Efficient Approximations for Call Admission Control Performance Evaluations in Multi-Service Networks

Emre A. Yavuz, and Victor C. M. Leung
Department of Electrical and Computer Engineering
The University of British Columbia
Vancouver, BC, Canada V6T 1Z4
Email: {emrey, vleung} @ece.ubc.ca

Abstract—Several dynamic call admission control (CAC) schemes for cellular networks have been proposed in the literature to reserve resources adaptively to provide the desired quality of service (QoS) to not only high priority calls but also to low priority ones. Efficient adaptive reservations depend on reliable and up-to-date system status feedback provided to the CAC mechanism. However exact analysis of these schemes using multidimensional Markov chain models are intractable in real time due to the need to solve large sets of flow equations. Hence performance metrics such as call blocking probabilities of various QoS classes are generally evaluated using one dimensional Markov chain models assuming that channel occupancy times for all QoS classes have equal mean values and all arriving calls have equal capacity requirements. In this paper we re-evaluate the analytical methods to compute call blocking probabilities of various QoS classes for several widely known CAC schemes by relaxing these assumptions, and propose a novel approximation method for performance evaluation with low computational complexity. Numerical results show that proposed method provides results that match well with the exact solutions.

I. INTRODUCTION

Next generation wireless networks will support not only voice telephony service but also a wide variety of data services for multimedia Internet applications. Satisfying the diverse quality of service (QoS) requirements of these services over cellular networks has become even more challenging due to reduced cell size and hence increased user mobility. Call admission control (CAC) schemes are deployed to selectivity limit the number of admitted calls from each QoS class to maximize the network utilization while satisfying the QoS constraints. CAC for wired and wireless networks has been intensively studied in the past and many priority based CAC schemes have been proposed [1]–[7]. Calls with more stringent QoS requirements are given higher priorities by having exclusive access to a number of reserved channels. Reducing blocking probabilities of calls with higher priorities increases the probability of blocking for calls with relatively lower priorities resulting in a trade off between QoS classes.

A set of guard channels are reserved for prioritized calls in Guard Channel (GC) schemes such as cutoff priority [1], fractional guard channel [2], new call bounding [3] and rigid division based [4] schemes. Many dynamic GC schemes have also been proposed to maximize network utilization adaptively [5]–[7]. Efficient adaptive reservation depends on reliable and up-to-date performance feedback; however exact analyses of these schemes using multidimensional Markov chain models are intractable in real time due to the need to solve large sets of flow equations [8][9]. Hence performance metrics such as call blocking probabilities are generally evaluated using one dimensional Markov chain models under the simplifying assumptions that call arrivals are Poisson, channel occupancy times are exponentially distributed with equal mean values, and traffic classes have the same capacity requirements. These assumptions may not be appropriate since calls with different priorities may have different average channel occupancy times, if not different distributions [10][11]. Existing performance evaluation methods, such as traditional and normalized, lead to significant discrepancies when average channel occupancy times for distinct QoS classes are different [12].

Performance evaluation approximation methods that have high accuracy and low computational cost are needed if dynamic CAC schemes are to be implemented in real time systems. In [13], Gersht and Lee proposed an iterative algorithm by modifying the approximation suggested by Roberts [14] to improve its accuracy when the service rates differ. However we showed in [12] that starting with an inappropriate initial value leads to significant discrepancies and thus proposed a closed form approximation method based on one dimensional Markov chain modeling, which we called effective holding time. We assume that all classes have same capacity requirements and independent and exponentially distributed channel occupancy times without the necessity of having the same average values. Numerical results showed that our approximation method gives more accurate results when compared with the previously proposed approximation methods. In the absence of a product form solution when various classes have distinct capacity requirements, calculating the channel occupancy distribution involves solving the demanding balance equations numerically. In [15], Borst and Mitra developed computational algorithms for the multi-service case by coupling the computation of joint channel occupancy probabilities with that of used capacity assuming that channels are occupied independently. The authors solved the resulting balance equations using numerical iterations.

In this paper, we classify CAC schemes into two novel categories based on the nature of communication links; symmetric and asymmetric. We define a CAC scheme as symmetric if all the communicating nodes in the state transition...
diagram of its Markov chain model have bidirectional links between them, such as in complete sharing (CS), complete partitioning (CP) and new call bounding schemes. We define a CAC scheme as asymmetric when the converse is true, such as in cutoff priority and fractional guard channel schemes. This paper is organized as follows. In the next section, we propose a novel performance evaluation approximation method for asymmetric CAC schemes. In Section III we compare the numerical results obtained from the proposed method with those obtained from a previously proposed approximation and from exact analysis. We conclude the paper in Section IV.

