Guaranteeing Seamless Mobility with User Redials and Automatic Handover Retrials[1]

Jose Manuel Gimenez-Guzman
(Dept. Comunicaciones, Universidad Politecnica de Valencia, Spain
jogiguz@upvnet.upv.es)

Mª Jose Domenech-Benlloch
(Dept. Comunicaciones, Universidad Politecnica de Valencia, Spain
mdoben@doctor.upv.es)

Vicent Pla
(Dept. Comunicaciones, Universidad Politecnica de Valencia, Spain
vpla@dcom.upv.es)

Vicente Casares-Giner
(Dept. Comunicaciones, Universidad Politecnica de Valencia, Spain
vcasares@dcom.upv.es)

Jorge Martinez-Bauset
(Dept. Comunicaciones, Universidad Politecnica de Valencia, Spain
jmartinez@upvnet.upv.es)

Abstract In communication systems that guarantee seamless mobility of users across service areas, repeated attempts occur as a result of user behavior but also as automatic retries of blocked requests. Both phenomena play an important role in the system performance and therefore should not be ignored in its analysis. On the other hand, an exact Markovian model analysis of such systems has proven to be infeasible and resorting to approximate techniques is mandatory. We propose an approximate methodology which substantially improves the accuracy of existing methods with a negligible increase of the computational time from the human point of view. A numerical evaluation of the model is carried out to investigate the impact on performance of the parameters related to the retry phenomena. As a result, some useful guidelines for setting up the automatic retries are provided. Finally, we also show how our model can be used to obtain a tight performance approximation in the case where reattempts have a deterministic nature.

Key Words: seamless mobility, Markovian model, finite truncated model, reattempt
Category: C.4, C.2.0

[1] A preliminar and shorter version of this work was presented at The 7th International Conference on Next Generation Teletraffic and Wired/Wireless Advanced Networking (NEW2AN) 2007.

1 Introduction

The retrial phenomenon appears in multiple situations in telecommunications, computer networking as well as in many other fields. In this paper we focus our attention in a generic communication network that guarantees seamless mobility to its customers by means of a cellular architecture. In these type of networks, the network coverage area is divided into services areas, known as cells, and customers, even those with an active communication, move across different service areas of the network producing handovers. When a customer with an active communication moves from one cell to another a so-called *handover* procedure is executed to allocate the necessary resources in the new cell and release the unused resources in the former cell. Nowadays, perhaps the most widespread and popular example of this type of networks are the telephone cellular networks (2G and 3G) but the current perspective is that in a near future a variety of technologies fitting into this category will be in place, e.g. Mobile IP, IEEE 802.16 (commercially known as WiMAX), which has recently incorporated mobility into the standard [IEEE 802.16] and IEEE 802.20 [Bolton et al. 2007] (Mobile Broadband Wireless Access, MBWA).

The phenomenon of repeated attempts has been studied, at least, since the early 70's [Jonin and Sedol 1970]. However the scenario under study in those works is that of a classical telephone network where the effect of reattempts is due to the customer's behavior. In contrast, this paper deals with the case in which reattempts appear not only when a customer is blocked but also when a handover is blocked. In this paper we refer to the former as *redials* and to the latter as *(automatic) retrials* and to both types together as reattempts. Blocked handovers will be automatically retried until a reattempt succeeds or the user moves outside the handover area. In the former case the session will continue without the user noticing any disruption, while in the latter the session will be abruptly terminated. In contrast, persistence of redials depends on the user patience and an eventual abandonment results in session setup failure. Another difference is that the maximum number of unsuccessful automatic retrials in a row has a distribution, which is generally deterministic [Onur et al. 2002], set by the network operator while redials are affected by the randomness of human behavior. In general, blocking a new session setup is considered to be less harmful than blocking a handover attempt. When the session under consideration is of streaming type, blocking a handover produces an abrupt termination of the ongoing session, which results more annoying from the user perspective than delaying the initiation of a new session. In the case of an elastic traffic session [Bonald and Roberts 2003] the effect of tearing down a session is even worse, as the amount of information transmitted so far is rendered completely useless. Therefore, both types of reattempts have different characteristics and should receive different priority by the admission controller, and as a conse-

quence two separate reattempt pools have to be considered in the analysis of the system.

Although not all the aforementioned technologies include the automatic retrial in their technical specifications, we believe it is an option worth considering in both standardization and network deployment as it reduces considerably the abrupt termination of ongoing sessions. An example of technology that enables automatic handover retrials is GSM. In this paper we give some tools and guidelines for configuring the automatic retrials properly.

To the best of our knowledge, the first and only paper that has considered the effect on network performance of both types of reattempts (retrials and redials) is [Onur et al. 2002]. The mathematical model of the system is of the type of the multiserver retrial queue, for which it is known that an analytical solution is not available [Artalejo and Pozo 2002] and one must resort to approximate models. Among the approximate models for the multiserver retrial queue those based on a generalized truncation [Artalejo and Pozo 2002, Falin 1983, Neuts and Rao 1990] offer a good tradeoff between precision an computational complexity. Unfortunately, in our case a generalized truncation approach cannot be applied since there are two different reattempt orbits (and consequently an infinite and non-homogeneous state space in two dimensions) and therefore we are forced to use an approximation based on finite truncated models, which is expected to be less accurate than generalized truncated models [Artalejo and Pozo 2002]. This type of approximation has already been employed in the context of cellular networks. Marsan et al. [Marsan et al. 2001] consider a cellular network with only customer retrials and propose an approximate technique for its analysis. In [Domenech et al. 2005] a generalization of the approximate method in [Marsan et al. 2001] is proposed for a system with only a single retrial orbit showing a substantial improvement in the accuracy at the expense of an acceptable increase of the computational cost. In this paper we extend the approximation technique of [Domenech et al. 2005] to a system with two different reattempt orbits (redials and retrials). The proposed method is employed to perform a numerical analysis of the system focusing on how redials and retrials impact on the system performance. As a result of the study we give some guidelines for setting up the automatic handover retrial capability. Additionally, we propose an accurate approximation method to analyze the performance of a system when retrials have a deterministic nature, i.e. the maximum number of retrials or the time between consecutive reattempts take fixed values. To the best of our knowledge all previous performance analysis of cellular systems with retrials [Marsan et al. 2001, Tran-Gia and Mandjes 1997, Domenech et al. 2005] assumed that the maximum number of retrials is geometrically distributed and the time between consecutive reattempts is exponentially distributed.

The rest of the paper is structured as follows. [Section 2] describes the system

under study, while [Section 3] discusses the system model and the analysis methodology. In [Section 4] the numerical analysis of the impact of retrials/redials is carried out. Final remarks and a summary of results are provided in [Section 5].

2 System description

We consider a system with C resource units, being the physical meaning of a unit of resources dependent on the specific technological implementation. Although the solution can be extended to multiservice systems increasing the mathematical complexity, we have considered only one service. More specifically, we consider an application that uses the conversational bearer service like a voice session, although it could be extended to an elastic traffic session. Moreover, and without loss of generality, we consider that each user occupies one resource unit. As shown in [Fig. 1] there are two arrival streams: the first one represents new sessions and the second one handovers from adjacent service areas. Both arrivals are considered Poisson processes with rates λ_n and λ_h respectively, being $\lambda = \lambda_n + \lambda_h$. For determining the value of λ_h we consider that the incoming handover stream is equal to the outgoing handover stream, due to the system homogeneity [Marsan et al. 1999]. For the sake of mathematical tractability, the session duration and the residence time are exponentially distributed with rates μ_s and μ_r, respectively. Hence, the channel holding time is also exponentially distributed with rate $\mu = \mu_r + \mu_s$ and the mean number of handovers per session when the number of resources is infinite is $N_H = \mu_r/\mu_s$. Note that, in order to keep the analytical model tractable, we have considered exponential distributions for all the rv that describe time magnitudes. While for some of the involved rv the statistical features may not be properly captured by an exponential distribution, when it comes to the impact on the performance parameters of interest the exponential approximation is shown to be reasonable in a wide range of scenarios [Orlik and Rappaport 2001, Khan and Zeghlache 1997].

Since the loss of a handover request or a retrial is less desirable than the loss of a new session setup or a redial, we must include any kind of admission control policy to guarentee a certain degree of Quality of Service (QoS). The most widespread technique is to reserve some resources to highest priority flows, being in our case handovers and their associated automatic retrials. This technique can be generalized including a fractional reservation, and then, is called Fractional Guard Channel (FGC) admission control policy. The FGC policy is characterized by only one parameter t ($0 \leq t \leq C$). New sessions and redials are accepted with probability 1 when there are less than $L = \lfloor t \rfloor$ resources being used and with probability $f = t - L$, when there are exactly L resources in use. If there are more than L busy resources, new sessions and redials are no longer accepted. Handovers and automatic retrials are only rejected when the system is completely

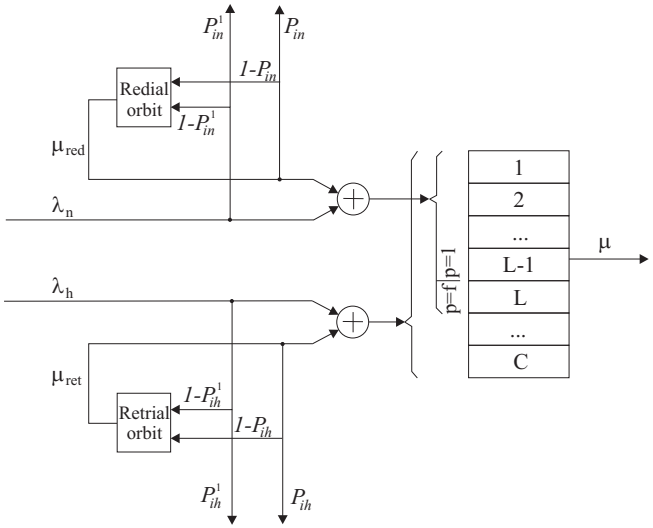


Figure 1: System model.

occupied. Note that to analyze a system in which there is not an admission control algorithm we must make $t = C$.

