Performance Analysis of Dedicated Signalling Channels in GERAN by Retrial Queues

Salvador Luna-Ramírez*, Matías Toril* and Volker Wille†

February 15, 2010

Abstract

GERAN (GSM-EDGE Radio Access Network) operators have traditionally used the Erlang B formula to estimate the number of signalling channels on a per-cell basis. Thus, it is assumed that the network behaves as a loss system with Poisson arrivals. However, the presence of automatic retrial mechanisms and correlated arrivals in these channels suggests that these assumptions might not be valid. This paper presents a performance analysis of the *Stand-alone Dedicated Control CHannel* (SDCCH) in GERAN. Preliminary analysis shows that the Erlang B formula underestimates congestion and blocking on this channel. To address this issue, a queueing model with retrials and correlated arrivals is proposed, where correlation between arrivals is modelled by a simple Markov-Modulated Poisson Process. The proposed model can be tuned on a per-cell basis by statistics in the Network Management System. Model assessment is based on performance statistics from a live GERAN system. Results show that a simple retrial queueing model fails to explain blocking in cells with a large number of channels. These limitations are overcome by adding correlated arrivals in the retrial model.

1 Introduction

In GERAN, signalling capacity largely depends on the capacity of the *Stand-alone Dedicated Control CHannel* (SDCCH). This radio interface channel is involved in mobility management procedures, namely call set-up, mobile station registration and location update, as well as in data services, such as *Short Message Service* (SMS), *Multimedia Messaging Service* (MMS) and *Wireless Application Protocol* (WAP), [1]. Hence, SDCCH congestion must be avoided to minimise revenue loss.

During network design, operators have to estimate the required number of SDCCHs on a per-cell[1] basis. Traditionally, the Erlang B formula has been used to estimate the minimum number of these channels based on predictions of the signalling traffic, [2]. The application of this formula assumes that: a) the request arrival process is a Poisson process, b) blocked attempts are cleared, and c) the number of users is large. Although these assumptions are known to be valid for voice traffic, [3], none of them necessarily holds true for the SDCCH traffic. On the one hand, automatic retrial/redial mechanisms incorporated in mobiles cause repeated attempts during congestion periods, [4]. On the other hand, SDCCH requests are correlated in some cells, as will be shown later. In addition, some cells show large signalling traffic from a few terminals. In all these cases, the Erlang B formula fails to give accurate predictions. Due to these limitations, operators used to over-dimension SDCCH resources in the early days of GERAN. However, such an approach is not financially viable anymore as cellular operators have to maximise the usage of each and every time slot in the network to maximise their return on investment.

*Communications Engineering Dept., University of Málaga, Málaga, Spain, {sluna,mtoril}@ic.uma.es

†Performance Services, Nokia Siemens Networks, Huntingdon, UK, volker.wille@nsn.com

[1]A cell is the geographical area served by a base station.

Many teletraffic models have been proposed for cellular networks since Hong and Rappaport proposed their Markov chain model, [5]. Subsequent studies improved the model by considering more general distributions of channel holding times, [6], correlated arrivals, [7][8], retrials, [9][10], multiple services, [11][12], and multiple network layers, [13][14]. All these models were conceived and tested for user traffic channels. However, to the authors' knowledge, no study has been published checking the validity of these models for dedicated signalling channels based on real network data.

In this paper, a comprehensive analysis of the SDCCH is performed over measurements from a live GERAN system. Preliminary analysis shows that the Erlang B formula fails to predict SDCCH congestion and blocking in many cells. Having identified retrials as a source of inaccuracy, a simplified queueing model is presented to evaluate the influence of retrials on SDCCH performance. Such a model extends that in [9] by considering a mixture of services with and without retrials. Then, the model is improved by including correlated arrivals by a Markov-Modulated Poisson Process, as in [15]. The resulting model includes parameters that can be tuned on a per-cell basis using statistics in the Network Management System. Model assessment is carried out by comparing performance estimates obtained by the model against measurements taken from a live network. Results show that, once the proposed model is tuned on a per-cell basis, it clearly outperforms models currently used by operators to re-plan SDCCH resources.

A large number of papers have studied the problem of retrials in both wired and wireless networks. For a survey on retrial queues, the reader is referred to [16]. Early references on the effects of retrials are [17] and [18]. Performance analysis of standard multi-server retrial systems, considering Poisson arrivals, exponential service times and exponential inter-retrial times, is presented in [19][20][21][22]. In [23], the analysis is extended to retrial systems with correlated arrivals. In the context of cellular networks, the previous models have been extended to consider handovers, [9][10][24], automatic retrials and user's redials, [4][25], and more general distributions of inter-arrival, service and inter-retrial times, [15][24]. This paper applies for the first time well-known principles and techniques of retrial queues to the analysis of the SDCCH. The main contributions of this paper are: a) to show the limitations of the Erlang loss model for dedicated signalling traffic in GERAN, b) to prove that such limitations are due to time correlation between arrivals, c) to propose an accurate retrial queueing model for the SDCCH, which, unlike more refined models, can easily be tuned from network statistics, and d) to compare SDCCH performance estimates obtained by different queueing models against real network measurements. The rest of the paper is organised as follows. Section 2 outlines the SDCCH re-planning problem from the operator's point of view. Section 3 presents two retrial queueing models for the SDCCH. Section 4 compares performance estimates obtained by the models with real network measurements. Finally, Section 5 presents the conclusions of this work.

2 The SDCCH Re-planning Problem

Cellular traffic tends to be unevenly distributed both in time, [26], and space, [27]. Fast fluctuations in traffic demand are dealt with by complex radio resource management features. In contrast, permanent congestion problems can only be solved by proper dimensioning of traffic resources on a cell basis.

The planning of SDCCH resources aims to minimise persistent congestion problems by a proper selection of SDCCH capacity on a per-cell basis. The main design parameter is the number of time slots dedicated to SDCCH on a permanent basis. Each time slot can comprise 4 or 8 sub-channels, [28]. Therefore, the number of SDCCH sub-channels in a cell, N, is a multiple of 4. In some networks, one of these sub-channels is used for the Cell Broadcast CHannel (CBCH), in which case N takes values in the set $4i-1$, $i \in \mathbb{N}_+$.

For re-planning purposes, SDCCH statistics are collected by the Network Management System (NMS) on an hourly basis. The key performance indicators are the SDCCH blocking ratio (i.e., ratio of blocked attempts), BR, and the SDCCH congestion ratio (i.e., ratio of time without free sub-

channels), CR. Vendor equipment also provides the average SDCCH carried traffic, A_c, the mean SDCCH holding time, MHT, and the number of offered, blocked and successfully carried SDCCH attempts per hour. It should be pointed out that the latter counters include both fresh and retrial attempts, as the network cannot differentiate between them. In addition, the number of carried attempts is also broken down by establishment causes: *Mobile Originated Call* (MOC), *Mobile Terminated Call* (MTC), *Emergency Call* (EC), *call Re-Establishment* (RE), *Location Update* (LU), *IMSI Detach* (ID), *Supplementary Service* (SS), *Short Message Service* (SMS) and *GHost seizure* (GH), [29]. The latter reflects SDCCH seizures that time out due to false requests in the Random Access CHannel (RACH).

Based on these measurements, operators re-assign SDCCH resources, at most, on a weekly basis. Due to bad planning, some cells experience unacceptable SDCCH blocking during operation. Network operators consider blocking ratios larger than 10^{-2} unacceptable. Note that if a MOC or MTC attempt is blocked on the SDCCH, the call is lost. Even if some schemes allow using spare traffic channels temporarily for signalling purposes, this cannot be relied on as peaks of signalling and call traffic tend to be correlated, [30]. Hence, operators counteract SDCCH blocking by increasing the number of sub-channels, N, in problematic cells. Subsequent addition of new cells often causes that SDCCH resources on existing cells become unnecessary, which cannot be detected without a precise performance model. Unfortunately, such a model is not currently available due to retrials and correlated arrivals in the SDCCH. As a result, SDCCH resources are over-dimensioned in many cells and under-dimensioned in others, [30]. This problem can be solved by an improved performance model that can be tuned on a per-cell basis. As main benefit, many time slots unnecessarily assigned to SDCCH could be converted into Traffic CHannels (TCHs).

The problem treated here has important similarities with that reported in [4]. In that paper, a simple analytical model was proposed to estimate, for each cell, the average number of retrials and redials per fresh call attempt in user traffic channels by using only NMS measurements. The main differences for the SDCCH are: a) the mixture of services with very different properties, and b) the presence of correlated arrivals.

3 System Models

This section describes two queueing models for the SDCCH. The first one considers retrials and the second one extends the previous one by adding correlated arrivals.