II. PERFORMANCE EVALUATION OF ASYMMETRIC CALL ADMISSION CONTROL SCHEMES

We consider a cellular system with two classes of calls: non-prioritized and prioritized calls, where the latter enjoy a higher service priority than the former. Let $\lambda_{np}$ and $\lambda_p$ denote the arrival rates, $1/\mu_{np}$ and $1/\mu_p$ denote the average channel occupancy times and $b_{np}$ and $b_p$ denote the required bandwidth in units for non-prioritized and prioritized calls, respectively. Let $C$ denote the total number of channels in a cell, and $q_p(j)$ and $q_{np}(r)$ denote the estimated equilibrium channel occupancy probabilities when $j$ prioritized calls and $r$ non-prioritized calls, respectively, exist in the system. Let $\beta_i$ denote the admission probability of an arriving non-prioritized call when the number of busy channels is $i$, and $k_i$ denote the admission probability of an arriving prioritized call when $j$ prioritized calls exist in the cell regardless of the number of existing non-prioritized calls.

We present the following novel performance evaluation approximation method, referred as state space decomposition. Instead of evaluating the system using a one dimensional Markov chain model by grouping the nodes with the same total number of occupied channels regardless of the types of call, we group the nodes with the same number of calls of a certain type to obtain “supernodes” to compose a one dimensional Markov chain model for each type of call.

By grouping nodes with the same number of prioritized calls together to obtain supernodes, as shown in Fig. 1, we can frame a one dimensional Markov chain model that we can solve to obtain the steady state probabilities of each of these supernodes. The same approach can be utilized to group nodes that have the same number of non-prioritized calls together as shown in Fig. 2. In Fig. 1, we observe that for all supernodes except the ones that have at least one member node that represents a system state in which the total number of occupied channels is equal to the total number of channels in the system, $C$, there exist $(m+1)$ pairs of transitional flows between their member nodes and the corresponding member nodes that belong to their neighboring supernodes. Conversely, for the rest of the supernodes there exist some member nodes that do not have transitional flows in between any of the corresponding nodes that belong to their neighboring supernodes. Same can also be observed for the supernodes shown in Fig. 2; however in addition to those mentioned above there exist some other member nodes with unidirectional transition flows.

In Fig. 2, $k_i$ is the admission probability for prioritized calls when the system is in a state that is a member of a particular supernode $j$ where $j = 0, 1, \ldots, \lceil C/b_p \rceil$. It is similar to $\beta_i$; however, $\beta_i$ is a predefined user controlled parameter that indicates whether an arriving non-prioritized call will be admitted or not based on the number of occupied channels in the system as opposed to $k_i$ which is extracted from the multidimensional model of the system. We determine the values of the admission probabilities for prioritized calls, $k_i$, by obtaining the ratio of the sum of occupancy probabilities of the feasible member nodes of a supernode, for which the system admits an arriving prioritized call, to the sum of occupancy probabilities of all feasible member nodes of that particular supernode. Thus, when $j = 0, 1, \ldots, \lceil C/b_p \rceil$, the admission probabilities for prioritized calls, $k_p$, are equal to 1. The equilibrium channel occupancy probability when exactly $j$ prioritized calls exist, $q_p(j)$, where $j = 0, 1, \ldots, \lceil C/b_p \rceil$, can be obtained from the following recursive equation.

\[(\rho_p \cdot k_{j-1}) \cdot q_p(j-1) = j \cdot q_p(j), j = 1, \ldots, \lceil C/b_p \rceil \tag{1}\]
admitted, multiplied with supernode, for which an arriving non-prioritized call is occupancy probabilities of the feasible member nodes of a determine the values of prioritized calls. Similar to, yet slightly different than probability of an arriving non-prioritized call when qnp supernode. The equilibrium channel occupancy probabilities, probabilities of all the feasible member nodes of that particular nodes)

Therefore we initiate )((

Fig. 2. Transition diagram for asymmetric call admission control schemes with supernodes for non-prioritized calls.

