Analysis of a Handover Procedure with Queueing, Retrials and Impatient Customers[*]

Jose Manuel Gimenez-Guzman[†], M^{a} Jose Domenech-Benlloch[‡],
Jorge Martínez-Bauset, Vicent Pla and Vicente Casares-Giner
Department of Communications
Universidad Politécnica de Valencia (UPV)
ETSIT Camí de Vera s/n, 46022 Valencia, Spain.
Telephone: +34 963879733, Fax: +34 963877309
E-mail: (jogiguz,mdoben)@doctor.upv.es
(jmartinez,vpla,vcasares)@dcom.upv.es

Abstract

We evaluated the impact that new session retrials have on the performance of a mobile cellular network which deploys a fractional number of guard channels, a queue for handover sessions and an exponential deadline for serving those requests, modeling in this way the overlapping area between cells. To solve the Markov model we introduced an approximate methodology which is substantially more accurate than previous ones, while increasing the computation cost only marginally. Results show that deploying a handover queue and a fractional number of guard channels help to improve system capacity while guaranteeing a given QoS objective. Finally, we evaluated the magnitude of the over-dimensioning that takes place when retrials are perceived as an increment in the arrival rate of new sessions, showing that it can be severe when the terminals retry persistently as occurs when equipped with automatic redialing.

***Keywords*— Cellular networks, fractional guard channel, retrials, quasi–birth–and–death process.**

1 Introduction

A common assumption when evaluating the performance of communication systems is that users that do not obtain an immediate service leave the system without retrying. However, due to the increasing number of users and the complexity of current systems the impact of retrials is no longer negligible, and this is particularly true in mobile cellular networks [1]. The impact of retrials has been extensively studied mostly in fixed networks, see for example [2] and references therein. Nevertheless, the problem of customer retrials in mobile networks is different from the problem in fixed networks due to the necessity of maintaining the communication while the terminal is moving.

Recently, different papers studied the customer retrial phenomenon in a mobile network context by analytical [3, 4] or simulation [5, 6] models. The common approach is to deploy a multiserver model with a finite or infinite customer population and no waiting facility, where blocked sessions either new or both new and handover, can retry indefinitely. Customer impatience is also taken into account by using a geometric distribution for the number of retrials. Also, an integer number of guard channels is deployed to limit the probability of forced session termination, because from the customer point of view the forced termination of an ongoing session is less desirable than blocking a new one.

Handover queueing schemes have also been proposed as a means to limit the probability of forced session termination, see for example [7] and references therein. Additionally, deploying a fractional amount of guard channels allows the operator to limit with more precision the probability of forced session termination and as a result to achieve a higher system capacity [8].

[*] This work has been supported by the Spanish Government (PGE, 30%) and the European Commission (FEDER, 70%) through the projects TIC2003-08272 and TEC2004-06437-C05-01.

[†] This author was supported by the Universidad Politécnica de Valencia under *Programa de Incentivo a la Investigación*.

[‡] This author was supported by the Spanish Ministry of Education and Science under contract AP-2004-3332

We analyze a model in which blocked new session arrivals retry and customer impatience is being accounted for by using a geometric distribution for the number of retrials. Blocked handover sessions are queued but an exponential deadline is defined beyond which a session is forced to terminate, which models the time spent by handover sessions in the overlapping area between adjacent cells. It is clear that a new session is not granted a server while queued handover sessions exist, and therefore queuing is a handover prioritizing scheme. Our model also supports the reservation of a fractional number of guard channels for handover sessions. We consider that this model has not been sufficiently studied.

In the models studied in the literature, the population can be very large or even infinite, so the numerical computation required to solve the model and to obtain values for the parameters of interest can be extremely large in terms of memory space and CPU time, or even impossible in many cases. Therefore, approximate methodologies are needed like the one proposed in [3], which is studied in a mobile cellular network scenario and it is based on grouping states according to the presence or not of users in the retrial orbit.

We have developed a methodology to reduce the state space of the Markov model in such a way that the accuracy is not compromised and the computation cost is greatly reduced. Our approximation is a generalization of the one in [3], where it was shown to be a good approximation for the blocking probabilities. Notwithstanding, the approximation in [3] is too simple to obtain accurate values for other common performance parameter used in retrial systems like the immediate service probability (P_{is}), the delayed service probability (P_{ds}) and the non-service probability (P_{ns}), being $P_{is} + P_{ds} + P_{ns} = 1$. An additional feature of our proposal is that allows a gradual transition from the model in [3] towards the exact model.

The rest of the paper is structured as follows. Section 2 describes the model of the system, while Section 3 introduces the proposed approximation methodology, evaluates its performance and defines the performance parameters of interest. Section 4 evaluates the impact that the different features of the model have on system performance. It also evaluates the magnitude of the over-dimensioning required to meet a given QoS objective when retrials are perceived as an increment in the arrival rate of new sessions. Finally, Section 5 summarizes the paper.

2 Model Description

The model under consideration is a cellular system where each cell is served by a unique base station. We consider the homogeneous case where all cells are statistically identical, and consequently the global performance of the system can be analyzed focusing on a single cell, under the assumption that neighboring cells show independent random behavior.

Figure 1 displays the model that characterizes the cell under study, in which we consider two different arrival streams. The first one with rate λ_n represents new sessions that are initiated in the cell, and the second one with rate λ_h represents the incoming flow of handovers entering the cell. The value of λ_h is determined by assuming that the system is in statistical equilibrium and therefore the rate at which handover sessions enter and exit a cell are equal [9]. Consequently, the incoming handover rate for the cell under study must be evaluated numerically using a fixed point approximation. For the sake of mathematical tractability we make the common assumptions of Poisson arrival processes and exponentially distributed random variables.