When an incoming new session is blocked, according to [Fig. 1], it joins the redial orbit with probability $(1 - P_{in}^1)$ or leaves the system with probability P_{in}^1. If a redial is not successful, the session returns to the redial orbit with probability $(1-P_{in})$, redialing after an exponentially distributed time with rate μ_{red}. Redials are able to access to the same resources as the new sessions.

Similarly, P_{ih}^1, P_{ih} and μ_{ret} are the analogous parameters for automatic retrials. Making $P_{ih}^1 = 0$, at least one retrial will be performed. In that case, if the system were so loaded that the probability of a successful retrial could be considered negligible, the time elapsed since the first handover attempt until the system finally gives up and the session is dropped will be a sum of X iid exponential rv of mean μ_{ret}^{-1}. In our model the discrete rv X follows a geometric distribution with mean $1/P_{ih}$, hence the total time from the first attempt until abandonment is described by an exponential rv of rate $\mu'_r = \mu_{ret} P_{ih}$. In the light of the above discussion, our model represents a situation in which the blocked handover requests will keep retrying while the user remains within the handover area, being the sojourn time modeled as an exponential rv of rate μ'_r. In cellular networks, this assumption has been shown to have a low impact on the performance measures of interest [Pla and Casares-Giner 2002].

Transition	Condition	Rate
$(k,m,s) \rightarrow (k+1,m,s)$	$0 \leq k \leq L-1$	λ
	$k = L$	$\lambda_h + f\lambda_n$
	$L < k \leq C$	λ_h
$(k,m,s) \rightarrow (k+1,m,s-1)$	$0 \leq k \leq C-1$	$s\mu_{ret}$
$(k,m,s) \rightarrow (k,m,s-1)$	$k = C$	$s\mu_{ret}P_{ih}$
$(k,m,s) \rightarrow (k+1,m-1,s)$	$0 \leq k \leq L-1$	$m\mu_{red}$
	$k = L$	$m\mu_{red}f$
$(k,m,s) \rightarrow (k,m-1,s)$	$k = L$	$m\mu_{red}(1-f)P_{in}$
	$L < k \leq C$	$m\mu_{red}P_{in}$
$(k,m,s) \rightarrow (k-1,m,s)$	$1 \leq k \leq C$	$k\mu$
$(k,m,s) \rightarrow (k,m,s+1)$	$k = C$	$\lambda_h(1-P_{ih}^1)$
$(k,m,s) \rightarrow (k,m+1,s)$	$k = L$	$\lambda_n(1-P_{in}^1)(1-f)$
	$L < k \leq C$	$\lambda_n(1-P_{in}^1)$

Table 1: Transition rates of the exact model.

3 System model and performance analysis

The model considered can be represented as a tridimensional (k,m,s) Continuous Time Markov Chain (CTMC), being the first dimension (k) the number of sessions being served, the second dimension (m) the number of sessions in the redial orbit and the third dimension (s) the number of sessions in the retrial orbit. The transition rates of this model are represented in [Table 3]. Additionally, in [Fig. 2] we can see the transition diagram. The main mathematical features of this queueing model are the fact of having two infinite dimensions (the state space of the model is $\{0,\ldots,C\} \times \mathbb{Z}_+ \times \mathbb{Z}_+$) and the space-heterogeneity along them. This heterogeneity is produced by the retrial and redial rates, which respectively depend on the number of customers in the retrial and the redial orbits.

It is known that the classical theory (see, e.g., [Neuts 1981]) is developed for random walks on the semi-strip $\{0,\ldots,C\} \times \mathbb{Z}_+$ with infinitesimal transitions subject to conditions of space-homogeneity. When the space-homogeneity condition does not hold the problem of calculating the equilibrium distribution has not been addressed beyond approximate methods [Bright and Taylor 1995, Latouche and Ramaswami 1999]. Indeed, if we focus on the simpler case of multiserver retrial queues (with only one retrial orbit) it can emphasize the absence of closed form solutions for the main performance characteristics when $C > 2$ [Artalejo and Pozo 2002].

As it is clear that in our case it is necessary to resort to approximate models and numerical methods of solution, in [Domenech et al. 2005] we developed a

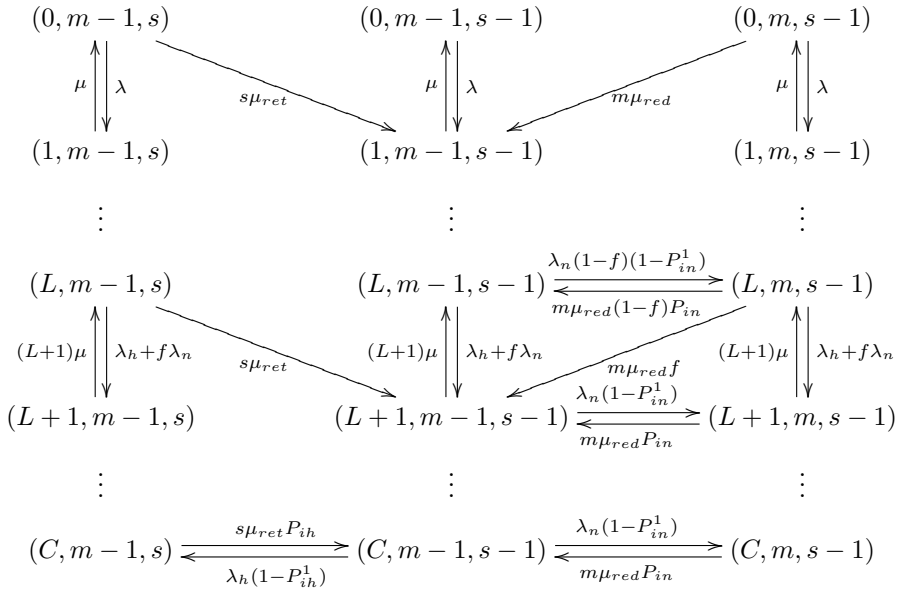


Figure 2: Transition diagram of the exact model.

generalization of the approximation method proposed in [Marsan et al. 2001]. The new methodology can be applied to both retrial and redial orbits, reducing the state space to a finite set by aggregating all states beyond a given occupancy of the orbits: Q_n (Q_h) defines the occupancy from which the states in the redial (retrial) orbit are aggregated. By increasing the values of Q_n and/or Q_h the considered state space in the approximation is enlarged and the accuracy of the solution improves at the expense of a higher computational cost.

Due to that aggregation two new parameters for each orbit are introduced. The parameter M_n denotes the mean number of users in the redial orbit conditioned to those states where there are at least Q_n users in the orbit, i.e. $M_n = E(m|m \geq Q_n)$. The probability that after a successful redial the number of users in the redial orbit does not drop below Q_n is represented by p_n. For the retrial orbit the parameters M_h and p_h are defined analogously.

As a result of the aggregation the state space of the approximate model is $S = \{(k, m, s) : 0 \leq k \leq C; 0 \leq m \leq Q_n; 0 \leq s \leq Q_h\}$ where states of the form $(\cdot, Q_n, \cdot)$ represent the situation where at least Q_n users are in the redial orbit. Likewise the states of the form $(\cdot, \cdot, Q_h)$ represent the situation where at least Q_h users are in the retrial orbit.

