The Impact of Customer Retrial on the Capacity Redimensioning Process of Cellular Mobile Networks

Lucian Ioan, Andreea Dumitru, Grazziela Niculescu*

Abstract – The paper propose a method for capacity redimensioning process of cellular mobile networks which take into account the customer retrial phenomenon. The main parameter of this method is the real call blocking probability which can be estimate by analyzing a two dimension Markov chain. Its transition probabilities are governed by two important factors related to the customer behavior: the retrial probability and the mean retrial rate. The influence of these factors on the capacity redimensioning process is analyzed by means of some numerical results.

Keywords – call blocking probability, performance, retrial, Markov chain, QoS, capacity assignment.

I. INTRODUCTION

For planning the growth of telecommunications networks it is important to have a good knowledge about the traffic offered, for instance, to circuit groups or signaling links [1]. This quantity are normally estimated from measurements of busy-hour carried traffic and call attempts, [2], but there are a number of factors which may need to be taken into account within the estimation procedures. One of these factors is the congestion. This will generally result in a decrease in carried traffic, influenced by customer’s repeat attempts and by the actions (e.g., automatic repeat attempts) of other network components [3].

The modeling of customer’s repeat attempts has been a subject of a large number of papers, [4] – [8]. Most of them resort of queuing theory concepts and use continuous Markov chain to model the impact of customer behavior on the system performance.

In this paper we consider the two dimensional Markov chain proposed by [7], which are briefly described in section II. Then, in the third section we derived the formulas for system performance evaluation, including the offered traffic. An application of the derived formulas in network planning activity is described and numerical results are analyzed in section 4, and, finally, in section 5, appropriate conclusions are done.

II. MODEL DESCRIPTION

The abstract structure of the traffic system and the possible ways to access its pool of resources by the customers are presented in Fig. 1. We consider a number of \( n \) channels and a finite number of \( m \) customers, greater than \( n \). According to the communication system an attached mobile station can be found in one of three states: IDLE (non calling/called customer), BUSY (active customer, in a calling or called situation, his call occupying one communication channel) or WAITING for attempts. When the customer is "active" he uses the allocated channel whose holding time has an exponentially distribution with mean \( 1/\mu \). After finishing a call, a customer will enter to the Idle customers' source state and he will stay there until generating a new call. The interarrival times of the first attempt calls are supposed to be exponentially distributed with mean \( 1/\alpha \).

The model assumes, that every time a customer is blocked, he will try to repeat the request with probability \( \theta \) (the persistence intensity) and will give up with probability \( 1 - \theta \). If the customer decides to reattempt, he will wait an exponentially distributed time with mean \( 1/\alpha_0 \) before repeating his request. The redial probability \( \theta \) is not affected by the number of redial already done. Considering the above assumption, the traffic processing system can be modeled as a continuous-time Markov chain [10]. The state transition diagram is depicted in Fig. 2. The current state of the system is described by means of two values. The first one, \( i \), represents the number of active customers (the number of occupied channels). This number varies between 0 and the...
number of channels available, $n$. The second value, $j$, represents the number of customers waiting for reattempt and it takes values between 0 and $m-n$, $m$ representing the dimension of the customer source.

The steady state probabilities, that is the chance to have $i$ channels in use and $j$ customers waiting for reattempt, are denoted as:

$$p(i, j) = P[X = i, Y = j], 0 \leq i \leq n, 0 \leq j \leq m - n \quad (1)$$

These probabilities are found by solving the global balance equations which are written around each state of the system. These equations are of the next forms:

- around state $(i, j)$:
  $$[\mu + (m - i - j)\alpha + \alpha_0]p(i, j) = (m - i - j)\alpha p(i - 1, j) + (j + 1)\mu p(i, j + 1)$$
  $$\quad (2)$$

- around state $(n, 0)$:
  $$(n\mu + (n - m)\alpha\alpha_0)p(n, 0) = (m - n + 1)\alpha p(n - 1, 0) + \alpha_0 p(n, 1) \quad (3)$$

- around state $(n, n - m)$:
  $$[\mu + (1 - 0)(m - n)\alpha_0]p(n, m - n) = \alpha p(n - 1, m - n) + \alpha_0 p(n, m - n - 1)$$
  $$\quad (4)$$

For the computation of the state probabilities one of these equations must be replaced with the normalization equation, because the original equations system is linear dependent:

$$\sum_{i=0}^{n} \sum_{j=0}^{m-n} p(i, j) = 1 \quad (5)$$

The calculus may be performed by using a matrix approach or by applying the recursive algorithm proposed in [7] in order to overcome problems when inverting large matrix.