3.1 Retrial Model

The basic retrial model is based on that presented in [9]. Such a model considers a single cell in which repeated attempts occur. Each terminal can be in one of three states: *idle*, *active* and *wait-for-reattempt*. After finishing a transaction, a terminal is in idle state until generating the next fresh attempt. In case of rejection, the terminal enters the wait-for-reattempt state (also referred to as *retrial orbit*) with retrial probability θ or abandons with probability (1-θ). The durations of all states are assumed to be exponentially distributed, and, hence, the system can be modelled by a Markov chain. For simplicity, it is assumed here that the population in a cell is infinite.

The above-described model is extended here to differentiate between services with and without retrials in the SDCCH. The resulting model, referred to as *Retrial Model* (RM), is shown in Figure 1 (a). The arrival flow is divided into two components, namely retrial and non-retrial traffic, depending on whether blocked attempts are repeated or not. For simplicity, it has been assumed that all services except GH are repeated until success (i.e., θ=1). This assumption is reasonable, because, even if automatic retrials by the terminal fail, re-dialing is only a matter of pushing a button in current handsets. For computational efficiency, the number of users in the orbit is artificially limited to M to keep the number of system states finite, [31]. The main parameters in the model are the total arrival rate for services with and without retrials, λ_r and λ_{nr}, the service rate (i.e., the inverse of the mean channel holding time), μ, the retrial rate (i.e., the inverse of the mean time between retrials), α, the number of sub-channels, N, and the size of the orbit, M. Figure 1 (b) shows the state transition diagram, where the state of the system (i, j) is described by the number of busy SDCCH sub-channels, i, and the number of requests waiting for re-attempt (i.e., the number of users in the orbit), j.

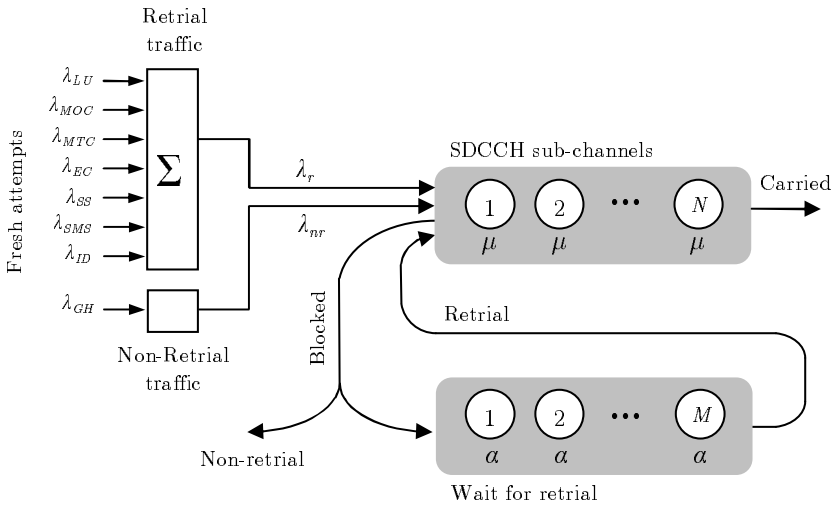
The values of all parameters in the model can be obtained from measurements gathered on a cell and hourly basis in the NMS. The total arrival rate for services with retrials, λ_r, is obtained directly from the number of seizures per hour (note that, for these services, offered and carried traffic coincides, since it has been assumed that θ=1). For services without retrials (i.e., GH), the arrival rate, λ_{nr}, is estimated from the congestion ratio, CR, as

$$\lambda_{nr} = \frac{N_{GH}}{3600(1 - CR)} \,, \tag{1}$$

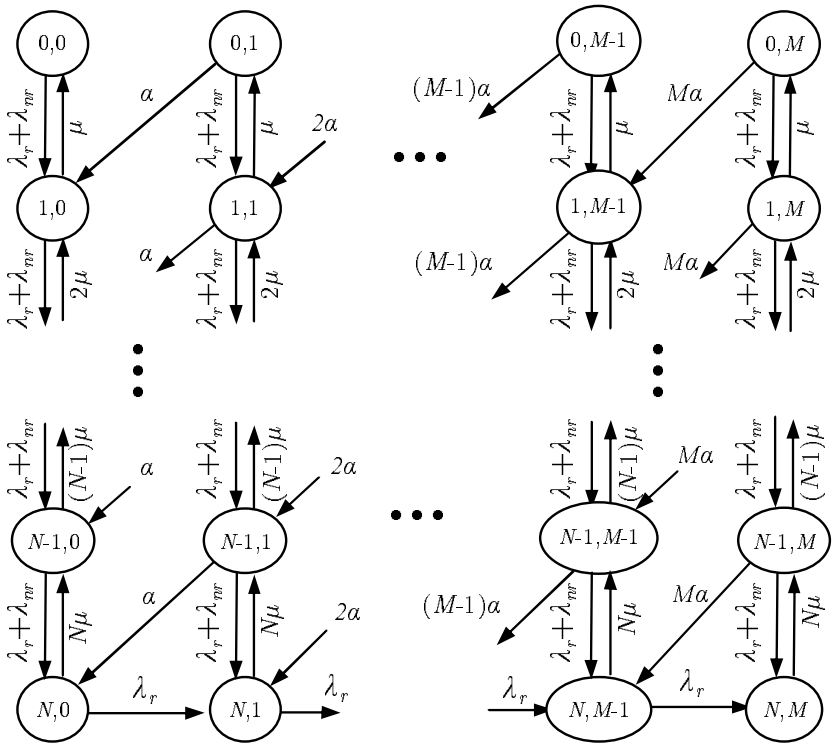
where N_{GH} is the number of ghost seizures in one hour. The service rate, μ, is the inverse of the SDCCH mean holding time. The retrial rate, α, is fixed to the value configured network wide by the operator. Finally, the size of the orbit, M, must be chosen so that the probability that the orbit is full is negligible, which depends on the traffic conditions. In the previous model, it is assumed that the time between consecutive retrials is exponentially distributed to ensure mathematical tractability, even if this parameter takes a deterministic value in a real system. It is expected that this assumption has a negligible impact on key performance indicators, as shown in [25].

The reader is referred to [16] for a performance analysis of this classical retrial queue. It should be pointed out that an analytical expression for steady-state probabilities in RM is only available for $N = 1$ and 2. For $N \geq 3$, the problem does not preserve the birth-and-death structure and, consequently, no closed-form expression can be found, [16]. Thus, teletraffic performance indicators can only be obtained by computing the stationary distribution of the Markov chain describing system dynamics by numerical methods. This is achieved by solving the system of linear equations

$$\Pi \, \mathbf{Q} = 0 \,, \;\; \Pi \, e = 1 \,, \;\; \Pi \geq 0 \,, \tag{2}$$



(a) System model



(b) State transition diagram

Figure 1: The basic retrial model.

where Π is the steady-state probability vector, $\mathbf{Q}$ is the infinitesimal generator matrix and e is a column vector of ones, [32]. The reader is referred to [21] for the value of $\mathbf{Q}$ for the retrial queue in Figure 1. Once the stationary distribution, Π, is obtained, the carried traffic, $A_{c_{rm}}$, the congestion ratio, CR_{rm}, and the blocking ratio, BR_{rm}, are calculated as

$$A_{c_{rm}} = \sum_{i=0}^{N} \sum_{j=0}^{M} i\, \Pi(i,j) \;, \tag{3}$$

$$CR_{rm} = \sum_{j=0}^{M} \Pi(N,j) \tag{4}$$

and

$$BR_{rm} = \frac{\sum\limits_{j=0}^{M} \left[(\lambda_r + \lambda_{nr} + j\alpha)\, \Pi(N,j)\right]}{\sum\limits_{i=0}^{N} \sum\limits_{j=0}^{M} \left[(\lambda_r + \lambda_{nr} + j\alpha)\, \Pi(i,j)\right]} \;, \tag{5}$$

where $\Pi(i,j)$ is the probability of having i busy sub-channels and j users in the orbit.

3.2 Retrial Model with Correlated Arrivals

In mobile networks, user location is constantly updated by LU requests sent on the SDCCH. A LU request is triggered when a subscriber crosses the border of location areas into which the network is divided. For subscribers moving in groups (e.g., in public transport), the boundary-crossing event is synchronised, [7]. As a result, LU requests tend to concentrate in short periods of time in cells on the border of location areas. To account for this effect, the proposed model considers time correlation of LU attempts.