The transition rates for the approximate model are shown in [Table 2]. The

Transition	Condition		Rate
$(k,m,s) \rightarrow (k+1,m,s)$	$0 \leq k \leq L-1$	$m < Q_n;\ s < Q_h$	λ
		$m < Q_n;\ s = Q_h$	$\lambda + \beta_h$
		$m = Q_n;\ s < Q_h$	$\lambda + \beta_n$
		$m = Q_n;\ s = Q_h$	$\lambda + \beta_n + \beta_h$
	$k = L$	$m < Q_n;\ s < Q_h$	$\lambda_h + f\lambda_n$
		$m < Q_n;\ s = Q_h$	$\lambda_h + \beta_h + f\lambda_n$
		$m = Q_n;\ s < Q_h$	$\lambda_h + f(\beta_n + \lambda_n)$
		$m = Q_n;\ s = Q_h$	$\lambda_h + \beta_h + f(\beta_n + \lambda_n)$
	$L < k \leq C$	$m < Q_n;\ s < Q_h$	λ_h
		$m < Q_n;\ s = Q_h$	$\lambda_h + \beta_h$
		$m = Q_n;\ s < Q_h$	λ_h
		$m = Q_n;\ s = Q_h$	$\lambda_h + \beta_h$
$(k,m,s) \rightarrow (k+1,m,s-1)$	$0 \leq k \leq C-1$	$1 \leq s \leq Q_h - 1$	$s\mu_{ret}$
		$s = Q_h$	α_h
$(k,m,s) \rightarrow (k,m,s-1)$	$k = C$	$1 \leq s \leq Q_h - 1$	$s\mu_{ret}P_{ih}$
		$s = Q_h$	$\alpha_h P_{ih}$
$(k,m,s) \rightarrow (k+1,m-1,s)$	$0 \leq k \leq L-1$	$1 \leq m \leq Q_n - 1$	$m\mu_{red}$
		$m = Q_n$	α_n
	$k = L$	$1 \leq m \leq Q_n - 1$	$m\mu_{red}f$
		$m = Q_n$	$\alpha_n f$
$(k,m,s) \rightarrow (k,m-1,s)$	$k = L$	$1 \leq m \leq Q_n - 1$	$m\mu_{red}(1-f)P_{in}$
		$m = Q_n$	$\alpha_n(1-f)P_{in}$
	$L < k \leq C$	$1 \leq m \leq Q_n - 1$	$m\mu_{red}P_{in}$
		$m = Q_n$	$\alpha_n P_{in}$
$(k,m,s) \rightarrow (k-1,m,s)$	$1 \leq k \leq C$		$k\mu$
$(k,m,s) \rightarrow (k,m,s+1)$	$k = C$		$\lambda_h(1-P_{ih}^1)$
$(k,m,s) \rightarrow (k,m+1,s)$	$k = L$		$\lambda_n(1-P_{in}^1)(1-f)$
	$L < k \leq C$		$\lambda_n(1-P_{in}^1)$
Note:	$\alpha_n = M_n\mu_{red}(1-p_n), \quad \beta_n = M_n\mu_{red}p_n$		
	$\alpha_h = M_h\mu_{ret}(1-p_h), \quad \beta_h = M_h\mu_{ret}p_h$		

Table 2: Transition rates of the proposed model.

associated infinitesimal generator, $\mathbf{Q}$, for the proposed analytic model has the following block tridiagonal structure, whose block size is $(Q_n + 1)(Q_h + 1) \times$

$(Q_n+1)(Q_h+1)$:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{v}_0^0 & \mathbf{v}_0^+ & \dots & 0 & 0 \\ \mathbf{v}_1^- & \mathbf{v}_1^0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{v}_{C-1}^0 & \mathbf{v}_{C-1}^+ \\ 0 & 0 & \dots & \mathbf{v}_C^- & \mathbf{v}_C^0 \end{bmatrix}$$

Matrices $\mathbf{v}$ define the different transitions between blocks, being $\mathbf{v}^+$ the matrices that define transitions $k \rightarrow k+1$ users in the system, $\mathbf{v}^-$ the matrices that define transitions $k \rightarrow k-1$, and $\mathbf{v}^0$ define transitions between states with the same number of users.

In order to compute the steady-state probabilities of the system ($\pi(k,m,s)$) the actual values of the parameters M_n, p_n, M_h and p_h should be known. By balancing the probability fluxes and equating the rate of *blocked first attempts that reattempt* to the sum of the rates of successful and abandoning reattempts, parameters M_n, p_n, M_h and p_h can be expressed in terms of the steady-state probabilities. Following the procedure shown in Appendix A we obtain:

$$p_h = \frac{\sum_{m=0}^{Q_n} \pi(C,m,Q_h)}{\sum_{m=0}^{Q_n} \left[\pi(C,m,Q_h) + \pi(C,m,Q_h-1)\right]} \tag{1}$$

$$M_h = \frac{\lambda_h(1-P_{ih}^1)\left(\sum_{m=0}^{Q_n}\left[(\pi(C,m,Q_h)+\pi(C,m,Q_h-1)\right]\right)}{\mu_{ret}\left(\sum_{k=0}^{C-1}\sum_{m=0}^{Q_n}\pi(k,m,Q_h) + P_{ih}\sum_{m=0}^{Q_n}\pi(C,m,Q_h)\right)} \tag{2}$$

$$p_n = \frac{\zeta_1}{\zeta_2} \quad ; \quad M_n = \frac{\lambda_n(1-P_{in}^1)\zeta_2}{\mu_{red}\zeta_3} \tag{3}$$

where

$$\zeta_1 = \sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s) + (1-f)\sum_{s=0}^{Q_h}\pi(L,Q_n,s)$$

$$\zeta_2 = \sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\left[\pi(k,Q_n-1,s)+\pi(k,Q_n,s)\right] + (1-f)\sum_{s=0}^{Q_h}\left[\pi(L,Q_n-1,s)+\pi(L,Q_n,s)\right]$$

$$\zeta_3 = \sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,Q_n,s) + [f+(1-f)P_{in}]\sum_{s=0}^{Q_h}\pi(L,Q_n,s) + P_{in}\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)$$

The global balance equations, the normalization equation and Eqs. (1)–(3) form a system of simultaneous non-linear equations, which can be solved using —for instance— the iterative procedure sketched next: set $p_n = p_h = 0$, $M_n = Q_n$ and $M_h = Q_h$ and compute the steady-state probabilities using the algorithm defined in [Servi 2002], now compute M_n, p_n, M_h, p_h using Eqs. (1)–(3) and start again. In all of our numerical experiments we repeated the iterative procedure until the relative difference between two consecutive iterations was less than 10^{-4} for all four parameters.

The most common performance parameters used in communication systems are the blocking probabilities of both new sessions (P_b^n) and handovers (P_b^h), which are defined as the probability of being the system in a state where new sessions or handovers are not accepted, respectively. Additionally, it is also used the probability of having a handover failure, denoted as forced termination probability (P_{ft}). Notwithstanding, other performance parameters can describe the behavior of retrial systems more accurately. That performance parameters are the immediate service probability (P_{is}^x), the delayed service probability (P_{ds}^x) and the non-service probability (P_{ns}^x), where x denotes new sessions ($x = n$) or handover requests ($x = h$). P_{is}^x is defined as the probability of a session being served in its first attempt, P_{ds}^x as the probability of obtaining service but not in its first attempt and P_{ns}^x the probability of leaving the system due to impatience without having been served. Obviously, it must be met $P_{is}^x + P_{ds}^x + P_{ns}^x = 1$. Moreover, we define the mean number of redials (retrials) that performs every new session (handover), being u_n (u_h) and the mean number of users redialing (retrying), defined by N_{red} (N_{ret}).

Therefore, the computation of new session performance parameters is done by using:

$$P_{is}^n = \sum_{k=0}^{L-1}\sum_{m=0}^{Q_n}\sum_{s=0}^{Q_h}\pi(k,m,s) + f\sum_{m=0}^{Q_n}\sum_{s=0}^{Q_h}\pi(L,m,s)$$

$$P_{ds}^n = \lambda_n^{-1}\mu_{red}\Big[\sum_{k=0}^{L-1}\sum_{m=0}^{Q_n-1}\sum_{s=0}^{Q_h} m\pi(k,m,s) + M_n\sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,Q_n,s) +$$
$$+ f\sum_{m=0}^{Q_n-1}\sum_{s=0}^{Q_h} m\pi(L,m,s) + M_n f\sum_{s=0}^{Q_h}\pi(L,Q_n,s)\Big]$$

$$P_{ns}^n = \lambda_n^{-1}\mu_{red}P_{in}\Big[\sum_{k=L+1}^{C}\sum_{m=0}^{Q_n-1}\sum_{s=0}^{Q_h} m\pi(k,m,s) + M_n\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s) +$$
$$+ (1-f)\sum_{m=0}^{Q_n-1}\sum_{s=0}^{Q_h} m\pi(L,m,s) + M_n(1-f)\sum_{s=0}^{Q_h}\pi(L,Q_n,s)\Big] +$$
$$+ P_{in}^1\Big[(1-f)\sum_{m=0}^{Q_n}\sum_{s=0}^{Q_h}\pi(L,m,s) + \sum_{k=L+1}^{C}\sum_{m=0}^{Q_n}\sum_{s=0}^{Q_h}\pi(k,m,s)\Big]$$

$$P_b^n = P_{ds}^n + P_{ns}^n = \sum_{k=L+1}^{C} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(k,m,s) + (1-f) \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(L,m,s)$$

$$u_n = \frac{\mu_{red}}{\lambda_n}(1-P_{in})\Big[\sum_{k=L+1}^{C} \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} m\pi(k,m,s) + M_n\zeta_1 + (1-f) \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} m\pi(L,m,s)\Big] +$$

$$+ (1-P_{in}^1)\Big[(1-f) \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(L,m,s) + \sum_{k=L+1}^{C} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(k,m,s)\Big]$$

$$N_{red} = \sum_{k=0}^{C} \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} m\pi(k,m,s) + M_n \sum_{k=0}^{C} \sum_{s=0}^{Q_h} \pi(k,Q_n,s)$$