In our model, when a new session request is blocked the customer retries, at least once. In the case of successive blockings, the customer reattempts with probability $(1 - P_i)$. The time between reattempts of the same customer is exponentially distributed with rate (μ_{ret}).

For handover sessions, we consider that the resource allocation in the destination cell can be delayed while crossing the overlapping area between adjacent cells. We model that scenario by incorporating a FIFO queue of finite capacity Q_h and by considering that the sojourn time in the overlapping area is exponentially distributed with rate μ_r', which has been shown to be a good approximation [10].

The cell under study has a total of C resource units, being the physical meaning of a unit of resources dependent on the specific technological implementation of the radio interface. Without loss of generality, we consider that each session occupies one resource unit. The session duration is exponentially distributed with rate μ_s and the cell residence time is exponentially distributed with rate μ_r. Hence, the resource holding time in a cell is exponentially distributed with rate $\mu = \mu_s + \mu_r$. The maximum time a handover request can be queued is also exponentially distributed with rate $\gamma = \mu_s + \mu_r'$.

We deploy a Fractional Guard Channel (FGC) [11] admission control policy, which divides the resource

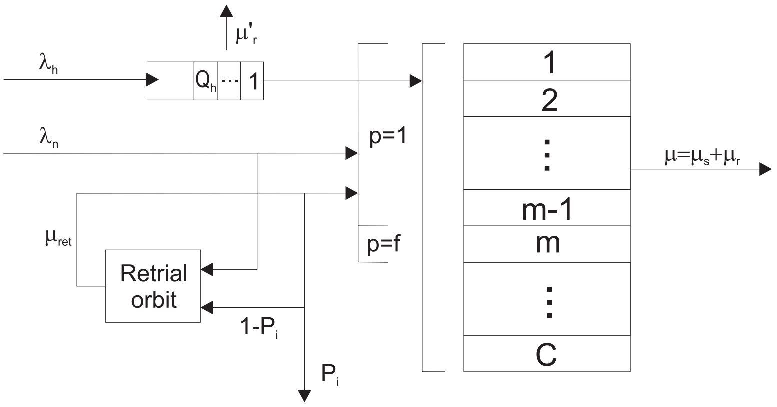


Figure 1: System model.

units into three groups: primary, secondary and one partially reserved resource unit (PRU). Primary resource units can be assigned to both new and handover requests, while the secondary group is reserved for handovers. The PRU can be assigned to both handover and new session requests, but in this last case, only with a certain probability. The FGC policy is characterized by only one parameter t ($0 \leq t \leq C$), from which the number of resource units in the primary (m) and secondary group (n) and the probability that the PRU is allocated to a new sessions (f) can be determined in the following way: $m = \lfloor t \rfloor$, $f = t - m$ and $n = C - (m + 1)$.

The order followed to allocate a resource unit when a handover request arrives is: primary group, PRU, secondary group and, lastly, a position in the queue. If the assignment fails then the session is forced to terminate. The resource allocation order followed when a new session request or a reattempt arrives is: primary group and, if it is not possible, then the PRU with probability f. If the assignment is not possible the request is rejected and the customer either retries again or abandon. When a resource unit is released, it is assigned to the handover request at the head of queue, if the queue is not empty.

3 Performance Analysis

Due to the model complexity, approximate methodologies are required to reduce the computation cost. We have developed [12] a generalization of the approximation proposed in [3], which aggregates levels of the Markov model beyond level Q.

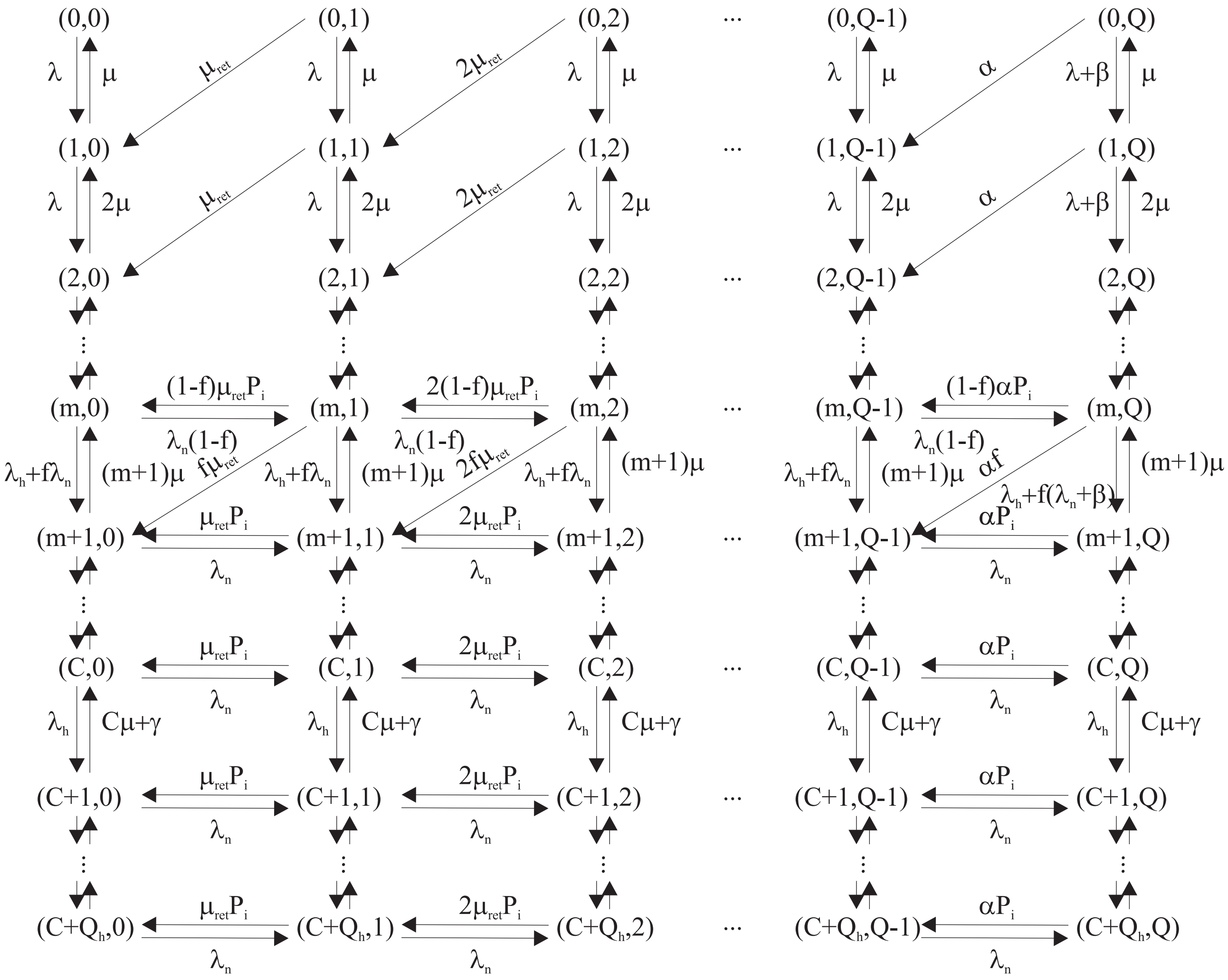
The system described can be modelled as a quasi–birth–and–death (QBD) [13] process with a state space given by $(i, j) : 0 \leq i \leq C + Q_h; 0 \leq j \leq Q$, where i is the number of busy servers plus the number of handovers queued and j is the number of new sessions retrying, when $j < Q$. States (i, Q) correspond to the situation where Q or more users are retrying.