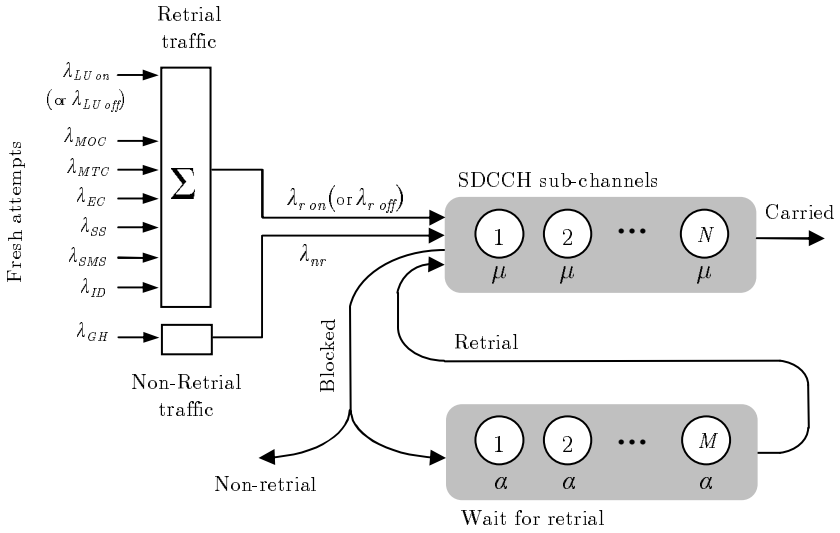
For simplicity, time correlation between fresh LU attempts is modelled by a switched Poisson process, [33]. In this case, the LU arrival rate switches between two values (denoted as *on* and *off* values) with a certain frequency. It is assumed that the duration of the *off* and *on* periods is exponentially distributed. The resulting model, referred to as *Retrial Model with Correlated Arrivals* (RMCA), is shown in Figure 2. The new parameters in the model are: a) the LU arrival rate during the *on* and *off* periods, $\lambda_{LU_{on}}$ and $\lambda_{LU_{off}}$, and b) the switching rates between LU activity states, $\rho_{on\text{-}off}$ and $\rho_{off\text{-}on}$ (or, equivalently, the mean duration of the *on* and *off* states, τ_{on} and τ_{off}). Note that the measured average LU arrival rate, λ_{LU}, is the weighted average of the attempt rates during the *on* and *off* periods

$$\lambda_{LU} = \frac{\lambda_{LU_{on}} \tau_{on} + \lambda_{LU_{off}} \tau_{off}}{\tau_{on} + \tau_{off}} \;. \tag{6}$$

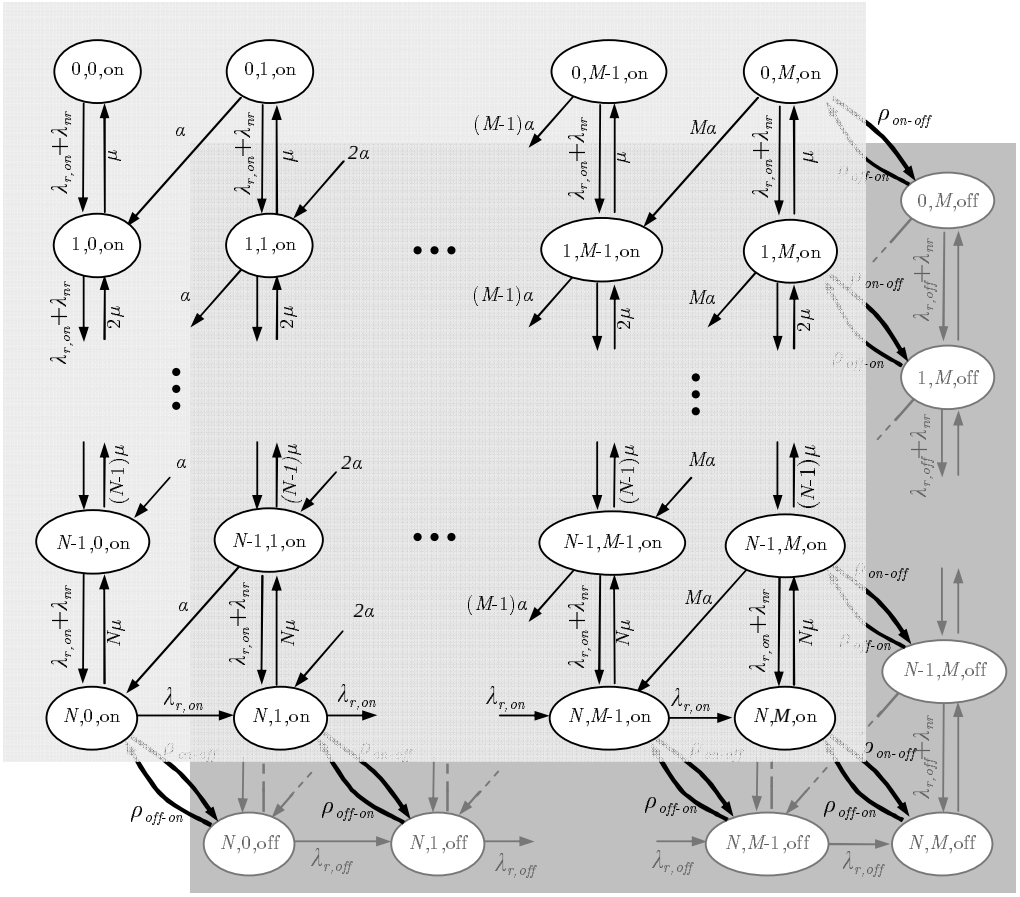
Figure 2 (b) presents the tri-dimensional state transition diagram, which extends that in Figure 1 (b) by considering two LU traffic intensity states, denoted as *on* and *off*. Thus, the state of the system (i,j,k) is described by the number of busy SDCCH sub-channels, i, the number of requests waiting for re-attempt, j, and the LU-activity state, k.

3.2.1 Computation of Queueing Performance Indicators

As in RM, there is no closed-form expression that relates performance indicators and model parameters in RMCA, since RMCA is a generalisation of RM. Thus, queueing performance can only be estimated by computing the stationary distribution numerically from (2). For brevity, the reader is referred to [23] for the value of $\mathbf{Q}$ for the retrial queue in Figure 2, as a special case of the MAP/M/C retrial queue. In both RM and RMCA, $\mathbf{Q}$ has a block tri-diagonal structure if states



(a) System model



(b) State transition diagram

Figure 2: The proposed retrial model with correlated arrivals.

are enumerated in lexicographic order, suggesting the use of block gaussian elimination for solving (2). This is the approach followed by Gaver, Jacobs and Latouche, [34]. The numerical complexity and stability of their approach is analysed in [34]. Numerical stability is ensured for square block matrixes and positive values in $\mathbf{Q}$, which is the case for RM and RMCA. The complexity of solving (2) is $O(MN^3)$ for RM and $O(M(2N)^3)$ for RMCA, where M is the size of the orbit and N the number of sub-channels. This complexity is similar to other block gaussian elimination methods (e.g., [35][36]) and much lower than classical Gauss-Jordan techniques, whose complexity is $O((MN)^3)$ and $O((2MN)^3)$ for RM and RMCA, respectively, [37].

The resulting stationary distribution, Π, is used to compute the carried traffic, $A_{c_{rmca}}$, the congestion ratio, CR_{rmca}, and the blocking ratio, BR_{rmca}, as

$$A_{c_{rmca}} = \sum_{i=0}^{N} \sum_{j=0}^{M} \sum_{k \in \{on, off\}} i\, \Pi(i,j,k) \;, \tag{7}$$

$$CR_{rmca} = \sum_{j=0}^{M} \sum_{k \in \{on, off\}} \Pi(N,j,k) \tag{8}$$

and

$$BR_{rmca} = \frac{\sum_{j=0}^{M} \left[(\lambda_{r_{on}} + \lambda_{nr} + j\alpha)\, \Pi(N,j,on) \;+\; (\lambda_{r_{off}} + \lambda_{nr} + j\alpha)\, \Pi(N,j,off) \right]}{\sum_{i=0}^{N} \sum_{j=0}^{M} \left[(\lambda_{r_{on}} + \lambda_{nr} + j\alpha)\, \Pi(i,j,on) \;+\; (\lambda_{r_{off}} + \lambda_{nr} + j\alpha)\, \Pi(i,j,off) \right]} \;, \tag{9}$$

where $\Pi(i,j,k)$ is the probability of having i busy sub-channels and j users in the orbit in the LU-activity state k, where $k \in \{on, off\}$.