And the expressions for handovers are:

$$P_{is}^h = \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(k,m,s)$$

$$P_{ds}^h = \lambda_h^{-1}\mu_{ret}\Big[\sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(k,m,s) + M_h \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \pi(k,m,Q_h)\Big]$$

$$P_{ns}^h = \frac{\mu_{ret}}{\lambda_h} P_{ih}\Big[\sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(C,m,s) + M_h \sum_{m=0}^{Q_n} \pi(C,m,Q_h)\Big] + P_{ih}^1 \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(C,m,s)$$

$$P_b^h = \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(C,m,s)$$

$$P_{ft} = \frac{N_H P_{ns}^h}{1 + N_H P_{ns}^h}$$

$$u_h = \frac{\mu_{ret}}{\lambda_h}(1-P_{ih})\Big[\sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(C,m,s) + M_h \sum_{m=0}^{Q_n} \pi(C,m,Q_h)\Big] + (1-P_{ih}^1) \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(C,m,s)$$

$$N_{ret} = \sum_{k=0}^{C} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(k,m,s) + M_h \sum_{k=0}^{C} \sum_{m=0}^{Q_n} \pi(k,m,Q_h)$$

4 Results and discussion

In this section a number of numerical examples are presented with the purpose of illustrating the capabilities and versatility of our model and the analysis methodology. The numerical analysis is also aimed at assessing the impact on performance of varying the values and/or distributions of the system parameters.

For the numerical experiments a basic configuration is used and then the different parameters are varied, normally a single variation is introduced in each experiment. Thus, unless otherwise indicated, the value of the parameters will be those of the basic configuration: $C = 32$, $N_H = \mu_r/\mu_s = 2$, $\mu = \mu_r + \mu_s = 1$, $t = 31$, $P_{ih} = P_{in} = 0.2$, $\mu_{red} = 20$, $P_{ih}^1 = P_{in}^1 = 0$, $\mu_r' = 10\mu_r$ and then $\mu_{ret} = 100/3$.

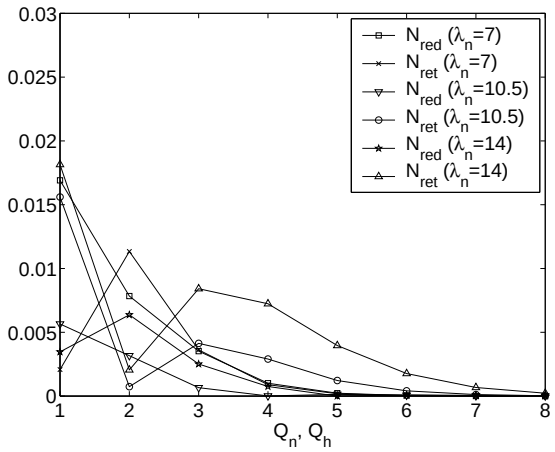


Figure 3: Accuracy of the approximate methodology.

4.1 Accuracy of the proposed methodology

Here we evaluate the accuracy of the approximate analysis as a function of Q_h and Q_n. For a given performance indicator I and given values of Q_h and Q_n the relative error introduced by the approximate model is estimated by $\epsilon_I(Q_n, Q_h) = \left| \frac{I(Q_n+1,Q_h+1)}{I(Q_n,Q_h)} - 1 \right|$.

In [Fig. 3] the relative error estimate is plotted as a function of $Q = Q_h = Q_n$ taking as performance indicators N_{red} and N_{ret}. We have checked the behavior of the proposed methodology using the basic configuration with three different arrival rates, $\lambda_n = 7$, $\lambda_n = 10.5$ and $\lambda_n = 14$ in order to study the system in a wide range of system load. With these values P_b^n (P_b^h) takes values from 1.75% (0.62%) to 44.91% (21.26%) when using $\lambda_n = 7$ and $\lambda_n = 14$ respectively which should capture most of the scenarios of interest.

As it might be expected, except for a very short transient phase, the value of $\epsilon_I(Q_n, Q_h)$ decreases when the values of Q_h and Q_n increase, and also, that a higher load (given by λ_n) results in a poorer accuracy. The curves also show that a good accuracy can be achieved with relatively low values of Q, having been observed in all the numerical examples we have carried out. Moreover, in all the numerical results shown hereafter the values of Q_h and Q_n have been chosen so that $\epsilon_{N_{red}}(Q_n, Q_h) < 10^{-4}$ and $\epsilon_{N_{ret}}(Q_n, Q_h) < 10^{-4}$.

4.2 Redimensioning with redials

Due to the human behavior, users normally redial if a previous attempt has been blocked. Network operators, however, do not consider redials as such simply because they are not able to distinguish between first attempts and redials,

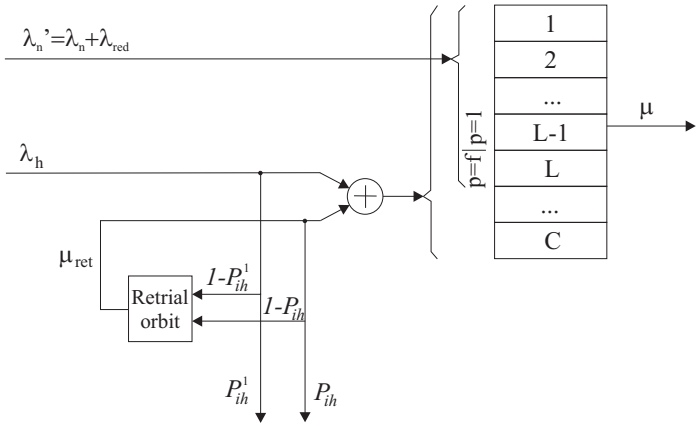


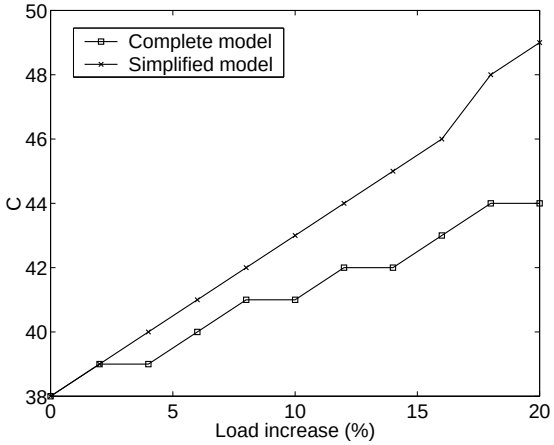
Figure 4: Network operator model where redials are considered as fresh new sessions.

therefore every incoming session is regarded as a first attempt. Without that distinction, a resource over-provisioning can occur because for each user requesting a session whose first attempt is blocked several new session requests are actually accounted (one per attempt). Obviously, the load that redials introduce into the system must be taken into account but we must not consider them as fresh new sessions.

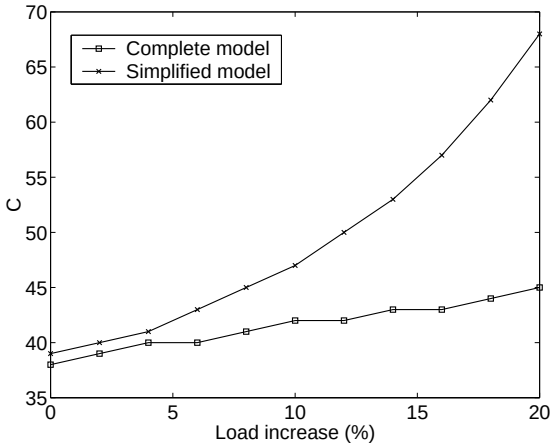
In order to evaluate the magnitude of over-provisioning the following experiment was carried out. We start from a basic situation in which the QoS objectives ($P_b^n \leq 0.05$ and $P_{ft} \leq 0.005$) are fulfilled and consider several values of load growth. For each value of the load increment, the amount of resources (C) is redimensioned in order to meet the QoS objectives. The redimensioning process is done using the complete model and a simplified model where redials are considered as added to fresh new calls, i.e. $\lambda'_n = \lambda_n + \lambda_{red}$, where $\lambda_{red} = \mu_{red} N_{red}$. This last simplified model is shown in [Fig. 4]. [Fig. 5] shows a sample of results from the redimensioning process which reveal that ignoring the existence of redials can produce a significant over-provisioning.

4.3 Impact of automatic retrial configuration

If the network operator enables the automatic retrial option the blocked handover attempts will be automatically retried while the user remains within the handoff area. We consider a fixed mean sojourn time in the handover area ($\mu'_r = 20/3$) and study the impact of varying the retrial rate (μ_{ret}). Note that for varying μ_{ret} while μ'_r is kept constant the value of P_{ih} is varied accordingly using their relationship, $\mu'_r = \mu_{ret} P_{ih}$.



(a) $P_{in} = 0.1$.



(b) $P_{in} = 0$.

Figure 5: Resource redimensioning with redials.