Solving for \( q_p(0) \) in the equation \( \sum_{j=0}^{\lfloor c/b \rfloor} q_p(j) = 1 \), we obtain

\[
q_p(j) = \frac{1}{p_j} \cdot q_p(0), \quad 1 \leq j \leq \lfloor C/b \rfloor
\]  

(2)

where

\[
q_p(0) = \left[ 1 + \sum_{j=0}^{\lfloor c/b \rfloor} \frac{p_j \cdot k_j}{j!} \right]^{-1}
\]

(3)

Let \( h_r \), where \( r = 0, 1, \ldots, \lfloor m/b \rfloor - 1 \), denote the admission probability of an arriving non-prioritized call when \( r \) non-prioritized calls exist, regardless of the number of existing prioritized calls. Similar to, yet slightly different than \( k_j \), we determine the values of \( h_r \) by obtaining the ratio of the sum of occupancy probabilities of the feasible member nodes of a supernode, for which an arriving non-prioritized call is admitted, multiplied with \( b_j \) to the sum of occupancy probabilities of all the feasible member nodes of that particular supernode. The equilibrium channel occupancy probabilities, \( q_p(r) \), could be obtained similarly to prioritized calls if unidirectional transition flows, shown in Fig. 2, did not exist. However their existence needs to be taken into account by adjusting \( \mu_p \) affiliated with each supernode appropriately. Therefore we initiate \( \mu'_p(r) \) to replace \( \mu_p \) affiliated with each supernode in the model and determine its value by dividing the number of transition flows departing from the associated supernode with the number of pairs of bidirectional transition flows in between the same particular supernodes.

\[
\mu'_p(r) = \left[ \frac{(C-r)/b_p + 1}{(m-(r-1))/b_p} \right] \quad r = 1, \ldots, \lfloor m/b \rfloor
\]

(4)

Then we can obtain the occupancy probabilities \( q_p(r) \), \( r = 0, 1, \ldots, \lfloor m/b \rfloor - 1 \), which satisfy the recursive equation:

\[
(\lambda_{np} \cdot h_{r-1}) \cdot q_p(r) = r \cdot \mu'_p(r) \cdot q_p(r), \quad r = 1, \ldots, \lfloor m/b \rfloor
\]

(5)

Solving for \( q_p(0) \) in the equation \( \sum_{r=0}^{\lfloor m/b \rfloor} q_p(r) = 1 \), we obtain

\[
q_p(r) = \frac{\prod_{r=0}^{\lfloor m/b \rfloor} (\lambda_{np} \cdot h_r)}{r!} \cdot q_p(0), \quad 1 \leq r \leq \lfloor m/b \rfloor
\]

(6)

where

\[
q_p(0) = \left[ 1 + \sum_{r=0}^{\lfloor m/b \rfloor} \frac{\lambda_{np} \cdot h_r}{r!} \right]^{-1}
\]

(7)

The admission probabilities for both prioritized, \( k_j \), and non-prioritized calls, \( h_r \), cannot be obtained without computing the occupancy probability of each feasible node. Even if the occupancy probabilities of supernodes for prioritized and non-prioritized calls can be obtained using this method, we still need to compute the occupancy probabilities of certain feasible nodes since joint occupancy probabilities of these supernodes cannot be used due to their dependencies. To overcome these difficulties, we suggest the following iterative approach:

1. Initialize the value of estimated equilibrium occupancy probabilities \( \hat{q}(n_{np}, n_p) \) for \( n_{np} = 0, 1, \ldots, \lfloor m/b \rfloor \) and \( n_p = 0, 1, \ldots, \lfloor C/b \rfloor \) by setting them equal to \( 1/ (\text{total number of feasible nodes}) \).

2. Calculate \( \mu'_p(r) \) for \( r = 1, \ldots, \lfloor m/b \rfloor \) using (4).

3. Iterate with the following steps until the changes in the updated values of \( k_j \) and \( h_r \) are less than a chosen resolution.
3.1. Calculate and update \( k_j \) for \( j = 0,1,\ldots \) \( \lfloor C/b_p \rfloor - 1 \) and \( q_p(j) \) for \( j = 0,1,\ldots \) \( \lfloor C/b_p \rfloor \) using (2) and (3).

3.2. Update the values of the estimated occupancy probabilities, \( \hat{q}(n_{np}, n_p) \), by apportioning the value of the last updated occupancy probability, \( q_p(j) \), of the corresponding supernode for prioritized calls amongst its nodes with respect to the value of the last updated occupancy probability, \( q_{np}(r) \), of the corresponding supernode for non-prioritized calls.

3.3. Calculate and update \( h_r \) for \( r = 0,1,\ldots \) \( \lfloor m/b_{np} \rfloor - 1 \) and \( q_{np}(r) \) for \( r = 0,1,\ldots \) \( \lfloor m/b_{np} \rfloor \) using (6) and (7).

