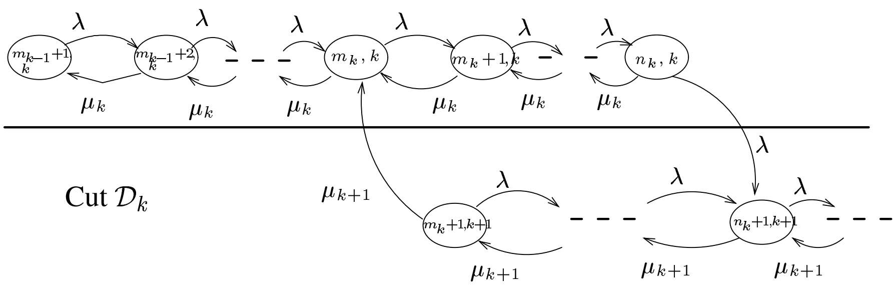


Figure 5: Cut $\mathcal{D}_k$

Using it, Equation (5) becomes, if $n_{k-1}+1 \leq m \leq n_k - 1$,

$$\pi_{m+1,k} = \rho_k \pi_{m,k} - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1}, \tag{7}$$

#### 3.1.5 Cuts $\mathcal{E}_m$

If we cut as indicated in Figure 6, we obtain

$$\mu_2 \pi_{m_1+1,2} + \mu_1 \pi_{m+1,1} = \lambda \pi_{m,1} \text{ for } m_1 \leq m \leq n_1 - 1,$$

i.e., $\forall m_1 \leq m \leq n_1 - 1$,

$$\pi_{m+1,1} = \rho_1 \pi_{m,1} - \frac{\rho_1}{\rho_2} \pi_{m_1+1,2} \tag{8}$$

## 3.2 The proof

The algorithm to compute the steady-state probabilities is the following:

1. Determine $\pi_{m,1}$ in terms of $\pi_{0,0}$ for $1 \leq m \leq m_1$ using equation (1).
2. Determine $\pi_{m,1}$ for $m_1 + 1 \leq m \leq n_1$ and $\pi_{m_1+1,2}$, all in terms of $\pi_{0,0}$, by solving the system of equations given by (8) and (6).
3. By induction, on the level $k$ ($2 \leq k \leq K-1$), we use relations (3), (4), (7) and (6) to determine $\pi_{m,k}$ ($m_{k-1}+2 \leq m \leq n_k$) and $\pi_{m_k+1,k+1}$ in terms of $\pi_{0,0}$.
4. At level $K$, we use relation (3) to determine $\pi_{m,K}$ ($m_{K-1}+2 \leq m \leq n_{K-1}+1$) in terms of $\pi_{0,0}$.
5. Finally, we determine $\pi_{m,K}$ in terms of $\pi_{0,0}$ for $m \geq n_{K-1}+2$ using equation (2).
6. Compute $\pi_{0,0}$ using $\sum_{i,j} \pi_{i,j} = 1$ and we obtain then all the steady-state probabilities.

Now, let us detail the proof.

1. Using recursively Equation (1), we have $\pi_{m,1} = \rho_1^m \pi_{0,0}$ for $0 \leq m \leq m_1$.
2. By induction using Equation (8), we have $\forall m > m_1$

$$\begin{aligned}\pi_{m,1} &= \rho_1^{m-m_1} \pi_{m_1,1} - \frac{\rho_1}{\rho_2} \pi_{m_1+1,2} \sum_{k=0}^{m-m_1-1} \rho_1^k \\ &= \rho_1^{m-m_1} \pi_{m_1,1} - \frac{\rho_1}{\rho_2} \pi_{m_1+1,2} \frac{1-\rho_1^{m-m_1}}{1-\rho_1}.\end{aligned}$$

In particular,

$$\pi_{n_1,1} = \rho_1^{n_1-m_1} \pi_{m_1,1} - \frac{\rho_1}{\rho_2} \pi_{m_1+1,2} \frac{1-\rho_1^{n_1-m_1}}{1-\rho_1}.$$