[Fig. 6] shows that a higher value of μ_{ret} results in a lower forced termination probability but also a higher mean number of retrials per session. While the former is a positive effect the later is not that much as it entails an increased signaling load. In order to gain a further insight into the existing tradeoff between P_{ft} and u_h we define the overall cost function $C_T = \beta\lambda_n P_{ft} + \lambda_h u_h$. The choice of the value for β may depend on many factors and a suitable value can vary widely from one situation to another, thus we have used a wide range of values, $\beta = \{2, 5, 10, 15, 20, 50, 100\}$. We also explored the effect of varying the mean sojourn time in the handover area $1/\mu_r'$ (actually a normalized parameter with

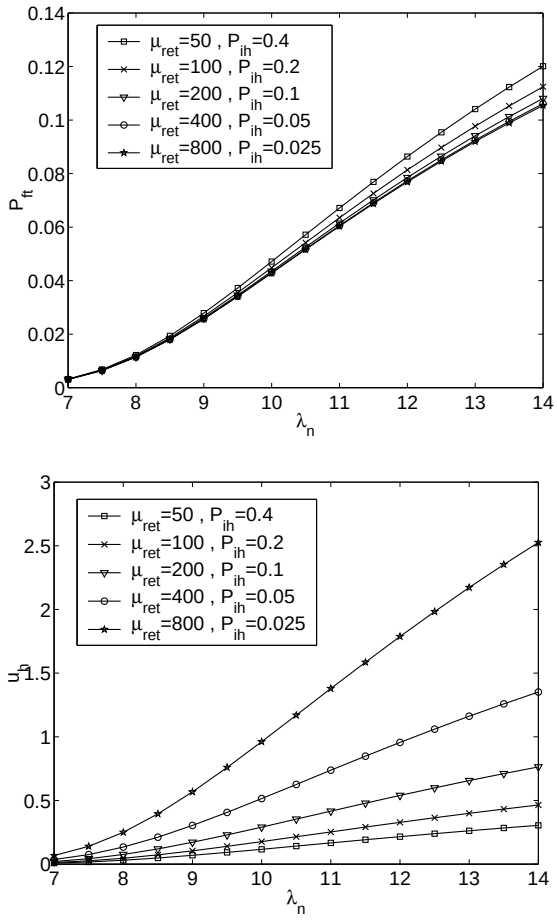


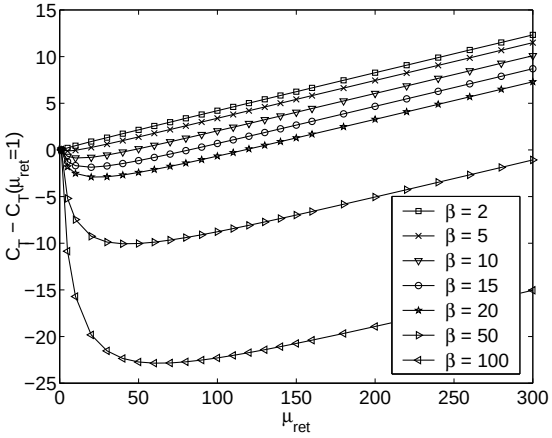
Figure 6: Performance parameters for different retrial configurations.

respect to $1/(C\mu)$ has been used $\Gamma = C\mu/\mu'_r$).

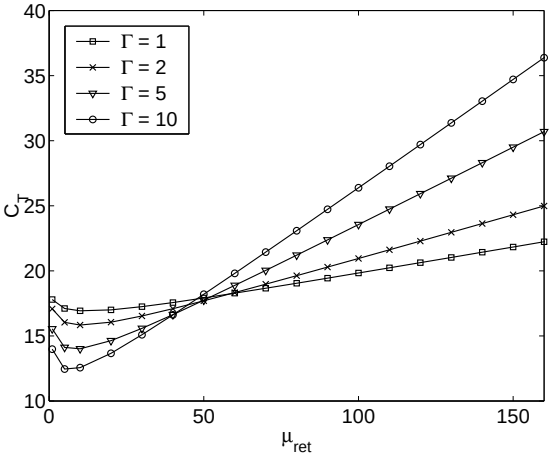
The shape of cost curves in [Fig. 7(a)] shows the existence of an optimal configuration point. Both the relevance of the optimal configuration point and the value of the retrial rate at which it is attained increase when the weight factor β is increased. Moreover, [Fig. 7(b)] shows that the optimal value of μ_{ret} is rather insensitive to the mean value of the sojourn time in the handover area.

4.4 Distribution of the maximum number and time between reattemps

In real systems and for simplicity, the time between retrials as well as the maximum number of retrials per request use to take a deterministic value instead



(a) Increment $(C_T(\mu_{ret}) - C_T(1))$ when μ_{ret} varies, $\Gamma = 1$.



(b) Absolute value, $\beta = 10$.

Figure 7: Cost function, $\lambda_n = 12$.

of an stochastic one [Onur et al. 2002]. In our model, however, in order to keep the mathematical analysis tractable, we used an exponentially distributed time between retrials and a geometric distribution for the maximum number of reattempts. Here we validate these two assumptions with the help of a simulation model. More concretely, a specific discrete-event simulator has been implemented in C. The simulation model is the same as the analytical model, except that the less stringent assumptions necessary in the latter allowed to consider a wider range of probabilistic distributions, namely: deterministic, erlangian and hype-

rexponential distributions, in addition to exponential distributions. Simulation results have been obtained generating 10^7 new session arrivals to the system, which offer very low confidence intervals in the performance parameters of interest.

4.4.1 Time distribution between redials/retrials

We analyze the values of P_b^n, P_b^h, u_n and u_h when the distribution of the time between redials, retrials, or both are switched from exponential to deterministic, keeping constant its mean value. From the results in [Figs. 8, 9] and other that show the same conclusions, we determine that assuming an exponential distribution for the time between redials and/or retrials has a negligible impact in all the performance parameters of interest.

In addition to the exponential ($CV = 1$) and deterministic ($CV = 0$) distributions we also considered the hyperponential and erlang distributions. Note that the coefficient of variation (CV) of random variable X is the ratio of its standard deviation to its mean, $CV_X = \sigma_X / E[X]$. The variability, i.e. the CV, of the erlang distribution lies between that of the deterministic and the exponential distribution, $0 < CV < 1$. In turn, the hyperponential distribution allows a higher variability, $CV > 1$. Using the aforementioned distributions, we ran a series of simulations covering a wide range of scenarios in order to evaluate the impact that the variability of the time between redials has on the performance parameters. In all cases, the mean of the distribution was the same, i.e. $1/\mu_{red}$. [Figs. 10, 11] show a typical example of the results obtained. In general it can be concluded that in practice, for all the performance parameters studied, the sensitivity to the distribution of the time between redials is negligible. Obviously, similar conclusions can be drawn for retrials.

As it can be noted in [Fig. 10] we have varied the system load to obtain a wide range of blocking probabilities. Although blocking probabilities above 20% are not acceptable working points, we have considered them because it is in this situation when the retrial phenomenon becomes more noticeable and then it is suitable for evaluation purposes. Besides, such severe overloads appear in some scenarios, like emergency situations or in special dates.

4.4.2 Distribution of the maximum number of reattempts

We compare a geometric distribution (after each unsuccessful attempt the user decides to abandon the system with probability P_i) with a deterministic distribution (the users leave the system after d unsuccessful attempts). For making these two options comparable the mean number of reattempts must be the same in both cases. Note it is not the same as both distributions having the same

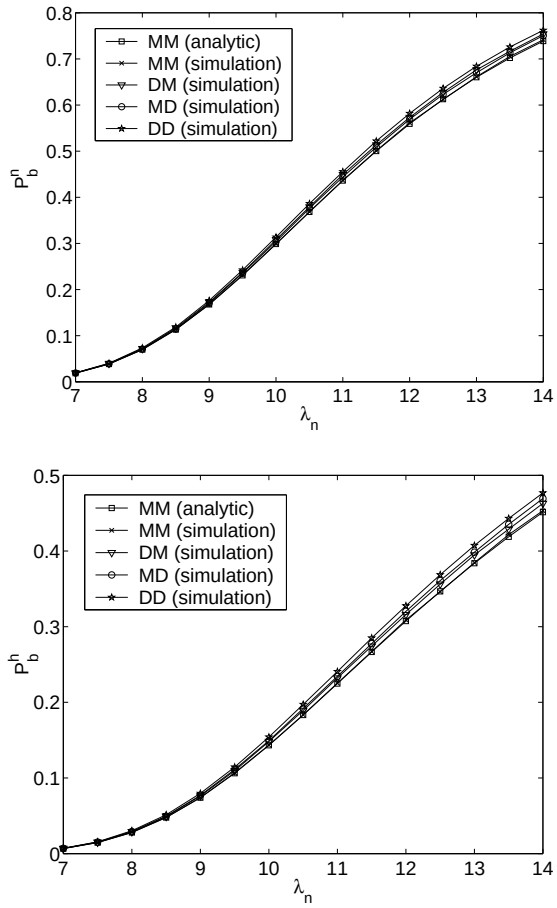


Figure 8: Distribution of the time between reattempts: impact on P_b^n and P_b^h. Legend: XY, X (Y) $\equiv$ distribution for redials (retrials); $M \equiv$ exponential, $D \equiv$ deterministic.

mean as the distributions refer to the maximum number of reattempts and not to the actual number of reattempts.