3.2.2 Performance Sensitivity Analysis

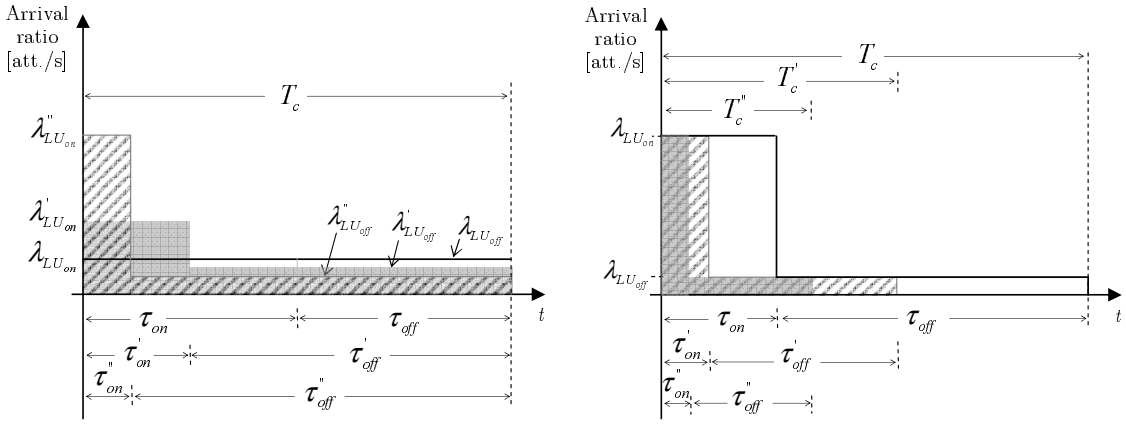
It is interesting to describe the effect of model parameters on queueing performance for RMCA. In the absence of an analytical expression that relates system performance and model parameters, a sensitivity analysis is presented here. The analysis is focused on parameters that reflect time correlation between new SDCCH requests (i.e., $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off}). Note that RMCA reduces to RM when $\lambda_{LU_{on}} = \lambda_{LU_{off}}$. For each combination of parameter values, a new matrix $\mathbf{Q}$ is generated and the steady-state probabilities are obtained by the Gaver, Jacobs and Latouche's algorithm. For simplicity, it is assumed in the sensitivity analysis that all SDCCH traffic is due to LUs (i.e., $\lambda_r = \lambda_{LU}$, $\lambda_{nr} = 0$). The retrial rate, α, is set to $1/6\, s^{-1}$, as fixed by cellular operators. Likewise, the service rate, μ, is set to $1/6\, s^{-1}$, according to real measurements of the SDCCH mean holding time.

The first experiment, described in Figure 3 (a), aims to quantify the impact of concentrating traffic demand in short periods of time. For this purpose, the duration of an *on-off* cycle, T_c $(= \tau_{on} + \tau_{off})$, is fixed to one measurement period in the NMS (i.e., one hour). Then, the number of attempts in each state, $\lambda_{LU_{on}} \cdot \tau_{on}$ and $\lambda_{LU_{off}} \cdot \tau_{off}$, is fixed to be the same, i.e.,

$$\lambda_{LU_{on}} \tau_{on} = \lambda_{LU_{off}} \tau_{off} \;. \tag{10}$$

Finally, the length of the *on* period is progressively reduced. As observed in Figure 3 (a), as τ_{on} is reduced, the value of $\lambda_{LU_{on}}$ increases and the value of $\lambda_{LU_{off}}$ decreases in order to satisfy (10). Thus, the degree of time correlation between arrivals is controlled by the ratio $r = \tau_{on}/\tau_{off}$. From (6) and (10), it can be derived that

$$\lambda_{LU_{on}} = \frac{\lambda_{LU}(\tau_{on} + \tau_{off})}{2\tau_{on}} = \frac{\lambda_{LU}}{2}(1 + \frac{1}{r}) \tag{11}$$



(a) Modification of r (T_c fixed)

(b) Modification of T_c (r fixed)

Figure 3: Sensitivity analysis for RMCA model.

and

$$\lambda_{LU_{off}} = \frac{\lambda_{LU}(\tau_{on} + \tau_{off})}{2\tau_{on}} = \frac{\lambda_{LU}}{2}(1 + r)\,. \tag{12}$$

From (11) and (12), it can be deduced that, if $r = 1$ (i.e., $\tau_{on} = \tau_{off}$), $\lambda_{LU_{on}} = \lambda_{LU_{off}} = \lambda_{LU}$ (i.e., new arrivals are uncorrelated) and RMCA is reduced to RM. Recall that, unlike the Erlang loss model, RM still considers retrials. In contrast, if $r \to 0$ (i.e., $\tau_{on} \to 0$), $\lambda_{LU_{on}} \to \infty$, so that new arrivals are highly correlated. From (11) and (12), it can also be deduced that r must not be greater than 1 to ensure that $\lambda_{LU_{on}} \geq \lambda_{LU_{off}}$.

Figure 4 shows congestion and blocking ratios (solid and dotted lines, respectively) with increasing traffic demand for different values of r for the particular case $N = 3$. For comparison purposes, the Erlang Loss model is also included in the figure (denoted as Erl-B). In the latter, $BR = CR$. As expected, the Erlang Loss Model, considering neither retrials nor correlated arrivals, has the lowest BR and CR. In RMCA, the lower the value of r (i.e., the higher the concentration of traffic demand), the larger BR. The same trend is observed for CR for small offered traffic, whereas the opposite is observed for large offered traffic. This result can be explained by the fact that time correlation between new arrivals makes congestion possible, even for very small average offered traffic. For very large average offered traffic, time correlation between arrivals ensure long periods of low activity, which leads to a low CR. More formally, (11) shows that, if $r \to 0$, $\lambda_{LU_{on}} \to \infty$, and, obviously, $\tau_{off} \to T_c$, since $\tau_{on} \to 0$.

The second experiment, described in Figure 3 (b), evaluates the influence of the switching rate between the *on* and *off* periods. For this purpose, the ratio $r = \tau_{on}/\tau_{off}$ is fixed, while still satisfying (10), and the duration of an *on-off* cycle, T_c, is progressively reduced.

Figure 5 shows congestion and blocking ratios with increasing traffic demand for different values of T_c (in seconds) for $r = 0.1$ and $N = 3$. In the figure, it is observed that the larger T_c (i.e., the lower the switching rate), the higher BR. This is due to the fact that, for large switching rates, the effects of the transient regime between the *on* and *off* states become more evident. Thus, the *on* period might not be long enough to cause congestion after the queue becomes empty during the *off* period. It can also be observed that, as T_c decreases, RMCA tends to perform as RM ($r = 1$ in Figure 4), despite the fact that $r << 1$ (i.e., $\lambda_{LU_{on}} >> \lambda_{LU_{off}}$).

3.2.3 Model Tuning

Unlike RM, not all parameters in RMCA can be obtained directly from measurements in the NMS. Note that the values of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} are not available, since only the average LU

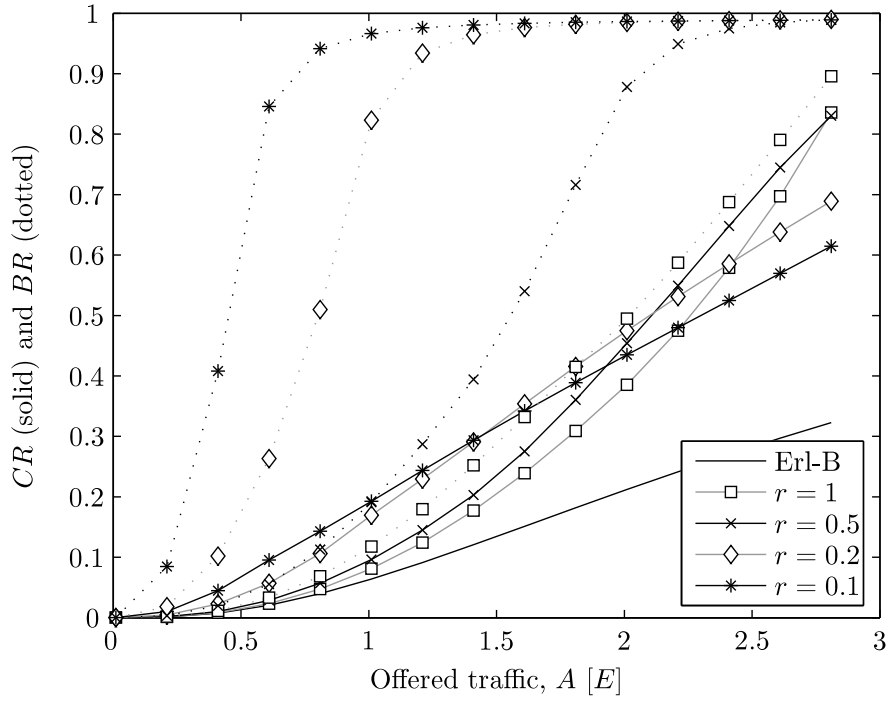


Figure 4: Influence of time correlation between new arrivals on SDCCH performance (case $N = 3$).