III. SYSTEM PERFORMANCE CHARACTERISTICS

The performance of the system is given by the following characteristics [9]:

- the **carried (served) traffic**, defined as the mean number of channels in use:
  $$A_s = \sum_{i=0}^{n} \sum_{j=0}^{m-n} p(i, j) \quad (6)$$

- the **apparent offered traffic**, defined as the mean number of request (first and repeated) received by the system service units along a mean service time:
  $$A_{ap\_of} = \rho \sum_{i=0}^{n} \sum_{j=0}^{m-n} [(m - i - j) + \beta] p(i, j) \quad (7)$$

- the **real offered traffic**, derived from the model M/M/m/m, that is when all clients’ request are accepted by the system service units:
  $$A_{rl\_of} = \sum_{j=0}^{m-n} \sum_{i=0}^{n} p(i, j) \rho^{m-j} \quad (8)$$

- the **apparent call blocking (reject) probability** (measured) defined as the long term ratio between the number of rejected request and the number of all requests:
  $$p_{ap\_bl} = 1 - \frac{A_s}{A_{ap\_of}} \quad (9)$$

- the **real call blocking probability** defined as the long term ratio between the number of first attempt rejected request and the number of first attempt request:
  $$p_{rl\_bl} = 1 - \frac{A_s}{A_{rl\_of}} \quad (10)$$

All the above computational formulas for traffic estimation are expressed as functions of the next two parameters:

- the **normalized individual rate of first attempt service request**, denoted and defined as:
  $$\rho = \alpha / \mu \quad (11)$$

- the **retrial intensity**, denoted and defined as:

![Fig. 2. State transition diagram](Image)
\[ \beta = \rho_0 / \rho \] (12)
where \( \rho_0 \) is the normalized individual rate of repeated service request, defined as:
\[ \rho_0 = \alpha_0 / \mu \] (13)

IV. THE CAPACITY DIMENSIONING PROCESS

The network planning process is depicted in Fig. 3. The parameters used in the diagram have the following signification:
- \( T \) – the time interval between two traffic estimations from measurements;
- \( n_k \) – the number of service units at time \( t_k \);
- \( p_d \) – the call blocking probability threshold considered in the phase of (re)dimensioning;
- \( p_{\text{max}} \) – the maximum accepted value for the call blocking probability;
- \( \rho_k \) - normalized individual rate of first attempt service request that is evaluated at time \( t_k \);
- \( \text{Eng}(n,m,\rho) \) – Engset formula with its recurrent form:
\[ \text{Eng}(i,m,\rho) = \rho \cdot (m - i + 1) \cdot \text{Eng}(i-1,m,\rho) / (1 + (m - i + 1) \cdot \text{Eng}(i-1,m,\rho)) \] (14)

To see how this process works, we consider the normalized individual rate of first attempt service request, \( \rho \), as variable (that is, customer rate demands change in time) and we keep the normalized individual rate of repeated service request, \( \rho_0 \), the service completion rate, \( \mu \), the persistence intensity, \( \theta \), and the dimension of the customers source, \( m \), as constants. In these circumstances we can imagine the next scenario.

At the first step of the dimensioning process, \( k = 0 \), the number of service units is evaluated using Engset formula, imposing that the call blocking probability is below a predefined threshold \( p_d \). Following the predetermined time interval, \( T \), the apparent call blocking probability is computed from equation (9) and compared with the limit value, \( p_{\text{max}} \). If the limit value of the call blocking probability is exceeded, a new dimensioning might be required, starting with a new evaluation of \( \rho \).