While the following discussion deals only with retrials it can be easily extended to redials as well. Let q denote the blocking probability for retrials (note that in general $q \neq P_b^h$), the average number of retrials is, in the geometric case,

$$u_h^{Geo} = \sum_{n \geq 1} P_b^h (1 - P_{ih}^1)((1 - P_{ih})q)^{n-1}(1 - (1 - P_{ih})q) = \frac{(1 - P_{ih}^1)P_b^h}{1 - (1 - P_{ih})q} \quad (4)$$

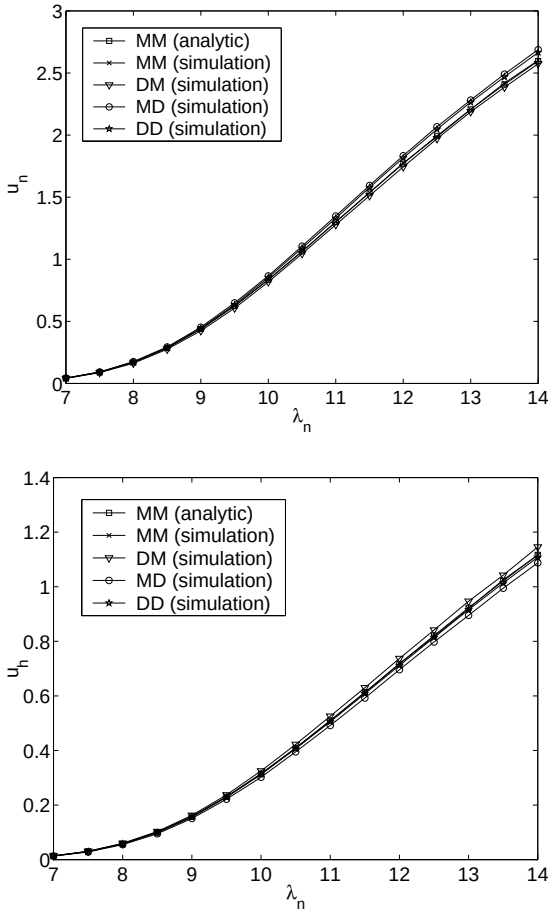


Figure 9: Distribution of the time between reattempts: impact on u_n and u_h. Legend: XY, X (Y) $\equiv$ distribution for redials (retrials); $M \equiv$ exponential, $D \equiv$ deterministic.

and in the deterministic case,

$$u_h^D = (1-q)P_b^h[1 + 2q + 3q^2 + \ldots + (d-1)q^{d-2}] + dP_b^h q^{d-1} = P_b^h \frac{1-q^d}{1-q} \quad (5)$$

If we assume that both q and P_b^h take approximately the same value in both cases, by equating the right hand side of Eq. (4) and Eq. (5) we obtain

$$P_{ih} = \frac{1-q}{q(1-q^d)}(q^d - P_{ih}^1) \quad (6)$$

For a given value of d, by using the expressions for P_b^h and u_h (see the end of

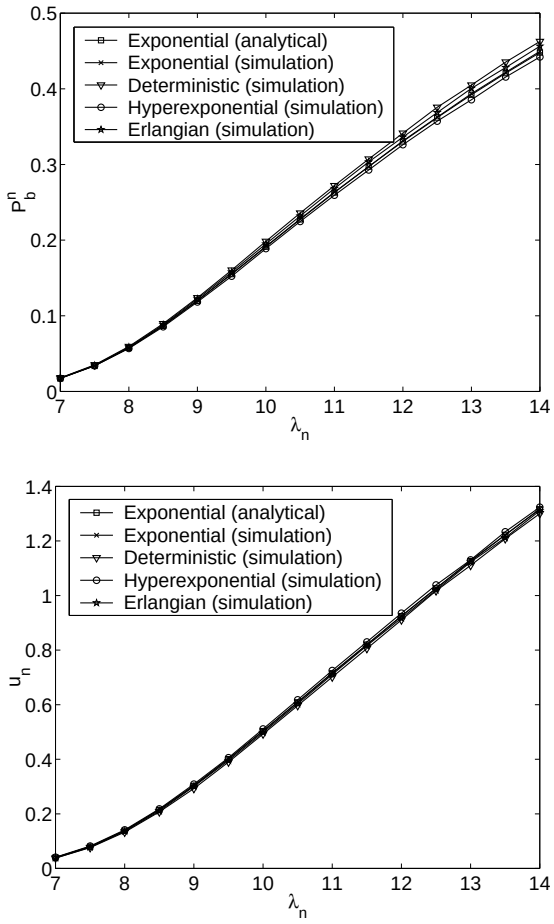


Figure 10: Other distributions of the time between redials. The coefficient of variation of the 2-phases erlangian distribution used has been $CV = 1/\sqrt{2}$ and $CV = 1.5$ for the 2-branches hyperexponential distribution.

[Section 3]) and Eqs. (4) and (6), the value of P_{ih} that yields $u_h^{Geo} = u_h^D$ can be iteratively computed. The results shown in [Fig. 12] demonstrate that using the adjusting procedure described above, our model can provide an excellent approximation for the performance analysis of a system in which the maximum number of retrials is a fixed number.

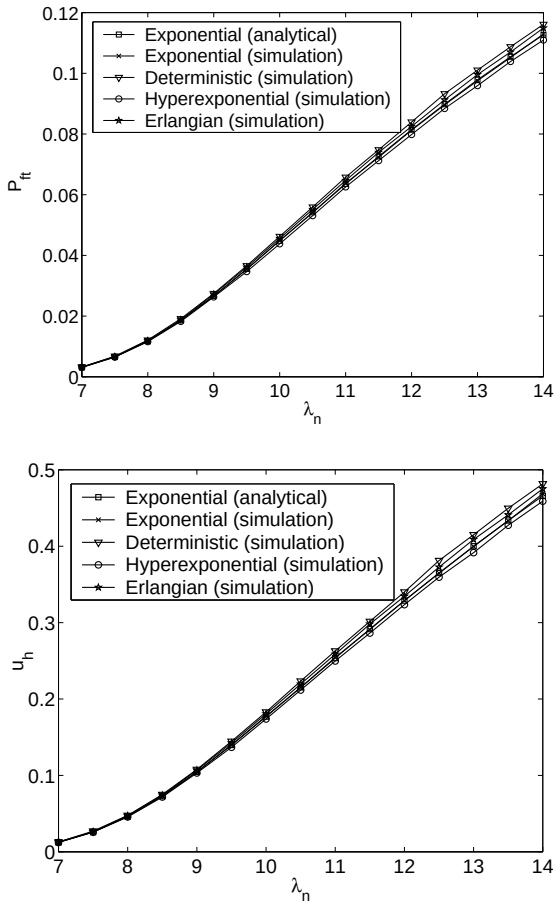


Figure 11: Other distributions of the time between redials. The coefficient of variation of the 2-phases erlangian distribution used has been $CV = 1/\sqrt{2}$ and $CV = 1.5$ for the 2-branches hyperexponential distribution.

5 Conclusions

In mobile communication systems like cellular networks, Mobile IP or the recently defined IEEE 802.16e and IEEE 802.20 networks, mobile operators must guarantee seamless mobility to its customers. In these networks, repeated attempts occur due to user redials when their session establishments are blocked and also due to automatic retries when a handover fails. The impact of both phenomena plays an important role in the system performance and, therefore, it should not be ignored. However, the main feature of the Markovian model describing such a complex system is the space-heterogeneity along the two infinite

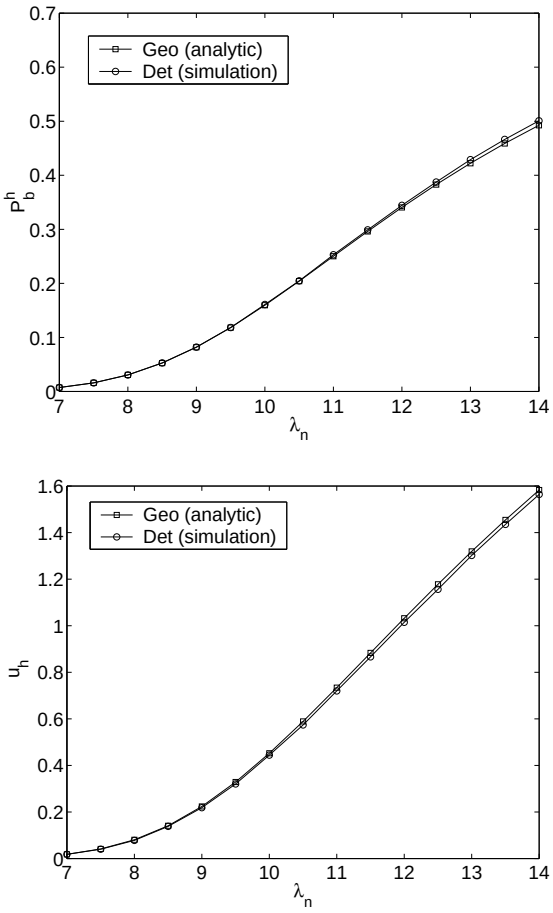


Figure 12: Analytical approximation of a deterministic maximum number of retrials; $d = 5$, $P_{ih}^1 = 0$.

dimensions. Due to this fact, we have developed an approximate methodology that aggregates users in the retrial/redial orbit beyond a given occupancy and that is able to get as much accuracy as desired.