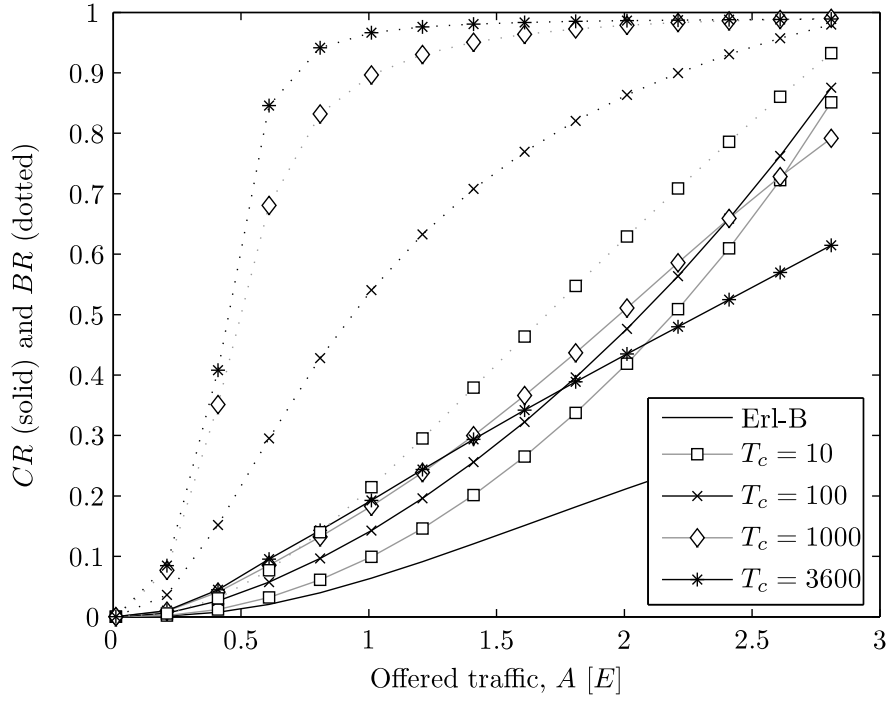


Figure 5: Influence of switching rate between *on* and *off* periods on SDCCH performance (case $N = 3$).

arrival rate, λ_{LU}, is measured. Neither it is possible to define default values for these parameters as experience shows that they greatly vary from cell to cell. Hence, these parameters must be estimated from SDCCH performance measurements. Such a problem is referred to as *inverse problem* in queueing theory.

In (6), it can be observed that several combinations of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} might give the same λ_{LU}, but completely different CR and BR, as shown in Section 3.2.2. For instance, if $\lambda_{LU_{on}} >> \lambda_{LU_{off}}$ and $\tau_{on} << \tau_{off}$ (i.e., most traffic demand is concentrated in a short period of time), CR is moderately low and BR is high. Conversely, a value of BR much larger than the value of CR is an indication of time correlation between new arrivals. Based on this observation, the values of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} in RMCA can be tuned on a cell and hourly basis so that key performance indicators given by RMCA, namely average traffic load, blocking ratio and congestion ratio, resembles as much as possible those measured in the real network. Such a tuning process can be performed by considering the values of $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} as the components of a 4-vector (i.e., a point in $\Re^4$) and solving the nonlinear programming problem, [38],

$$\begin{aligned}
\text{Minimise} \quad & \left(\frac{A_c - A_{c_{rmca}}(\lambda_{LU_{on}}, \lambda_{LU_{off}}, \tau_{on}, \tau_{off})}{N}\right)^2 \\
& + \left(CR - CR_{rmca}(\lambda_{LU_{on}}, \lambda_{LU_{off}}, \tau_{on}, \tau_{off})\right)^2 \\
& + \left(BR - BR_{rmca}(\lambda_{LU_{on}}, \lambda_{LU_{off}}, \tau_{on}, \tau_{off})\right)^2 && (13) \\
\text{subject to} \quad & \frac{\lambda_{LU_{on}}\tau_{on} + \lambda_{LU_{off}}\tau_{off}}{\tau_{on} + \tau_{off}} = \lambda_{LU}\,, && (14) \\
& \lambda_{LU_{on}} \geq \lambda_{LU_{off}}\,, && (15) \\
& \tau_{on} + \tau_{off} \leq 3600\,, && (16) \\
& \lambda_{LU_{on}}, \lambda_{LU_{off}} \geq 0,\ \ \tau_{on}, \tau_{off} \geq 1\,, && (17)
\end{aligned}$$

where $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} are the decision variables, A_c, CR and BR are measurements in the Network Management System (i.e., constants), and $A_{c_{rmca}}$, CR_{rmca} and BR_{rmca} are performance estimates given by RMCA (i.e., functions of the decision variables). Briefly, the objective function (13) reflects the goal of minimising the sum of squared errors between measurements and RMCA estimations for the average load, congestion ratio and blocking ratio. Note that estimates do not only depend on the LU traffic characteristics (defined by the decision variables), but also depend on the attempt rates due to other causes (e.g., MOC, MTC, ...). Since the latter are measurements (i.e., constants), they have been omitted in (13) for clarity. The nonlinear equality constraint (14) enforces the relationship between decision variables so that the average LU arrival rate coincides with the measured value. The linear inequality constraint (15) eliminates the symmetry in the model by forcing that the *on* period is that with the largest LU traffic. The remaining inequality constraints (16) and (17) ensure that the values of the decision variables correspond to a realistic case, where correlated LU arrivals are due to group boundary-crossing events. In practice, a typical crossing event lasts a few seconds, whereas the period between crossing events may be of up to several minutes. Constraint (16) discards excessively low switching rates, whose period would be larger than the measurement window in the NMS (i.e., one hour), while (17) are the lower-bound constraints ensuring that attempt rates are non-negative and the *on* and *off* periods last more than 1 second.

3.2.4 Feasibility of Optimisation Problem

It is interesting to prove that there is always a feasible solution to the problem (13)–(17). Constraint (14) is the only nonlinear equation, while (15)–(17) are linear inequalities. $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} are the unknowns in the optimisation process, while λ_{LU} is the only input parameter in the constraints. Note that λ_{LU} is the measured LU arrival rate and, therefore, λ_{LU} is a non-negative

real value fixed in advance. Constraints (16) and (17-right) are satisfied at points where $\tau_{on} = 1$ and $\tau_{off} = 3599$. If $\lambda_{LU_{on}} = \lambda_{LU_{off}} = \lambda_{LU}$ is also set (i.e., uncorrelated new arrivals), then all the remaining restrictions are satisfied and hence there exists at least one feasible solution, regardless of the value of $\lambda_{LU} (\geq 0)$. In fact, this solution is used in the next section as the starting point for the iterative optimisation algorithm.

4 Model Assessment

The aim of the following analysis is three-fold: a) to show the limitations of the Erlang B formula for dedicated signalling channels in GERAN, b) to prove that such limitations are due to time correlation between attempts, and c) to prove that both retrials and correlated arrivals must be considered to estimate performance on these channels accurately. For this purpose, SDCCH measurements were collected in a large geographical area of a live GERAN system. Such measurements comprise both SDCCH traffic demand and queueing performance on a cell basis and hourly intervals. From traffic demand values, key performance indicators were estimated on a cell and hourly basis by the theoretical models with and without retrials and correlated arrivals. Finally, model assessment was carried out by comparing performance estimates and real measurements throughout the network. For clarity, the analysis set-up is first introduced and results are then presented.

4.1 Analysis Set-up

The analysis is based on data collected over 8 days from 1730 cells in a live GERAN system. Such data is stored in a NMS covering about half of the operator's network. Each sample corresponds to the SDCCH Busy Hour (SDCCH-BH) of a day in a cell. Thus, the original dataset consists of 13840 samples of SDCCH performance data, namely total number of attempts, number of attempts per cause, carried traffic, congestion ratio and blocking ratio. To obtain reliable estimations, samples with SDCCH traffic less than 0.1 Erlang and ratio of ghost attempts larger than 50% are discarded, resulting in 10241 valid samples. The analysis is focused on cells with $N = 3$, 7 and 15, as these are the most common SDCCH configurations in the network. This dataset is representative of the whole network area as it comprises 75% of cells and these samples comprise 90% of the total SDCCH traffic. Likewise, robust estimations are expected, since the dataset covers a large geographical area (i.e., 120000 km^2) with very different traffic and user mobility characteristics.