One possible estimation way of \( \rho \), is the graphic method. The measured value for the apparent offered traffic and the values computed with equation (7), are plotted as a function of \( \rho \) on the same graph (see Fig. 4). In this case the solution will be the abscissa of the crossing point of these two curves, for instance: \( \rho = 0.12 \), when we considered that the measured apparent offered traffic is \( 10E \), \( m = 100 \), \( n = 15 \), \( \rho_0 = 5 \) and \( \theta = 0.8 \).

With the current value of \( \rho \) we determine the real call blocking probability by using equation (10) and we pass to the last decision block of the dimensioning diagram.

As an example of dimensioning process evolution, we consider at step \( k = 0 \): \( m = 100 \), \( \rho = 0.05 \), \( \rho_0 = 5 \) and \( \theta = 0.8 \). Imposing that the call blocking probability is below \( p_d = 10^{-4} \), we set the number of service units \( n_0 = 15 \). Then, we consider two possible cases:

- After time interval \( T \), the dimensioning process goes to the next step, \( k = k + 1 \), and the apparent call blocking probability is \( p_{\text{ap} \_bl} = 2 \cdot 10^{-3} \). If \( p_{\text{max}} \) is set to \( 10^{-3} \), the condition \( p_{\text{ap} \_bl} > p_{\text{max}} \) is satisfied, and the next action is to evaluate \( \rho \). We obtain \( \rho = 0.07 \) and the real call blocking probability \( p_{rl \_bl} = 4.05 \cdot 10^{-4} \). Because \( p_{rl \_bl} < p_{\text{max}} \), we restart to collect the traffic data for the next performance evaluation.
- After time interval \( T \), the apparent call blocking probability is \( p_{\text{ap} \_bl} = 6.4 \cdot 10^{-3} \). This value satisfied the condition \( p_{\text{ap} \_bl} > p_{\text{max}} \), so we must evaluate \( \rho \). We obtain \( \rho = 0.08 \) and the real call blocking probability \( p_{rl \_bl} = 1.3 \cdot 10^{-3} \). Because \( p_{rl \_bl} > p_{\text{max}} \), we must reevaluate the number of service units. We obtain \( n_1 = 19 \).

The capacity dimensioning process is straightforward related to the customer behavior which can be quantified by two parameters: the persistence intensity, \( \theta \), and the retrial intensity, \( \beta \).

To analyse the impact of the customer behavior we consider the next conditions: \( m = 100 \), \( n = 15 \), \( \rho = 0.05 \), \( 1 / \mu = 120 \text{ sec} \) and we plot in Fig. 5 the main system performance characteristics, that is apparent and real call blocking probabilities versus those two parameters. The
range of $\beta$ parameter ($25 \div 200$) is equivalent to the fact that the mean retrial rate take values between 0.625 calls/min and 10 calls/min.

From Fig. 5 results that the apparent blocking probability takes greater and greater values than the real blocking probability as the two considered parameters increase. In the worst case, $\theta = 1$ and $\beta = 200$, the apparent blocking probability is two time greater than the real one, this deviation inducing significant error in redimensioning process.

V. CONCLUSION

In this paper, we analyzed the impact of the redial behavior on network performance. In overload conditions, the repeated attempts increase the blocking probability. Therefore, in order to guarantee the desired quality of service (QoS), it is very important not to ignore the retrial phenomenon.

To model the retrial phenomenon, two parameters are considered in this paper: the persistence intensity and the retrial intensity. These parameters control the call blocking probability, which can be evaluated from measurements. Unfortunately, the measurements include reattempts and rejected reattempts, which disturb the estimation. Thus, in systems with limited number of resources, as cellular networks we must evaluate the real blocking probability in order to perform a correct capacity redimensioning.

Future work will focus on the evaluation of the real offered traffic and the redimensioning process in a more complex environment (like 3G mobile networks or ATM networks), where customers are streams of data packets, possible with different quality of service (QoS) requirements.

REFERENCES