Figure 2 shows the state transition diagram for the proposed model. Two new parameters have been introduced in the last column: M denotes the average number of users retrying when there are Q or more users retrying, and p is the probability that after a successful retrial the number of users retrying is bigger or equal than Q. By balancing the flux rates across the vertical cuts of the transition diagram it is not to difficult to show that

$$p = \frac{(1-f)\pi(m,Q) + \sum_{i=m+1}^{C+Q_h} \pi(i,Q)}{(1-f)[\pi(m,Q) + \pi(m,Q-1)] + \sum_{i=m+1}^{C+Q_h} [\pi(i,Q) + \pi(i,Q-1)]}$$

$$M = \frac{\lambda_n[(1-f)[\pi(m,Q) + \pi(m,Q-1)] + \sum_{i=m+1}^{C+Q_h}[\pi(i,Q) + \pi(i,Q-1)]]}{\mu_{ret}[\sum_{i=0}^{m-1} \pi(i,Q) + (1-f)P_i\pi(m,Q) + f\pi(m,Q) + P_i \sum_{i=m+1}^{C+Q_h} \pi(i,Q)]}$$

As can be observed in Fig. 2 only transitions between states of the same level or between adjacent levels



Being $\alpha = M\mu_{ret}(1-p)$ and $\beta = M\mu_{ret}p$.

Figure 2: State transition diagram.

are possible, thus the infinitesimal generator $\mathbf{Q}$ has the following block tridiagonal structure:

$$\mathbf{Q} = \begin{bmatrix} \mathbf{v}_0^0 & \mathbf{v}_0^+ & \dots & 0 & 0 \\ \mathbf{v}_1^- & \mathbf{v}_1^0 & \dots & 0 & 0 \\ 0 & \mathbf{v}_2^- & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \mathbf{v}_{Q-2}^+ & 0 \\ 0 & 0 & \dots & \mathbf{v}_{Q-1}^0 & \mathbf{v}_{Q-1}^+ \\ 0 & 0 & \dots & \mathbf{v}_Q^- & \mathbf{v}_Q^0 \end{bmatrix}$$

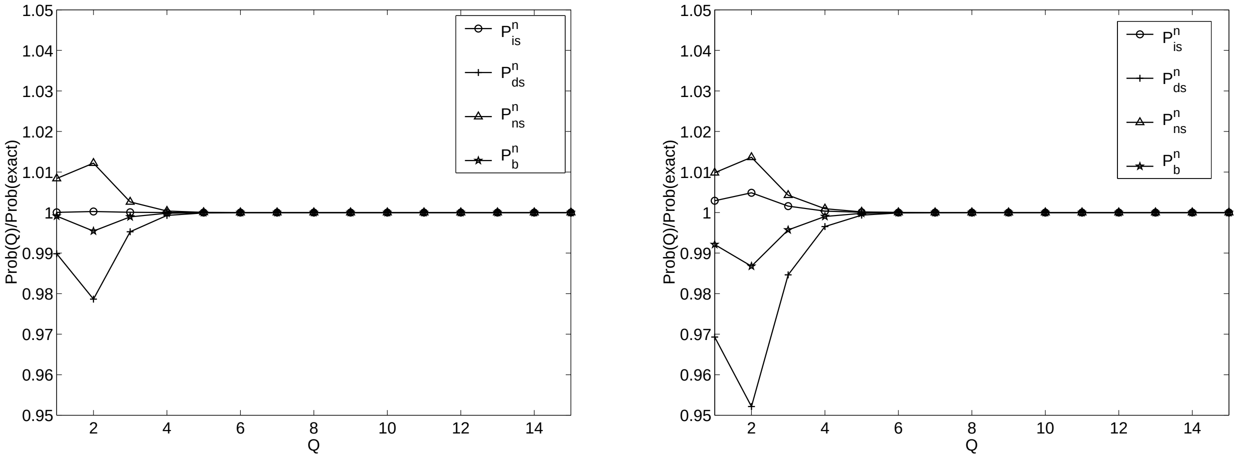
The blocks of $\mathbf{Q}$ are of size $(Q+1)\times(Q+1)$ and their contents are defined in A.