A numerical evaluation of the system has been performed in order to evaluate the impact of the reattempt phenomena in the system performance. We have studied the effect of automatic retrials for handovers while the user remains into the handover area, giving some guidelines to the network operators in order to configure this behavior optimally.

Finally, we have shown how our model can be used to obtain a tight performance approximation when the time between reattempts and maximum number

of reattempts are deterministic. Results of this approximate method are compared against those obtained by simulation, concluding that the proposed method is very accurate.

Acknowledgements

This work was supported by the Spanish Government (30% PGE) and the European Commission (70% FEDER) through projects TSI2005-07520-C03-03 and TEC2004-06437-C05-01 and by *Cátedra Telefónica de Internet y Banda Ancha* (e-BA) from the Universidad Politécnica de Valencia. Besides M. Jose Domenech-Benlloch was supported by the Spanish Ministry of Education and Science under contract AP-2004-3332.

References

[Artalejo and Pozo 2002] Artalejo, J.R., Pozo, M.: "Numerical calculation of the stationary distribution of the main multiserver retrial queue"; Annals of Operations Research, 116, 1–4 (2002), 41–56.

[Bolton et al. 2007] Bolton, W., Xiao, Y., Guizani, M.: "IEEE 802.20: mobile broadband wireless access"; IEEE Wireless Communications, 14 (2007), 84–95.

[Bonald and Roberts 2003] Bonald, T., Roberts, J.: "Congestion at flow level and the impact of user behaviour"; Computer Networks, 42 (2003), 521–536.

[Bright and Taylor 1995] Bright, L., Taylor, P.G.: "Calculating the equilibrium distribution of level dependent quasi-birth-and-death processes"; Communications in Statistics-Stochastic Models, 11, 3 (1995), 497–525.

[Domenech et al. 2005] Doménech-Benlloch, M.J., Giménez-Guzmán, J.M., Martínez-Bauset, J., Casares-Giner, V.: "Efficient and accurate methodology for solving multiserver retrial systems"; IEE Electronic Letters, 41, 17 (2005), 967–969.

[Falin 1983] Falin, G.: "Calculation of probability characteristics of a multiline system with repeat calls"; Moscow University Computational Mathematics and Cybernetics, 1 (1983), 43–49.

[IEEE 802.16] IEEE 802.16 standard. (http://www.ieee802.org/16/pubs/P80216e.html) May 2007.

[Jonin and Sedol 1970] Jonin, G., Sedol, J.: "Telephone systems with repeated calls"; Proc. 6th International Teletraffic Conference ITC'6 (1970), 435.1–435.5.

[Khan and Zeghlache 1997] Khan F., Zeghlache, D.: "Effect of Cell Residence Time Distribution on the Performance of Cellular Mobile Networks"; Proc. of IEEE VTC, (1997), 949–953.

[Latouche and Ramaswami 1999] Latouche, G., Ramaswami, V.: "Introduction to Matrix Analytic Methods in Stochastic Modeling"; ASA-SIAM (1999).

[Marsan et al. 1999] Marsan, M.A., Carolis, G.D., Leonardi, E., Cigno, R.L., Meo, M.: "How many cells should be considered to accurately predict the performance of cellular networks?"; Proc. European Wireless, (1999).

[Marsan et al. 2001] Marsan, M.A., Carolis, G.D., Leonardi, E., Cigno, R.L., Meo, M.: "Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials"; IEEE Journal on Selected Areas in Communications, 19, 2 (2001), 332–346.

[Neuts 1981] Neuts, M.: "Matrix-geometric Solutions in Stochastic Models: An Algorithmic Approach"; The Johns Hopkins University Press (1981).

[Neuts and Rao 1990] Neuts, M., Rao, B.: "Numerical investigation of a multiserver retrial model"; Queueing systems, 7 (1990), 169–190.

[Onur et al. 2002] Onur, E., Deliç, H., Ersoy, C., Çağlayan, M.U.: "Measurement-based replanning of cell capacities in GSM networks"; Computer Networks, 39 (2002), 749–767.
[Orlik and Rappaport 2001] Orlik, P., Rappaport, S.: "On the Handoff Arrival Process in Cellular Communications"; Wireless Networks Journal (WINET), 7, 2 (2001), 147–157.
[Pla and Casares-Giner 2002] Pla, V., Casares-Giner, V.: "Effect of the handoff area sojourn time distribution on the performance of cellular networks"; Proc. of IEEE MWCN, (2002), 401–405.
[Servi 2002] Servi, L.D.: "Algorithmic solutions to two-dimensional birth-death processes with application to capacity planning"; Telecommunication Systems, 21, 2–4 (2002), 205–212.
[Tran-Gia and Mandjes 1997] Tran-Gia, P., Mandjes, M.: "Modeling of customer retrial phenomenon"; IEEE Journal on Selected Areas in Communications, 15, 8 (1997), 1406–1414.

A Computation of the approximate model parameters

By balancing the probability flux into and out of a particular set of states we can compute M_n, p_n, M_h and p_h. In general, we define two sets of states S_a and S_b and equate the transition rates between them, i.e. the balance equations:

$$\sum_{x \in S_a, y \in S_b} q_{xy}\pi_x = \sum_{x \in S_a, y \in S_b} q_{yx}\pi_y$$

To compute parameters M_n and p_n we define Q_n different values for S_a and S_b, being

$$S_a^{(i)} = \{(k,m,s) : 0 \le k \le C; m = i-1; 0 \le s \le Q_h\}$$
$$S_b^{(i)} = \{(k,m,s) : 0 \le k \le C; m = i; 0 \le s \le Q_h\}$$

for $i \in [1, Q_n]$. Using these sets of states we will obtain a different balance equation for each value of i:

- $i = 1$:

$$\mu_{red}\sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,2,s) + f\mu_{red}\sum_{s=0}^{Q_h}\pi(L,2,s) + \mu_{red}(1-f)P_{in}\sum_{s=0}^{Q_h}\pi(L,2,s) +$$
$$+\mu_{red}P_{in}\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,2,s) = \lambda_n(1-f)(1-P_{in}^1)\sum_{s=0}^{Q_h}\pi(L,0,s) + \lambda_n(1-P_{in}^1)\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,0,s)$$

- $i = 2$:

$$2\mu_{red}\sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,1,s) + 2f\mu_{red}\sum_{s=0}^{Q_h}\pi(L,1,s) + 2\mu_{red}(1-f)P_{in}\sum_{s=0}^{Q_h}\pi(L,1,s) +$$
$$+2\mu_{red}P_{in}\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,1,s) = \lambda_n(1-f)(1-P_{in}^1)\sum_{s=0}^{Q_h}\pi(L,1,s) + \lambda_n(1-P_{in}^1)\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,1,s)$$

– ...

– $i = Q_n$:

$$\alpha_n \sum_{k=0}^{L-1} \sum_{s=0}^{Q_h} \pi(k, Q_n, s) + f\alpha_n \sum_{s=0}^{Q_h} \pi(L, Q_n, s) + \alpha_n (1-f) P_{in} \sum_{s=0}^{Q_h} \pi(L, Q_n, s) +$$

$$+\alpha_n P_{in} \sum_{k=L+1}^{C} \sum_{s=0}^{Q_h} \pi(k, Q_n, s) = \lambda_n (1-f)(1-P_{in}^1) \sum_{s=0}^{Q_h} \pi(L, Q_n - 1, s) +$$

$$+\lambda_n (1-P_{in}^1) \sum_{k=L+1}^{C} \sum_{s=0}^{Q_h} \pi(k, Q_n - 1, s)$$

Summing all the balance equations:

$$\mu_{red} \sum_{k=0}^{L-1} \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} m\pi(k, m, s) + f\mu_{red} \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} m\pi(L, m, s) + \mu_{red}(1-f) P_{in} \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} m\pi(L, m, s) +$$

$$+\mu_{red} P_{in} \sum_{k=L+1}^{C} \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} m\pi(k, m, s) + \alpha_n \sum_{k=0}^{L-1} \sum_{s=0}^{Q_h} \pi(k, Q_n, s) + f\alpha_n \sum_{s=0}^{Q_h} \pi(L, Q_n, s) +$$

$$+\alpha_n (1-f) P_{in} \sum_{s=0}^{Q_h} \pi(L, Q_n, s) + \alpha_n P_{in} \sum_{k=L+1}^{C} \sum_{s=0}^{Q_h} \pi(k, Q_n, s) =$$

$$= \lambda_n (1-f)(1 - P_{in}^1) \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} \pi(L, m, s) + \lambda_n (1-P_{in}^1) \sum_{k=L+1}^{C} \sum_{m=0}^{Q_n-1} \sum_{s=0}^{Q_h} \pi(k, m, s)$$