Preliminary analysis shows that 19% of the 1730 cells experience unacceptable averages of SDCCH-BH blocking (i.e., $BR > 1\%$). At the same time, 10% of the cells have SDCCH TSLs that remain unused even during the SDCCH BH. This a clear indication that the current dimensioning approach used by operators is not working properly.

In the analysis, three different queueing models are compared: the Erlang Loss Model (denoted as Erl-B), the basic retrial model (RM) and the retrial model with correlated arrivals (RMCA). In both retrial models, the retrial rate, α, is set to $1/6\ s^{-1}$, as configured by the operator. A heuristic algorithm is used to fix M on a per-sample basis (i.e., cell and hour) so as to reduce the number of states as much as possible while ensuring that the probability that the orbit is full is negligible. Thus, the value of M ranges from 1 to 100. It is worth noting that while RMCA is tuned for each cell and hour with real data, the same Erl-B and RM are applied to all cells in the network. Thus, $\lambda_{LU_{on}}$, $\lambda_{LU_{off}}$, τ_{on} and τ_{off} in RMCA are calculated for each cell and hour by solving the optimisation problem (13)–(17) with real network statistics, namely A_c, BR, CR and λ_{LU}.

Performance estimates for Erl-B are computed by the Erlang B formula. In the absence of an equivalent expression for RM and RMCA, performance is estimated by numerical methods. For each sample and retrial model, a new matrix $\mathbf{Q}$ is generated and the stationary distribution is computed by solving (2) by the Gaver, Jacobs and Latouche's algorithm. Then, key performance indicators are calculated as in (3)–(5) or (7)–(9). In the case of RMCA, the model is tuned by solving (13)–(17) with the *fmincon* function in MATLAB Optimization Toolbox, [39], initialised

to $\lambda_{LU_{on}} = \lambda_{LU_{off}} = \lambda_{LU}$, $\tau_{on} = 1$ and $\tau_{off} = 3599$. During the tuning process, (2) must be solved several times for each cell and hour, as $\mathbf{Q}$ in RMCA changes with different parameter settings.

Model assessment is based on two figures of merits. From the operator side, the most important criterion is the error in determining the number of fresh blocked attempts for revenue-generating services (i.e., MTC, MOC, EC, SS and SMS). Therefore, an adequate goodness-of-fit measure is the normalised sum of absolute errors for the blocked arrival rate of revenue-generating services, $NSAE_{brgs,m}$,

$$NSAE_{brgs,m} = \frac{\sum_{i=1}^{N_s} \lambda_{rgs}(i) \; |CR(i) - CR_m(i)|}{\sum_{i=1}^{N_s} \lambda_{rgs}(i) \, CR(i)} \,, \tag{18}$$

where N_s is the number of samples (i.e., cells and hours), $\lambda_{rgs}(i)$ is the total fresh arrival rate of revenue-generating services in sample i, $CR(i)$ is the measured congestion ratio in sample i, and $CR_m(i)$ is the congestion ratio for sample i suggested by model m, where $m \in \{\text{Erl-B}, \text{RM}, \text{RMCA}\}$. Note that fresh attempts of these services are Poisson arrivals and, therefore, CR equals the probability of finding all sub-channels busy. This figure of merit is dominated by cells and hours with a larger revenue-generating traffic. From the academic side, all cells, services and performance indicators are equally important. On this premise, a more adequate goodness-of-fit measure is the average sum of squared errors for the average load, congestion ratio and blocking ratio, $\overline{SSE}_m$,

$$\overline{SSE}_m = \frac{\sum_{i=1}^{N_s} \left(\left(\frac{A_c(i) - A_{c_m}(i)}{N(i)} \right)^2 + (CR(i) - CR_m(i))^2 + (BR(i) - BR_m(i))^2 \right)}{N_s} \,, \tag{19}$$

where $A_c(i)$, $CR(i)$ and $BR(i)$ are measurements, $A_{c_m}(i)$, $CR_m(i)$ and $BR_m(i)$ are estimates from model m, and $N(i)$ is the number of SDCCH sub-channels in the cell of sample i.

4.2 Analysis Results

The first experiment checks if the arrival process is Poisson distributed by checking the *Poisson Arrivals See Time Averages* (PASTA) property, [40], over real SDCCH measurements. Figure 6 shows a scatter plot of CR versus BR, together with a dashed line representing $BR = CR$. It is observed that, in many cases, BR is significantly larger than CR. This is a clear indication that the Poisson assumption does not hold for the SDCCH.

Figure 7 confirms this statement by representing CR and BR measurements in terms of the average carried traffic for cells with 3, 7 and 15 sub-channels. For comparison purposes, congestion values given by the Erlang B and C formulas (i.e., blocking and delay probability in a loss and delay system, respectively) are superimposed. The analysis is first focused on cells with N=3, represented in Figure 7 (a). It is observed that, in most cases, both CR and BR are above the Erlang B curve. Thus, for large traffic values, BR can be up to twice the value predicted by the Erlang B formula. Similar results are observed in cells with N=7 and 15, represented in Figure 7 (b)-(c). A more detailed analysis (not shown here) reveals that the value of the blocking probability given by the Erlang B formula falls outside the 95% confidence interval of BR in 15% of the samples. More important, the problematic samples are those with a larger blocking ratio, comprising 26% of the total SDCCH traffic in the network. Therefore, these samples are the main focus of network re-planning procedures.

These estimation errors are partly due to retrials. On the one hand, retrials make the network behave, in some sense, like a delay system. Thus, retrials tend to enlarge congestion periods (and, hence, increase CR), as channels are occupied by users in the orbit as soon as they become free. As a result, CR tends to be larger than Erlang B blocking probability for the same value of A_c. This is confirmed by the fact that most CR samples in Figure 7 (a) lie between Erlang B and C curves, which is a well-known property of retrial queues, [19]. On the other hand, due to retrials,

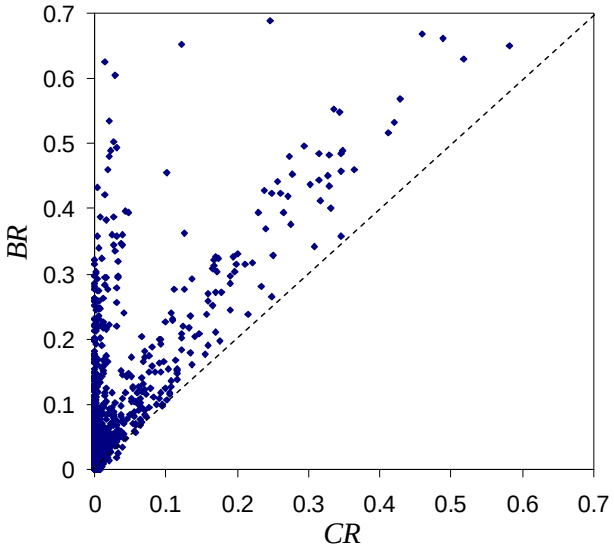
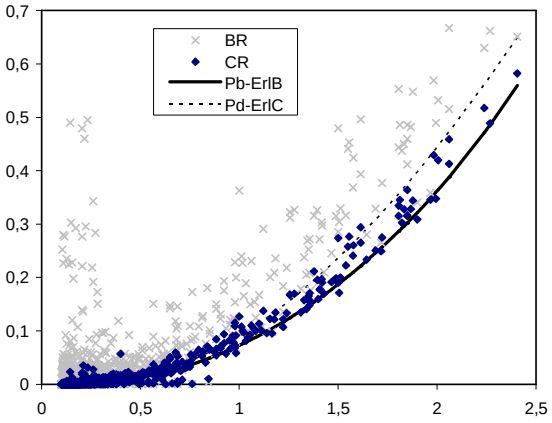


Figure 6: SDCCH congestion ratio versus blocking ratio in a live network.

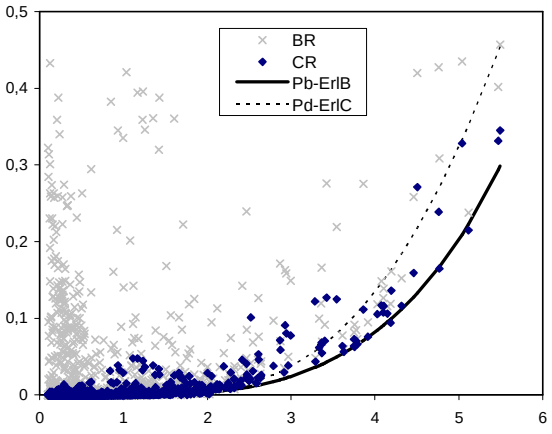
attempts are not statistically independent, but are concentrated around congestion periods. This justifies that $BR >> CR$, which is also a well-known effect of retrials, [16]. Similar trends are observed in cells with N=7 and 15, represented in Figure 7 (b)-(c). However, while $BR > CR$ in both figures, CR is well above the Erlang C curve in many samples in Figure 7 (c). From this observation, it is envisaged that RM will fail to explain congestion in cells with a large number of sub-channels.