The stationary probability distribution is obtained by solving $\pi\mathbf{Q} = \mathbf{0}$ with the normalization condition $\pi\mathbf{e} = 1$. If $\mathbf{Q}$ is a finite matrix, as in our case, this system can be solved by any of the standard methods defined in classical linear algebra. However, we can exploit the structure of $\mathbf{Q}$ using the algortihm 0 defined in [14], which allows us to reduce the computational cost.

Different performance parameters can be obtained from the stationary probability distribution like:

- Probability of a new session being blocked P_b^n and probability of a new session being served in its first attempt P_{is}^n,

$$P_b^n = \sum_{j=0}^{Q}(1-f)\pi(m,j) + \sum_{i=m+1}^{C+Q_h}\sum_{j=0}^{Q}\pi(i,j); \quad P_{is}^n = (1/\lambda_n)\lambda_n\left[\sum_{i=0}^{m-1}\sum_{j=0}^{Q}\pi(i,j) + f\sum_{j=0}^{Q}\pi(m,j)\right]$$



(a) $\lambda_n = 8$.

(b) $\lambda_n = 12$.

Figure 3: Impact of the approximation on the accuracy of results.

- Probability of obtaining service, but not in the first attempt,

$$P^n_{ds} = (1/\lambda_n)\mu_{ret}\left[\sum_{i=0}^{m-1}\sum_{j=0}^{Q-1} j\pi(i,j) + M\sum_{i=0}^{m-1}\pi(i,Q) + f\left[\sum_{j=0}^{Q-1} j\pi(m,j) + M\pi(m,Q)\right]\right]$$

- Probability of an impatient customer leaving the system without having been served,

$$P^n_{ns} = (1/\lambda_n)\mu_{ret}P_i\left[\sum_{i=m+1}^{C+Q_h}\sum_{j=0}^{Q-1} j\pi(i,j) + M\sum_{i=m+1}^{C+Q_h}\pi(i,Q) + (1-f)\left[\sum_{j=0}^{Q-1} j\pi(m,j) + M\pi(m,Q)\right]\right]$$

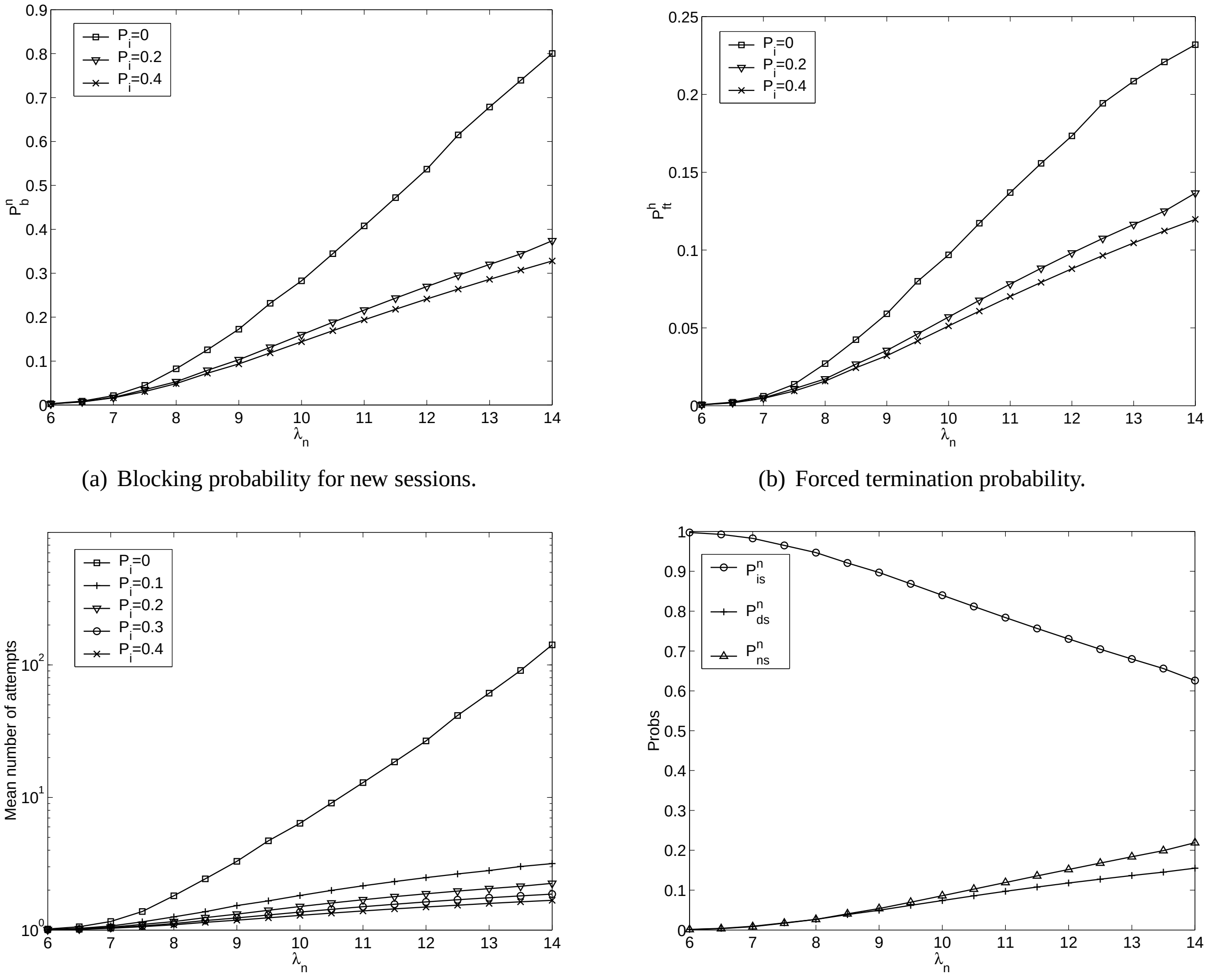
- For handover requests, we define the probability of a session being forced to terminate P^h_{ft} as a function of the probability of a handover being blocked because the queue is full P^h_b and the probability of a handover request abandoning the queue P^h_{ab} ,