Combining this with Equation (6) with $k = 1$, i.e., $\pi_{n_1,1} = \pi_{m_1+1,2}/\rho_2$, we obtain

$$\begin{aligned}\pi_{m_1+1,2} &= \frac{\rho_2 \rho_1^{n_1-m_1}(1-\rho_1)}{(1-\rho_1^{n_1-m_1+1})} \pi_{m_1,1} \\ &= \frac{\rho_2 \rho_1^{n_1}(1-\rho_1)}{(1-\rho_1^{n_1-m_1+1})} \pi_{0,0}\end{aligned} \tag{9}$$

and then $\forall m > m_1$, we obtain the result for $\pi_{m,1}$.

3. From (3), we get by induction, $\forall m_{k-1}+2 \leq m \leq m_k$,

$$\pi_{m,k} = \frac{1-\rho_k^{m-m_{k-1}}}{1-\rho_k} \pi_{m_{k-1}+1,k} \tag{10}$$

Next, using (4) and (10), we get still by induction that $\forall m_k + 1 \leq m \leq n_{k-1}+1$

$$\begin{aligned}\pi_{m,k} &= \pi_{m_{k-1}+1,k} \sum_{i=0}^{m-m_k-1} \rho_k^i + \pi_{m_k,k} \rho_k^{m-m_k} \\ &\quad - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1} \sum_{i=0}^{m-m_k-1} \rho_k^i \\ &= \pi_{m_{k-1}+1,k} \frac{1-\rho_k^{m-m_k}}{1-\rho_k} + \pi_{m_k,k} \rho_k^{m-m_k} \\ &\quad - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1} \frac{1-\rho_k^{m-m_k}}{1-\rho_k} \\ &= \pi_{m_{k-1}+1,k} \frac{1-\rho_k^{m-m_{k-1}}}{1-\rho_k} \\ &\quad - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1} \frac{1-\rho_k^{m-m_k}}{1-\rho_k}\end{aligned} \tag{11}$$

and using (7) and (11), we get $\forall n_{k-1}+2 \leq m \leq n_k$

$$\begin{aligned}\pi_{m,k} &= \pi_{n_{k-1}+1,k} \rho_k^{m-n_{k-1}-1} - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1} \sum_{i=0}^{m-n_{k-1}-2} \rho_k^i \\ &= \pi_{m_{k-1}+1,k} \rho_k^{m-n_{k-1}-1} \frac{1-\rho_k^{n_{k-1}-m_{k-1}+1}}{1-\rho_k} \\ &\quad - \frac{\rho_k}{\rho_{k+1}} \pi_{m_k+1,k+1} \frac{1-\rho_k^{m-m_k}}{1-\rho_k}.\end{aligned} \tag{12}$$

Using this formula for $m = n_k$ and (6), i.e., $\pi_{n_k,k} = \dfrac{1}{\rho_{k+1}} \pi_{m_k+1,k+1}$, we obtain $\pi_{m_k+1,k+1} =$

$$\rho_{k+1} \rho_k^{n_k-n_{k-1}-1} \frac{1-\rho_k^{n_{k-1}-m_{k-1}+1}}{1-\rho_k^{n_k-m_k+1}} \pi_{m_{k-1}+1,k}.$$

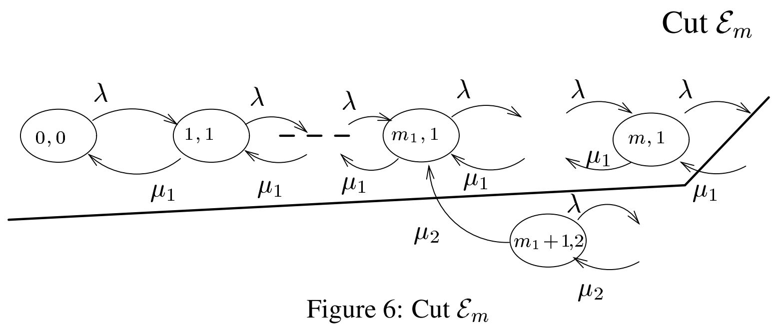


Figure 6: Cut $\mathcal{E}_m$

Using the definitions of $\alpha_k$ from the theorem, we have for each $3 \leq k \leq K$, using (9)