Reordering the last equation and taking into account that the rate of blocked first new sessions attempts is equal to the sum of the rates of successful and abandoning redials, we obtain

$$\lambda_n (1-f)(1-P_{in}^1) \sum_{s=0}^{Q_h} \pi(L, Q_n, s) + \lambda_n (1-P_{in}^1) \sum_{k=L+1}^{C} \sum_{s=0}^{Q_h} \pi(k, Q_n, s) =$$

$$= p_n M_n \mu_{red} \sum_{k=0}^{L-1} \sum_{s=0}^{Q_h} \pi(k, Q_n, s) + p_n f M_n \mu_{red} \sum_{s=0}^{Q_h} \pi(L, Q_n, s) +$$

$$+p_n M_n \mu_{red} (1-f) P_{in} \sum_{s=0}^{Q_h} \pi(L, Q_n, s) + p_n M_n \mu_{red} P_{in} \sum_{k=L+1}^{C} \sum_{s=0}^{Q_h} \pi(k, Q_n, s) \qquad (7)$$

From the last balance equation ($i = Q_n$):

$$\lambda_n(1-f)(1-P_{in}^1)\sum_{s=0}^{Q_h}\pi(L,Q_n-1,s)+\lambda_n(1-P_{in}^1)\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n-1,s)=$$
$$=M_n\mu_{red}\sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)+fM_n\mu_{red}\sum_{s=0}^{Q_h}\pi(L,Q_n,s)+M_n\mu_{red}(1-f)P_{in}\sum_{s=0}^{Q_h}\pi(L,Q_n,s)+$$
$$+M_n\mu_{red}P_{in}\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)-p_nM_n\mu_{red}\sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)-$$
$$-p_nfM_n\mu_{red}\sum_{s=0}^{Q_h}\pi(L,Q_n,s)-p_nM_n\mu_{red}(1-f)P_{in}\sum_{s=0}^{Q_h}\pi(L,Q_n,s)-$$
$$-p_nM_n\mu_{red}P_{in}\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s) \tag{8}$$

Summing Eqs. (7) and (8) we obtain:

$$\lambda_n(1-P_{in}^1)\Big[\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\big[\pi(k,Q_n-1,s)+\pi(k,Q_n,s)\big]+(1-f)\sum_{s=0}^{Q_h}\big[\pi(L,Q_n-1,s)+\pi(L,Q_n,s)\big]\Big]=$$
$$=M_n\mu_{red}\Big[\sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)+f\sum_{s=0}^{Q_h}\pi(L,Q_n,s)+(1-f)P_{in}\sum_{s=0}^{Q_h}\pi(L,Q_n,s)+$$
$$+P_{in}\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)\Big]$$

Finally, we can obtain the desired values for p_n and M_n:

$$p_n=\frac{\zeta_1}{\zeta_2}\quad;\quad M_n=\frac{\lambda_n(1-P_{in}^1)\zeta_2}{\mu_{red}\zeta_3}$$

where

$$\zeta_1=\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)+(1-f)\sum_{s=0}^{Q_h}\pi(L,Q_n,s)$$

$$\zeta_2=\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\big[\pi(k,Q_n-1,s)+\pi(k,Q_n,s)\big]+(1-f)\sum_{s=0}^{Q_h}\big[\pi(L,Q_n-1,s)+\pi(L,Q_n,s)\big]$$
$$\zeta_3=\sum_{k=0}^{L-1}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)+[f+(1-f)P_{in}]\sum_{s=0}^{Q_h}\pi(L,Q_n,s)+P_{in}\sum_{k=L+1}^{C}\sum_{s=0}^{Q_h}\pi(k,Q_n,s)$$

The computation of M_h and p_h is similar, considering the following values for S_a and S_b:

$$S_a^{(j)} = \{(k,m,s) : 0 \leq k \leq C; 0 \leq m \leq Q_n; s = j-1\}$$
$$S_b^{(j)} = \{(k,m,s) : 0 \leq k \leq C; 0 \leq m \leq Q_n; s = j\}$$

for $j \in [1, Q_h]$. Using these sets of states we will obtain the next balance equations:

- $j = 1$:

$$\mu_{ret} \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \pi(k,m,1) + \mu_{ret} P_{ih} \sum_{m=0}^{Q_n} \pi(C,m,1) = \lambda_h (1 - P_{ih}^1) \sum_{m=0}^{Q_n} \pi(C,m,0)$$

- $s = 2$:

$$2\mu_{ret} \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \pi(k,m,2) + 2\mu_{ret} P_{ih} \sum_{m=0}^{Q_n} \pi(C,m,2) = \lambda_h (1 - P_{ih}^1) \sum_{m=0}^{Q_n} \pi(C,m,1)$$

- ...

- $j = Q_h$:

$$\alpha_h \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \pi(k,m,Q_h) + \alpha_h P_{ih} \sum_{m=0}^{Q_n} \pi(C,m,Q_h) = \lambda_h (1 - P_{ih}^1) \sum_{m=0}^{Q_n} \pi(C,m,Q_h - 1)$$

The sum of these balance equations is

$$\lambda_h (1-P_{ih}^1) \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} \pi(C,m,s) = \mu_{ret} \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(k,m,s) + \mu_{ret} P_{ih} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(C,m,s) +$$
$$+ M_h \mu_{ret} (1-p_h) \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \pi(k,m,Q_h) + M_h \mu_{ret} (1-p_h) P_{ih} \sum_{m=0}^{Q_n} \pi(C,m,Q_h)$$

and reordering the last expression:

$$\lambda_h (1-P_{ih}^1) \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h} \pi(C,m,s) - \lambda_h (1-P_{ih}^1) \sum_{m=0}^{Q_n} \pi(C,m,Q_h) = \mu_{ret} \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(k,m,s) +$$
$$+ M_h \mu_{ret} \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \pi(C,m,Q_h) - p_h M_h \mu_{ret} \sum_{k=0}^{C-1} \sum_{m=0}^{Q_n} \pi(k,m,Q_h) + \mu_{ret} P_{ih} \sum_{m=0}^{Q_n} \sum_{s=0}^{Q_h-1} s\pi(C,m,s) +$$
$$+ M_h \mu_{ret} P_{ih} \sum_{m=0}^{Q_n} \pi(C,m,Qh) - p_h M_h P_{ih} \mu_{ret} \sum_{m=0}^{Q_n} \pi(C,m,Q_h)$$

Reordering the last equation and taking into account that the rate of blocked first attempts of handover sessions is equal to the sum of the rates of successful and abandoning automatic retrials, we obtain

$$\lambda_h(1-P_{ih}^1)\sum_{m=0}^{Q_n}\pi(C,m,Q_h)=p_hM_h\mu_{ret}\sum_{k=0}^{C-1}\sum_{m=0}^{Q_n}\pi(k,m,Q_h)+p_hM_hP_{ih}\mu_{ret}\sum_{m=0}^{Q_n}\pi(C,m,Q_h) \tag{9}$$

From the last balance equation:

$$\lambda_h(1-P_{ih}^1)\sum_{m=0}^{Q_n}\pi(C,m,Q_h-1)=\mu_{ret}\Big[M_h\sum_{k=0}^{C-1}\sum_{m=0}^{Q_n}\pi(k,m,Q_h)-p_hM_h\sum_{k=0}^{C-1}\sum_{m=0}^{Q_n}\pi(k,m,Q_h)+$$
$$+M_hP_{ih}\sum_{m=0}^{Q_n}\pi(C,m,Q_h)-p_hM_hP_{ih}\sum_{m=0}^{Q_n}\pi(C,m,Q_h)\Big] \tag{10}$$

Summing Eqs. (9) and (10):

$$\lambda_h(1-P_{ih}^1)\Big[\sum_{m=0}^{Q_n}\pi(C,m,Q_h)+\sum_{m=0}^{Q_n}C,m,Qh-1\Big]=$$
$$=M_h\mu_{ret}\Big[\sum_{k=0}^{C-1}\sum_{m=0}^{Q_n}\pi(k,m,Q_h)+P_{ih}\sum_{m=0}^{Q_n}\pi(C,m,Q_h)\Big]$$

Now, we are able to obtain the values for p_h and M_h:

$$p_h=\frac{\sum\limits_{m=0}^{Q_n}\pi(C,m,Q_h)}{\sum\limits_{m=0}^{Q_n}\big[\pi(C,m,Q_h)+\pi(C,m,Q_h-1)\big]}$$

$$M_h=\frac{\lambda_h(1-P_{ih}^1)\Big[\sum\limits_{m=0}^{Q_n}\big[\pi(C,m,Q_h)+\pi(C,m,Q_h-1)\big]}{\mu_{ret}\Big[\sum\limits_{k=0}^{C-1}\sum\limits_{m=0}^{Q_n}\pi(k,m,Q_h)+P_{ih}\sum\limits_{m=0}^{Q_n}\pi(C,m,Q_h)\Big]}$$