$$P^h_{ft} = \frac{\frac{\mu_r}{\mu_s}[P^h_b + P^h_{ab}]}{1 + \frac{\mu_r}{\mu_s}[P^h_b + P^h_{ab}]}; \quad P^h_b = \sum_{j=0}^{Q}\pi(C+Q_h, j); \quad P^h_{ab} = \frac{\mu'_r}{\lambda_h}\sum_{i=C+1}^{C+Q_h}\sum_{j=0}^{Q}(i-C)\pi(i,j)$$

4 Numerical Evaluation

In this section, we evaluate the impact that the different features of the model have on the performance by studying four different scenarios. In all the numerical examples we have used the following default values for the configuration parameters. The capacity of the system is $C = 32$ resource units with a fractional threshold of $t = 31$ and a handover queue of length $Q_h = 1$. The cell residence and session duration time satisfy: $\mu_r/\mu_s = 2$, $\mu_r + \mu_s = 1$ sessions/s. For the mean time in the handover area we used $\mu'_r/\mu_r = 10$. For retrials, we set $P_i = 0.2$ and $\mu_{ret} = 50$ sessions/s, which is around five times the arrival rate of fresh sessions for a load that represents the center of the load range of interest.

To solve the different QBD processes we deployed the approximation described in Section 3. Figure 3(a) and Fig. 3(b) show the variation of the relation $Prob(Q)/Prob(exact)$ with Q for two different new session arrival rates $\lambda_n = 8$ and 12 sessions/s, which correspond to scenarios that could be defined as quite loaded and heavily loaded. The parameter $Prob$ is one of the elements of the set $\{P^n_{is}, P^n_{ds}, P^n_{ns}, P^n_b\}$ being $P^n_b = P^n_{ds} + P^n_{ns}$, $Prob(exact)$ corresponds to the exact values and $Prob(Q)$ corresponds to values obtained for a given value of Q. Therefore, the relation $Prob(Q)/Prob(exact)$ expresses the relative error when using different values for Q. For the results displayed in the rest of the paper we used $Q = 6$ which provides an excellent precision with a very small computation cost.



(a) Blocking probability for new sessions.

(b) Forced termination probability.

(c) Mean number of attempts per new session.

(d) Immediate service, delayed service and non-service probabilities. $P_i = 0.2$.

Figure 4: Performance evaluation.

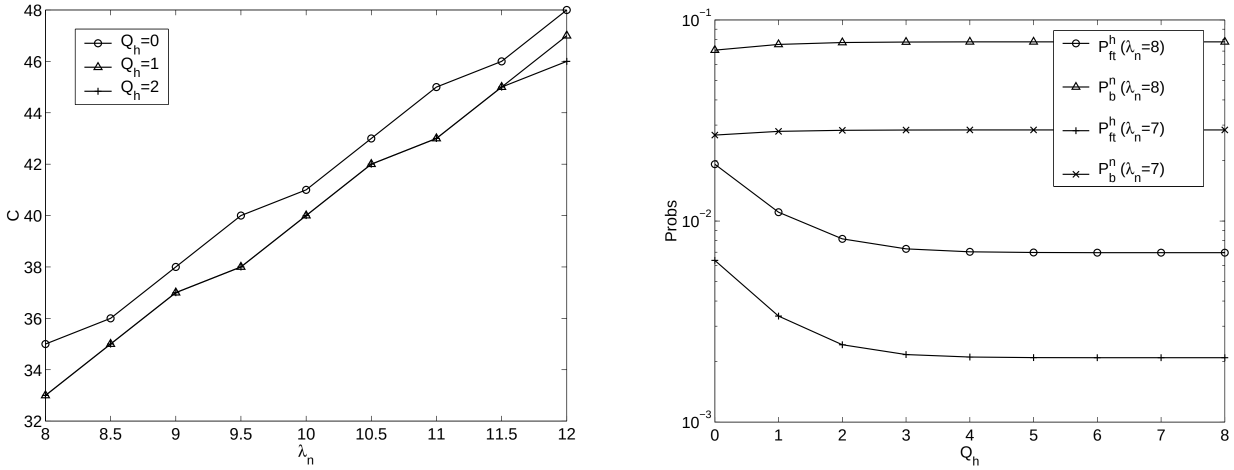
4.1 Model evaluation

In this subsection, we illustrate the impact that the retrial phenomenon has on the quality of service perceived by customers. Figure 4(a) shows the behaviour of the blocking probability for new sessions $P_b^n = P_{ds}^n + P_{ns}^n$ with the arrival rate of new sessions. As λ_n increases, the blocking probability also increases. Note that the blocking probability increases as P_i decreases, i.e. as the probability that users leave the system due to impatience decreases, which shows the negative impact of retrials in system performance. Similar conclusions can be drawn for the variation of the forced termination probability displayed in Fig. 4(b).