3.4. Update the values of estimated occupancy probabilities, \( \hat{q}(n_{np}, n_p) \), by apportioning the value of last updated occupancy probability, \( q_{np}(r) \), of the corresponding supernode for non-prioritized calls amongst its nodes with respect to the value of last updated occupancy probability, \( q_p(j) \), of the corresponding supernode for prioritized calls.

4. Obtain call blocking probabilities for prioritized and non-prioritized call using \( \hat{q}(n_{np}, n_p) \).

The call blocking probabilities for both types of calls are calculated as follows when the final estimated values of equilibrium occupancy probabilities, \( \hat{q}(n_{np}, n_p) \), are obtained.

\[
p_{pb} = \sum_{n_0} \hat{q}(n_0) \left( (C - (n \cdot b_p))/b_p \right)
\]

\[
p_{npb} = \left[ \lfloor m/b_{np} \rfloor \right] \sum_{a=0}^{\lfloor m/b_{np} \rfloor} \sum_{n=0}^{\left[ m/b_{np} \right]} \hat{q}(a,n) \quad (a \cdot b_{np} + n \cdot b_p) \geq \left[ m/b_{np} \right] \]

Despite its iterative nature, we expect the state space decomposition method to have a low computational complexity since decomposing the whole state space into subspaces and forming supernodes to apply one dimensional Markov chain modeling utilize the closed form formulas obtained from one dimensional Markov chain models and make the proposed method easy to implement for real time applications.

III. NUMERICAL RESULTS

In this section we compare the performance of the proposed method with Borst and Mitra’s approximation [15] and the direct numerical method. We investigate the cutoff priority scheme using the following set of parameters: \( C = 32, m = 24, \lambda_{np} = 0.1, 1/\mu_{np} = 200, 1/\mu_p = 50, b_p = 1 \) and \( 3, b_{np} = 1 \) and \( \lambda_p \) is varied from 1 to 0.05. Figs. 3 and 4 depict the prioritized and non-prioritized call blocking probabilities, respectively, under varying prioritized call traffic load. We observe for both values of \( b_p \) that when \( \rho_p > \rho_{np} \), both call blocking probabilities approximated by the proposed method match the exact results (direct) very well. However Borst and Mitra’s method overestimates the prioritized call blocking probability generously while it underestimates the non-prioritized one extensively. When \( \rho_p < \rho_{np} \), the proposed method slightly overestimates the prioritized call blocking probability while it underestimates the non-prioritized one with the discrepancy increasing as both traffic loads are decreasing. Yet, Borst and Mitra’s method gives a better approximation only when both traffic loads are very low due to its assumption on independent channel occupancy.

The discrepancy observed when \( \rho_p < \rho_{np} \) is due to an assumption that we made in the iterative solution described above, i.e., the steady state probabilities of all nodes that are members of the same particular supernode for prioritized calls are proportional to each other with the same ratio that exists between the steady state probabilities of the corresponding supernodes for non-prioritized calls, and vice versa. When the number of shared channels, \( m \), is increased to 28, call blocking probabilities for both types of calls are estimated more accurately since the corresponding transition diagram has more
supernodes for non-prioritized calls, each having relatively less number of member nodes. The following set of parameters are chosen in Figs. 5 and 6: $C = 32$, $m = 28$, $\lambda_{np} = 0.1$, $1/\mu_{np} = 200$, $1/\mu_p = 50$, $b_{np} = 1$ and $3$, $b_p = 1$ and $3$, $\lambda_p$ is varied from 1 to 0.05. The results are similar to the first case; however when $b_{np}$ increases, the discrepancy observed in the non-prioritized call blocking probabilities obtained from the proposed method increases while it decreases for the ones obtained from Borst and Mitra’s method. Still, this method gives a better approximation than the proposed method only when both traffic loads are very low. The computational complexity of the proposed method is comparable to the complexity of a closed form solution when the required computation time and memory are considered. The results are not presented here due to space constraints.

IV. CONCLUSION

In this paper we have proposed a novel computationally efficient approximation method that uses an iterative approach to evaluate the call blocking performance of asymmetric call admission control schemes in multi-service networks. Considering the high computational complexity of the existing numerical solution methods, we believe that providing performance evaluation approximation methods with low computational complexity will help motivate the practical implementation of dynamic call admission control schemes in cellular mobile networks.

REFERENCES