$$\begin{aligned}\pi_{m_{k-1}+1,k} &= \left(\prod_{j=2}^{k} \rho_j^{n_j-n_{j-1}-1} \frac{1-\rho_j^{n_{j-1}-m_{j-1}+1}}{1-\rho_j^{n_j-m_j+1}}\right) \\ &\quad \times \left(\prod_{j=3}^{k} \rho_j\right) \pi_{m_1+1,2} \\ &= \alpha_{k-1} \left(\prod_{j=2}^{k} \rho_j\right) \pi_{0,0}. \end{aligned} \tag{13}$$

Merging (13) with (10), (11) and (12), we get the values of $\pi_{m,k}$ $\forall 2 \leq k \leq K-1$.

4. Using (3), i.e., (10), $\forall m_K + 2 \leq m \leq n_{K-1}+1$,

$$\begin{aligned}\pi_{m,K} &= \frac{1-\rho_K^{m-m_{K-1}}}{1-\rho_K} \pi_{m_{K-1}+1,K} \\ &= \alpha_{K-1} \frac{1-\rho_K^{m-m_{K-1}}}{1-\rho_K} \left(\prod_{j=2}^{K} \rho_j\right) \pi_{0,0}.\end{aligned}$$

5. Finally, using (2), we get $\forall m \geq n_{K-1}+2$

$$\begin{aligned}\pi_{m,K} &= \rho_K^{m-n_{K-1}-1} \pi_{n_{K-1}+1,K} \\ &= \alpha_{K-1} \frac{\rho_K^{m-n_{K-1}-1}\left(1-\rho_K^{n_{K-1}+1-m_{K-1}}\right)}{1-\rho_K}\end{aligned}$$

which gives the result for $\pi_{m,K}$.

6. As the sum of all the terms gives one, we can easily compute $\pi_{0,0}$, this for a finite or infinite capacity.

# References

[1] O.C. Ibe and J. Keilson. Multi-server threshold queues with hysteresis. *Performance Evaluation*, 21:185–213, 1995.

[2] J.C.S. Lui and L. Golubchik. Stochastic complement analysis of multi-server threshold queues with hysteresis. *Performance Evaluation*, 35:185–213, 1999.

[3] R.L. Larsen. *Control of multiple exponential servers with application to computer systems*. PhD thesis, Dep. Comput. Sci., Univ. Maryland, Apr. 1981.

[4] R.L. Larsen and A.K. Agrawala. Control of a Heterogeneous Two-Server Exponential Queueing System. *IEEE Transactions on Software Engineering*, SE-9(4):522–526, 1983.

[5] W. Lin and P.R. Kumar. Optimal Control of a Queueing System with Two Heterogeneous Servers. *IEEE Transactions on Automatic Control*, AC-29(8):696–703, 1984.

[6] L-M. Le Ny. Probabilité stationnaire d'une file d'attente à 2 seuils. Technical Report 160, IRISA, 1987.

[7] J.A. Morrison. Two-server queue with one server idle below a threshold. *Queueing Systems: Theory and Applications*, 7:325–336, 1990.

[8] F.P. Kelly. *Reversibility and Stochastic Networks*. Wiley, Chichester, 1979.

# Biographies

**Bruno Tuffin** is an INRIA-researcher at IRISA, Rennes, France. He received his Ph.D. in applied Mathematics from Rennes 1 University in 1997. His areas of interest include performance evaluation, simulation methods (Monte Carlo, quasi-Monte Carlo...) and their applications in telecommunication.

**Louis-Marie Le Ny** is an Associate Professor in Applied Mathematics at Rennes 1 university (France). He is also part of the ARMOR (Networks architectures and Modeling) group at IRISA laboratory. His research interests include queueing networks analysis and performance evaluation of Internet Protocols.