We define $\eta = (\lambda_n + \lambda_{ret})/\lambda_n$ as the mean number of attempts per new session, i.e. the mean number of resource assignment requests per each fresh new session arrival, where $\lambda_{ret} = E[\zeta]\mu_{ret}$ is the average reattempt rate and $E[\zeta]$ is the mean number of users in the retrial orbit. Figure 4(c) displays η as a function of the new session arrival rate λ_n. It is clear that as P_i decreases the mean number attempts increases, but notice that the increase jumps drastically from $P_i = 0.1$ to $P_i = 0$. Obviously, η also increases with λ_n. We have also represented the common performance parameters for retrial systems in Fig. 4(d). As a typical scenario, we have taken $P_i = 0.2$.

4.2 Handover queue impact

In this subsection we quantify the impact of the handover queue length (Q_h) on system performance by means of two different studies.



(a) Impact of Q_h on system dimensioning.

(b) Impact of Q_h on the blocking probabilities.

Figure 5: Impact of the length of the queue for handover requests (Q_h).

First, Fig. 5(a) displays the impact of Q_h on the dimensioning process. The dimensioning process consists of computing the minimum number of resource units C and threshold t for a given new session arrival rate λ_n, in order to guarantee that $P_b^n \leq 0.05$ and $P_b^h \leq 0.005$. As expected, higher values of C are required as the system load increases. Note that using a queue for handover requests ($Q_h \neq 0$) has a positive effect, as less resource units are required to achieve the quality of service objective. Note also that only a few queue positions are required to perceive the benefit and that deploying a higher number of them do not have any impact, except in overloaded scenarios.

Second, Fig. 5(b) shows the impact of the handover queue length on the probability of forced termination. As seen, increasing Q_h has a positive influence on P_{ft}^h. As before, small values for Q_h are sufficient to benefit from a reduction in P_{ft}^h and higher values for Q_h do not have any impact. It is also interesting to observe that increasing Q_h has a negligible impact on P_b^n.

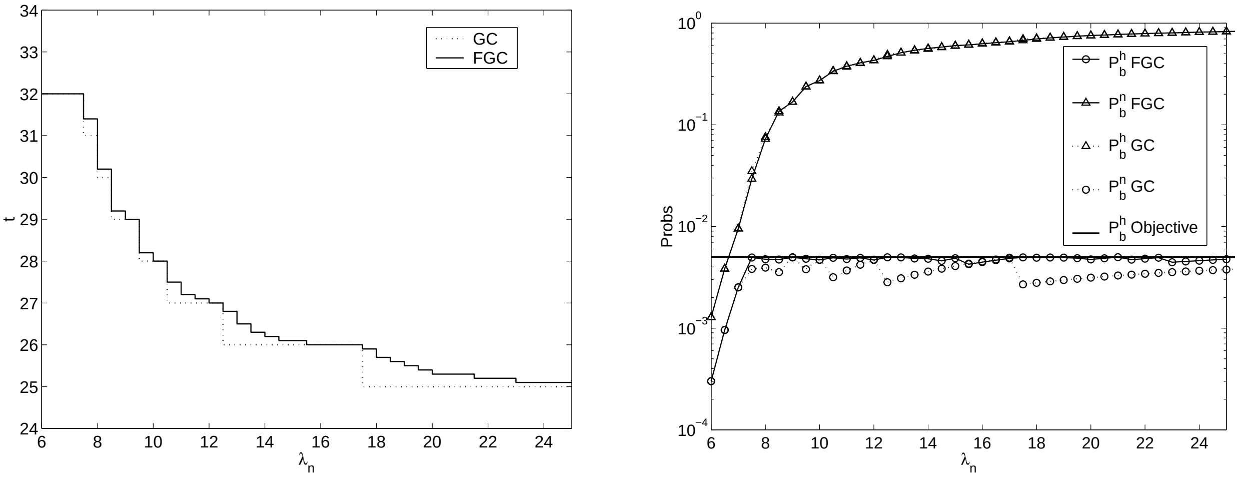
4.3 Fractional guard channel impact

In this subsection we study the impact of the number of fractional guard channels on the system performance. We use a dimensioning process similar to the one described in the previous subsection. Basically, we determine the optimum value of t in order to guarantee that $P_b^h \leq 0.005$.Figure 6(a) shows the required t to meet the quality of service objective, while Fig. 6(b) displays the variation of P_b^n and P_b^h with the arrival rate of new sessions when deploying the optimum value of t. In both figures, we display the results when reserving a fractional number of guard channels (FGC) and an integer number of them (GC). Observe that as system load increases the required value of t decreases in order to meet the P_b^h objective. This in turn decreases the mean number of resources that new sessions have access to, increasing in this way P_b^n. Note that the when deploying a fractional number of guard channels the objective is met with more precision and, although not shown, more traffic is carried.

4.4 Redimensioning with retrials

It is commonly accepted that if the retrial phenomenon is ignored during the planning phase a system over-dimensioning might occur, basically due to the extra load that the retries represent. In this subsection we evaluate the magnitude of the over-dimensioning. For this purpose we dimension the system in two scenarios, one in which the extra load is known to appear as a result of retrials and a second one in which it is perceived as an increment in the arrival rate of new sessions ($\lambda_n' = \lambda_n + \lambda_{ret}$).