The previous hypotheses are confirmed by the overall estimation results. Table 1 presents the values of the two goodness-of-fit measures for the three queueing models. Results have been broken down by number of sub-channels. In the table, it is observed that Erl-B is the worst model, as it gives the largest value of $NSAE_{brgs}$ and $\overline{SSE}$ for any number of channels. From the former indicator, it can be inferred that the error in estimating the number of blocked fresh attempts of revenue generating services by the Erlang B formula is 23%, 42% and 82% for N = 3, 7 and 15, respectively. As already pointed out, the error is much larger in cells with large N, where the Erlang B formula fails to explain congestion. In contrast, $\overline{SSE}$ for Erl-B decreases with N due to the fact that there are fewer cells with congestion problems when N is large. Similarly, RM shows large values of $NSAE_{brgs}$ and $\overline{SSE}$. Thus, it can be concluded that retrials can only explain a small fraction of blocking. In contrast, RMCA gives the lowest values of $NSAE_{brgs}$ and $\overline{SSE}$ for any number of sub-channels. From the last column, it can be deduced that, with RMCA, the overall $NSAE_{brgs}$ and $\overline{SSE}$ are reduced by 63% and 77%, respectively, when compared to Erl-B. This is a clear indication of the superior accuracy of RMCA. For $NSAE_{brgs}$, the benefit is more evident for large N, since this figure is dominated by a few cells where only RMCA can explain blocking. In contrast, the benefit in $\overline{SSE}$ is more evident for small N, as this indicator is an average of many cells and there are more cells with congestion problems in relative terms for small N.

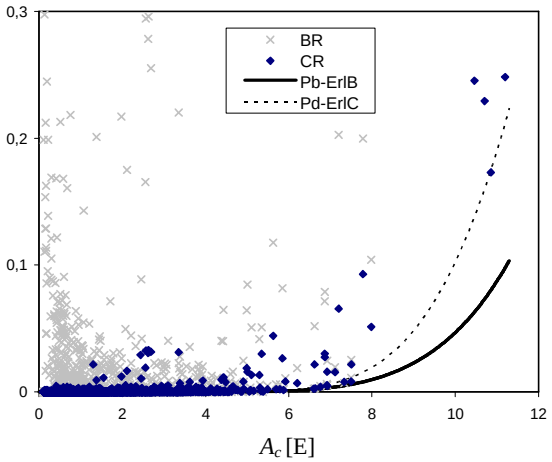
The previous result was expected, since parameters in RMCA are adjusted on a cell/hour basis to fit performance data. More interesting is the analysis of the final RMCA settings, which can reveal the amount of cells with correlated SDCCH arrivals. Such an analysis shows that 8% of the samples display values of $(\lambda_{r_{on}} + \lambda_{nr})/(\lambda_{r_{off}} + \lambda_{nr}) > 2$. Although this figure might seem relatively small, note that this is more than half of the samples where the Erlang B formula fails (i.e., 15%). More important, these samples comprise 59% of the total blocked attempts. These results clearly indicate the need for considering time correlation between new arrivals when estimating SDCCH performance.



(a) $N = 3$



(b) $N = 7$



(c) $N = 15$

Figure 7: SDCCH congestion and blocking performance.

Measure	Model	N=3	N=7	N=15	N={3,7,15}
$NSAE_{brgs}$	Erl-B	0.2272	0.4198	0.8242	0.4904
	RM	0.2176	0.4065	0.6500	0.4247
	RMCA	0.1519	0.1742	0.2244	0.1835
$\overline{SSE}$ $[\cdot 10^{-2}]$	Erl-B	0.4745	0.1812	0.0733	0.2430
	RM	0.3700	0.1742	0.0696	0.2046
	RMCA	0.0383	0.0863	0.0452	0.0566

Table 1: Performance of queueing models for different number of sub-channels

4.3 Implications for SDCCH Re-dimensioning

Having seen how poorly Erl-B/C predict SDCCH performance, it is clear that a precise performance model is needed to re-dimension SDCCH resources on a cell basis based on network statistics. The main challenge is to identify cells where the number of SDCCH sub-channels is unnecessarily high, as the opposite situation (i.e., excessively low) can simply be detected from BR statistics. Results have shown that the offered SDCCH traffic cannot be estimated by the sum of carried and blocked attempts, since part of the latter are retrials from the same original attempts. Likewise, the Erlang B formula currently in use is not adequate for many cells. To re-dimension SDCCH resources, the operator should first check if $CR \simeq BR$ and then check if both indicators coincide with the blocking probability obtained by the Erlang B formula, P_b. In cells where $CR < P_b < BR$ or $P_b < CR < BR$, RMCA can be adjusted from network statistics to derive the actual offered traffic and its temporal distribution. Once tuned, RMCA can be used to predict blocking performance with a different number of sub-channels. In the rare cases where $CR < BR < P_b$, none of the previous models is valid. The latter situation is typical of cells with limited population and large SMS traffic due to WAP-over-SMS traffic, [41]. For these cells, a finite source queueing model is more adequate.

The main drawback of the proposed methodology when compared to the current approach is the increased computational load. Specifically, the execution time for the 10241 samples (i.e., all the data in a NMS) in a 2.4GHz 2GB-RAM Windows-based computer is 2 hours, most of which is spent in samples with a large value of N and a high level of congestion, where M has to be set to 100 to get accurate results. Note that tuning RMCA requires running the Gaver-Jacobs-Latouche's algorithm several times per sample. The execution time might be reduced by substituting the current finite truncation approach by generalised truncated methods, [22]. Nonetheless, the current execution time is low enough for network re-planning purposes.

5 Conclusions

Due to financial pressures, operators are increasingly forced to maximise the financial return on their investment in GERAN. To achieve this aim, operators are trying to ensure that every time slot is assigned to the most suitable usage, i.e., SDCCH or TCH. To assist operators in this undertaking, a comprehensive performance analysis of dedicated signalling channels has been performed in a live GERAN system. Preliminary analysis has shown that the Erlang B formula, currently used by operators, fails to give adequate estimates of the SDCCH blocking ratio in 15% of measurements. More important, the problematic samples correspond to cells with the largest SDCCH blocking, receiving most of the attention from the operator. Likewise, it has been shown that considering retrials does not reduce estimation errors significantly. To overcome these limitations, a retrial queueing model with correlated arrivals has been proposed. Time correlation between new arrivals is modelled by a switched Poisson process. The resulting model is simple and flexible enough to be adjusted on a per-cell and per-hour basis using statistics in the Network Management System. With this model, the sum of squared residuals for the main performance indicators is reduced by

77% when compared to the current approach.

It is clear that more complex models considering differences between retrials and redials, [4][15][25], or more general distributions of inter-arrival, service and inter-retrial times, [15][23], would obtain more accurate predictions. However, such models are more difficult to tune, as they require knowledge of the traffic attributes on a cell basis. Such knowledge can only be obtained by a time-consuming analysis of traffic traces, which are seldom available.

An important issue is how to extend the analysis to other radio access technologies. The queueing models considered in this work have been conceived for dedicated signalling channels. While these channels in GERAN are based on a TDMA/FDMA scheme, their counterparts in *Universal Mobile Telecommunication System* (UMTS) and *Long Term Evolution* (LTE) (i.e., Dedicated Control CHannel, DCCH) use CDMA and OFDMA/TDMA schemes, respectively, [42][43]. In the literature, several attempts have been made to extend queueing models for TDMA/FDMA to CDMA and OFDMA for user traffic channels. For CDMA systems, the proposed model can be upgraded with state-dependent blocking probabilities to reflect that cell capacity depends on neighbour cell interference dynamically, [44][45]. A similar approach can be used in OFDMA-TDMA systems, where adaptive modulation and coding cause that the bandwidth allocated to each user is not deterministic, but dependent on channel conditions, [46]. It can be argued that the distinction between signalling and user traffic resources in UMTS and LTE is not as clear as in GERAN. Nonetheless, it is expected that a minimum share of cell capacity is reserved for signalling purposes by the operator. More important, this work has proved that new LU requests are time correlated in many cells of a live GERAN system. Such a behaviour is expected to be same in UMTS and LTE, since: a) idle user mobility does not depend on the radio access technology, b) UMTS and LTE networks are also divided into location, routing and tracking areas, and c) location-management procedures in UMTS and LTE are similar to those in GERAN, [47]. With the steady decrease in cell size, it is expected that signalling due to mobility management is a major traffic component in future mobile communication networks. The proposed methodology can help UMTS and LTE operators to re-dimension signalling resources in cells by detecting correlation between arrivals.