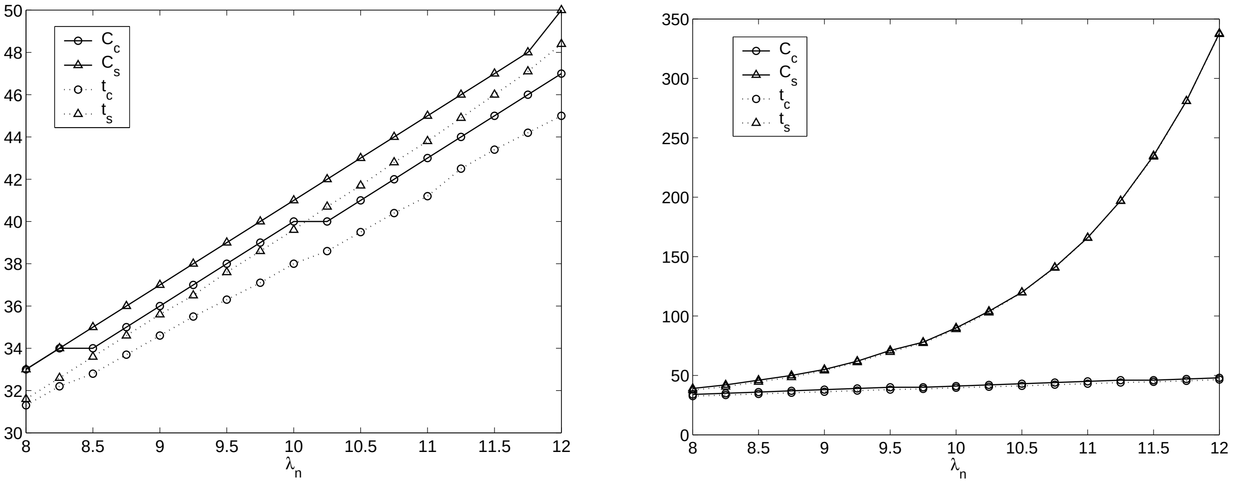
For each arrival rate of new sessions we obtain the retrial rate (λ_{ret}) according to the model of Fig. 1. In the scenario where retrials are known to happen, the handover rate is determined by balancing the input and output handover rates to the cell. The same value for the handover rate is used when retrials are perceived as an increment in the arrival rate of new sessions. Otherwise, a higher handover rate would had been obtained



(a) Optimum value of t to guarantee $P_b^h \leq 0.005$.

(b) Variation of P_b^n and P_b^h using the optimum t.

Figure 6: Impact of the number of fractional guard channels (t).



(a) $P_i = 0.2$

(b) $P_i = 0.0$

Figure 7: Redimensioning of resources taking and not taking retries into account.

making the comparison of both scenarios less realistic. Then we determine the number of resource units C and the optimum number of fractional guard channels t required to guarantee that $P_b^n \leq 0.05$ and $P_b^h = 0.005$.

Figure 7 displays the result of the dimensioning process in the two scenarios described above and for $P_i = 0.2$ and $P_i = 0$. C_c and t_c are the number of resource units and fractional guard channels required when using the model in which retrials are known to happen and C_s and t_s are the same parameters but for the scenario in which retrials are perceived as an increment in the arrival rate of new sessions.

As observed in Fig. 7, the number of resource units required increases as load increases in both scenarios, in order to met the quality of service requirements. However, perceiving retrials as an increment in the arrival rate of new sessions leads to a severe over-provisioning, specially when $P_i = 0$. Although the scenario with $P_i = 0$ might seem exaggerate, mobile terminals can be equipped with automatic redialing [1] and therefore, as observed in Fig. 4(c), considering a mean number of retrials around 10 for a moderate overload system ($\lambda_n = 10$) is not too unrealistic.

5 Conclusion

We evaluated the impact that new session retrials have on the performance of a mobile cellular network which deploys a fractional number of guard channels, a queue for handover sessions and an exponential deadline for serving those requests. These las features model the overlapping area between adjacent cells. We considered that such model has not been sufficiently explored in the literature.

We developed an approximate methodology that is a generalization of a previous proposal, which precision was not satisfactory when dealing with parameters different from blocking probabilities, like the probabilities of immediate service, delayed service and no service. Our proposal is substantially more accurate while increasing the computation cost only marginally.

Results show that only a few waiting positions are required to perceive a substantial reduction of the forced termination probability. We also showed that deploying a fractional number of guard channels allows the operator to adjust the handover blocking probability with more precision. Deploying both a queue for handover requests and a fractional number of guard channels helps to increase system capacity while meeting the required QoS objective.

Finally, we evaluated the magnitude of the over-provisioning required to meet a given QoS objective when retrials are perceived as an increment in the arrival rate of new sessions, showing that it can be severe when the terminals retry persistently as might occur when equipped with automatic redialing.

References

[1] E. Onur, H. Deliç, C. Ersoy and M.U. Çaglayan, "Measurement-based replanning of cell capacities in GSM networks," *Computer Networks*, vol. 39, pp. 749–767, 2002.

[2] G. Falin, "A survey of retrial queues," *Queueing Systems*, vol. 7, pp. 127–168, 1992.

[3] M.A. Marsan, G. De Carolis, E. Leonardi, R. Lo Cigno and M. Meo, "Efficient estimation of call blocking probabilities in cellular mobile telephony networks with customer retrials." *IEEE J. Sel. Areas in Commun.*, vol. 19, no. 2, pp. 332–346, Feb. 2001.

[4] P. Tran-Gia and M. Mandjes, "Modeling of customer retrial phenomenon in cellular mobile networks," *IEEE J. Sel. Areas in Commun.*, vol. 15, no. 8, pp. 1406–1414, Oct. 1997.

[5] C.D. Carothers, R.M. Fujimoto and Y-B. Lin, "A Re-dial Model for Personal Communications Services Networks," in Proc. IEEE 45th Vehicular Technology Conference (VTC '95) , Jul. 1995, pp. 135-139.

[6] N. Shingawa, T. Kobayashi, K. Nakano and M. Sengoku, "Teletraffic characteristics in prioritized handoff control method considering reattempt calls," *IEICE Trans. Commun.*, vol. E83-B, no. 8, pp. 1810-1818, Aug. 2000.

[7] A.E. Xhafa and O.K. Tonguz, "Dynamic priority queueing of handover calls in wireless networks: An analytical framework," *IEEE J. Sel. Areas in Commun.*, vol. 22, no. 5, pp. 904–916, Jun. 2004.

[8] D. García, J. Martínez and V. Pla, "Admission control policies in multiservice cellular networks: optimum configuration and sensitivity," Wireless Systems and Mobility in Next Generation Internet, Gabriele Kotsis and Otto Spaniol (eds.), *Lecture Notes in Computer Science (LNCS)*, vol. 3427, pp. 121–135, Springer Verlag 2005.

[9] M. A. Marsan, G. De Carolis, E. Leonardi, R. Lo Cigno and M. Meo, "How many cells should be considered to accurately predict the performance of cellular networks?," *in Proc. European Wireless*, Munich, Germany, October 1999.

[10] V. Pla and V. Casares-Giner, "Analytical-numerical study of the handoff area sojourn time," *in Proc. IEEE GLOBECOM*, Nov. 2002, pp. 886-–890.

[11] R. Ramjee, R. Nagarajan and D. Towsley, "On optimal call admission control in cellular networks," *Wireless Networks Journal (WINET)*, vol. 3, no. 1, pp. 29–41, 1997.

[12] M. J. Domenech-Benlloch, J. M. Gimenez-Guzman and V. Casares-Giner, "Modelos Markovianos para la resolución de sistemas con reintentos. Evaluación de diferentes metodologías," *in Proc. IV Jornadas de Ingeniería Telemática. JITEL'03*, pp. 9–16, Gran Canaria, Spain, September 2003.

[13] M. F. Neuts, *“Matrix Geometric Solutions in Stochastic Models: An Algorithmic Approach,”* The John Hopkins University Press, Baltimore, 1981.

[14] L.D. Servi, “Algorithmic solutions to two-dimensional birth-death processes with application to capacity planning,” *Telecommunication Systems*, vol. 21, no. 2-4, pp. 205–212, 2002.

A Q blocks

Matrices $\mathbf{v}_i^+$ define the transitions from states $(i,j) \rightarrow (i+1,k)$ and take the next values:

$$
\begin{aligned}
\mathbf{v}_i^+ &= \begin{bmatrix} \lambda & 0 & 0 & \dots & 0 & 0 \\ \mu_{ret} & \lambda & 0 & \dots & 0 & 0 \\ 0 & 2\mu_{ret} & \lambda & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda & 0 \\ 0 & 0 & 0 & \dots & \alpha & \lambda+\beta \end{bmatrix} \text{ when } i \in [0, m-1] \\
\mathbf{v}_m^+ &= \begin{bmatrix} \lambda_h + f\lambda_n & 0 & 0 & \dots & 0 & 0 \\ f\mu_{ret} & \lambda_h + f\lambda_n & 0 & \dots & 0 & 0 \\ 0 & 2f\mu_{ret} & \lambda_h + f\lambda_n & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \lambda_h + f\lambda_n & 0 \\ 0 & 0 & 0 & \dots & \alpha f & \lambda_h + f(\beta+\lambda_n) \end{bmatrix} \\
\mathbf{v}_i^+ &= \lambda_h \mathbf{I} \text{ when } i \in [m+1, C+Q_h]
\end{aligned}
$$

Where **I** is a $(Q+1) \times (Q+1)$ identity matrix.
Matrices $\mathbf{v}_i^0$ define the transitions from states $(i,j) \rightarrow (i,k)$ taking the next values:

$$
\begin{aligned}
\mathbf{v}_i^0 &= *\mathbf{I} \text{ when } i \in [0, m-1] \\
\mathbf{v}_m^0 &= \begin{bmatrix} * & (1-f)\lambda_n & 0 & \dots & 0 & 0 \\ (1-f)\mu_{ret}P_i & * & (1-f)\lambda_n & \dots & 0 & 0 \\ 0 & 2(1-f)\mu_{ret}P_i & * & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & * & (1-f)\lambda_n \\ 0 & 0 & 0 & \dots & (1-f)\alpha P_i & * \end{bmatrix} \\
\mathbf{v}_m^0 &= \begin{bmatrix} * & \lambda_n & 0 & \dots & 0 & 0 \\ \mu_{ret}P_i & * & \lambda_n & \dots & 0 & 0 \\ 0 & 2\mu_{ret}P_i & * & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & * & \lambda_n \\ 0 & 0 & 0 & \dots & \alpha P_i & * \end{bmatrix} \text{ when } i \in [m+1, C+Q_h]
\end{aligned}
$$

Note that $*$ are the values that make the sum of every row of **Q** equal to 0.
Finally, matrices $\mathbf{v}_i^-$ define the transitions from states $(i,j) \rightarrow (i-1,k)$ taking the next values:

$$
\begin{aligned}
\mathbf{v}_i^- &= i\mu\mathbf{I} \text{ when } i \in [1, C] \\
\mathbf{v}_i^- &= [C\mu + (i-C)\gamma]\mathbf{I} \text{ when } i \in [C+1, C+Q_h]
\end{aligned}
$$