6 Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation (grant TEC2008-06216). The authors would also like to thank Dr. Pablo Guerrero García for his valuable comments on the feasibility of the optimisation problem.

References

[1] M. Mouly, M. B. Pautet, The GSM system for mobile communications, Cell & Sys, 1992.

[2] T. Halonen, J. Melero, J. Romero, GSM, GPRS and EDGE Performance: Evolution Toward 3G/UMTS, John Wiley & Sons, 2002.

[3] G. Tunnicliffe, A. Murch, A. Sathyendran, P. Smith, Analysis of traffic distribution in cellular networks, in: Proc. 48th IEEE Vehicular Technology Conference, Vol. 3, 1998, pp. 1984–1988.

[4] E. Onur, H. Delic, C. Ersoy, M. U. Caglayan, Measurement-based replanning of cell capacities in GSM networks, Computer Networks 39 (6) (2002) 749–767.

[5] D. Hong, S. S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized and nonprioritized handoff procedures, IEEE Transactions on Vehicular Technology 35 (3) (1986) 77–92.

[6] Y. Fang, I. Chlamtac, Teletraffic analysis and mobility modeling of PCS networks, IEEE Transactions on Communications 47 (7) (1999) 1062–1071.

[7] S. Rappaport, The multiple-call hand-off problem in high-capacity cellular communications systems, in: Proc. 40th IEEE Vehicular Technology Conference, 1990, pp. 287–294.

[8] W. Li, A. S. Alfa, A PCS network with correlated arrival process and splitted-rating channels, IEEE Journal on Selected Areas in Communications 17 (7) (1999) 1318–1325.

[9] P. Tran-Gia, M. Mandjes, Modeling of customer retrial phenomenon in cellular mobile networks, IEEE Journal on Selected Areas in Communications 15 (8) (1997) 1406–1414.

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

[11] S. Rappaport, Traffic performance of cellular communication systems with heterogeneous call and platform types, in: Conf. Record, 2nd Int. Conf. on Universal Personal Communications, Vol. 2, 1993, pp. 690–695.

[12] Y. Fang, Thinning schemes for call admission control in wireless networks, IEEE Transactions on Computers 52 (5) (2003) 685–687.

[13] X. Lagrange, P. Godlewski, Teletraffic analysis of a hierarchical cellular network, in: Proc. 45th IEEE Vehicular Technology Conference, Vol. 2, 1995, pp. 882–886.

[14] P. Fitzpatrick, C. S. Lee, B. Warfield, Teletraffic performance of mobile radio networks with hierarchical cells and overflow, IEEE Journal on Selected Areas in Communications 15 (8) (1997) 1549–1557.

[15] A. S. Alfa, W. Li, PCS networks with correlated arrival process and retrial phenomenon, IEEE Transactions on Wireless Communications 1 (4) (2002) 630–637.

[16] J. R. Artalejo, A. Gómez-Corral, Retrial Queueing Systems, Springer, 2008.

[17] N. W. Macfadyen, Statistical observation of repeated attempts in the arrival process, in: Proc. 9th International Teletraffic Congress (ITC), 1979.

[18] K. Liu, Direct distance dialing: Call completion and customer retrial behavior, Bell System Technical Journal 59 (1980) 295–311.

[19] M. Nesenbergs, A hybrid of Erlang B and C formulas and its applications, IEEE Transactions on Communications 27 (1979) 59–68.

[20] M. Neuts, B. M. Rao, Numerical investigation of a multiserver retrial model, Queueing Systems 7 (2) (1990) 169–189.

[21] J. Artalejo, M. Pozo, Numerical calculation of the stationary distribution of the main multiserver retrial queue, Annals of Operations Research 116 (2002) 41–56.

[22] M. Domenech-Benlloch, J. Giménez-Guzmán, V. Pla, J. Martínez-Bauset, V. Casares-Giner, Generalized truncated methods for an efficient solution of retrial systems, Mathematical Problems in Engineering 2008.

[23] S. R. Chakravarthy, A. Krishnamoorthy, V. C. Joshua, Analysis of a multi-server retrial queue with search of customers from the orbit, Performance Evaluation 63 (8) (2006) 776–798.

[24] M. Domenech-Benlloch, J. Giménez-Guzmán, J. Martínez-Bauset, V. Casares-Giner, Efficient and accurate methodology for solving multiserver retrial systems, IEE Electronics Letters 41 (17) (2005) 967–969.

[25] J. Giménez-Guzmán, M. Domenech-Benlloch, V. Pla, V. Casares-Giner, J. Martínez-Bauset, Analysis of cellular network with user redials and automatic handover retrials, Lecture Notes in Computer Science, Next Generation Teletraffic and Wired/Wireless Advanced Networking 4712/2007 (2007) 210–222.

[26] U. Gotzner, A. Gamst, R. Rathgeber, Spatial traffic distribution in cellular networks, in: Proc. 48th IEEE Vehicular Technology Conference, Vol. 2, 1998, pp. 1994–1998.

[27] S. Almeida, J. Queijo, L. Correia, Spatial and temporal traffic distribution models for GSM, in: Proc. 50th IEEE Vehicular Technology Conference, Vol. 1, 1999, pp. 131–135.

[28] 3GPP TS 05.02 (v6.6.0), Multiplexing and multiple access on the radio path; GSM-Phase2+, Release 97 (Nov 1999).

[29] 3GPP TS 04.08 (v7.20.1), Mobile radio interface layer 3 specification; GSM-Phase2+, Release 98 (Sep 2003).

[30] S. Pedraza, V. Wille, M. Toril, R. Ferrer, J. Escobar, Dimensioning of signaling capacity on a cell basis in GSM/GPRS, in: Proc. 54th IEEE Vehicular Technology Conference, Vol. 1, 2003, pp. 155–159.

[31] R. Wilkinson, Theories for toll traffic engineering in the U.S.A., Bell System Technical Journal 35 (2) (1956) 421–514.

[32] W. J. Stewart, Introduction to the numerical solution of Markov chains, Princeton University Press, 1994.

[33] K. Meier-Hellstern, W. Fischer, The Markov-Modulated Poisson Process (MMPP) cookbook, Performance Evaluation 18 (1992) 149–171.

[34] D. Gaver, P. Jacobs, G. Latouche, Finite birth-and-death models in randomly changing environments, Advances in Applied Probability 16 (4) (1984) 715–731.

[35] W. Grassmann, M. Taksar, D. Heyman, Regenerative analysis and steady state distributions for markov chains, Operations Research 33 (1985) 1107–1116.

[36] L. D. Servi, Algorithmic solutions to two-dimensional birth-death processes with application to capacity planning, Telecommunications Systems 21 (2-4) (2002) 205–212.

[37] R. L. Burden, D. J. Faires, Numerical Analysis, Brooks Cole, 2004.

[38] J. Nocedal, S. J. Wright, Numerical Optimization, 2nd Edition, Springer, 2006.

[39] The MathWorks, Optimization Toolbox 4, User's Guide (2008).

[40] D. Gross, C. M. Harris, Fundamentals of Queueing Theory, 3rd Edition, Wiley, 1998.

[41] A. Andreadis, G. Benelli, G. Giambene, B. Marzucchi, Analysis of the WAP protocol over SMS in GSM networks, Wireless Communications and Mobile Computing 1 (2001) 381–395.

[42] H. Holma, A. Toskala, WCDMA for UMTS: Radio Access for Third Generation Mobile Communications, 3rd Edition, John Wiley & Sons, 2004.

[43] H. Holma, A. Toskala, LTE for UMTS. OFDMA and SC-FDMA based radio access, John Wiley & Sons, 2009.

[44] D. Staehle, A. Mäder, An analytic approximation of the uplink capacity in a UMTS network with heterogeneous traffic, in: Proc. 18th International Teletraffic Congress (ITC), 2003, pp. 81–91.

[45] V. Iversen, V. Benetis, N. Ha, V. Ha, S. Stepanov, Evaluation of multi-service CDMA networks with soft blocking, in: Proc. 16th International Teletraffic Congress (ITC), 2004, pp. 212–216.

[46] H. Wang, V. Iversen, Erlang capacity of multi-class TDMA systems with adaptive modulation and coding, in: Proc. IEEE International Conference on Communications (ICC), 2008, pp. 115–119.

[47] 3GPP TS 23.012 (v8.2.0), Technical specification group core network; Location management procedures, Release 8 (Jun 2009).