Discrete Time Optimal Solution for the Connection Admission Control Problem

C. Bruni, F. Delli Priscoli, G. Koch, I. Marchetti

Abstract—The Connection Admission Control (CAC) problem is formulated in this paper as a discrete time optimal control problem. The control variables account for the acceptance/rejection of new connections and forced dropping of in-progress connections. These variables are constrained to meet suitable conditions which account for the QoS requirements (Link Availability, Blocking Probability, Dropping Probability). The performance index evaluates the total throughput. At each discrete time, the problem is solved as an integer-valued linear programming one. The proposed procedure was successfully tested against suitably simulated data.

Keywords—Connection Admission Control - Optimal Control - Integer valued Linear Programming - Quality of Service Requirements – Robust Control.

I. INTRODUCTION

This paper focuses on the Connection Admission Control (CAC), which is a key resource management procedure, and proposes a solution to the CAC problem based on control modeling and methodologies. In particular, the paper refers to the CAC procedure for an asynchronous Code Division Multiple Access (CDMA) based cellular system. For instance, the proposed CAC can be applied to the uplink of the Wideband CDMA (W-CDMA) technique adopted by the Universal Mobile Telecommunications Standard (UMTS), and it is suitable for both terrestrial and satellite networks. Recent literature on the CAC problem includes [1,2,3]. A general continuous time CAC problem was already discussed and solved by the authors in [4] with promising results.

As detailed in the following, the CAC is a fundamental procedure aiming at maximizing the exploitation of the available bandwidth of a cellular system and, at the same time, assuring the respect of the Quality of Service (QoS) requirements of the various Service Classes. By Service Class we mean the set of connections characterized by the same QoS requirements; in the following, \( k (k = 1, 2, ..., C) \) indicates the generic Service Class.

Whenever a stand-by user requests a connection set-up, the CAC is in charge of deciding whether to accept or to reject the connection set-up.

If the CAC accepts a new connection, the connection becomes "in progress" and remains in this status up to either its natural completion (in this case, the user decides to terminate the connection), or its forced dropping (in this case, as explained below, the network decides to terminate the connection); whichever is the reason, whenever a connection is terminated, the corresponding user comes back in the stand-by status.

While a connection is in progress, the data transmitted by the connection in question enhance the exploitation of the available bandwidth, but produce interference (self-noise) which contributes to decrease the ratio between the received energy per bit and the spectral noise density. In order to consider the link available, the above ratio must be higher than a threshold value. A fundamental QoS requirement is that the probability that the link is available (hereafter referred to as link availability requirement) does not become lower than a fixed threshold.

If the CAC rejects a new connection, a so-called block occurs and the user which has attempted the connection remains in the stand-by status. In this respect, a fundamental QoS requirement is that the probability of experiencing a block while attempting to set-up a connection (hereafter referred to as blocking probability requirement) does not exceed a threshold value.

Moreover, situations can occur in which, in order to avoid the infringement of the link availability requirement, it is necessary to forcibly drop one or more connections. In this respect, a fundamental QoS requirement is that the probability of experiencing a connection drop (hereafter referred to as dropping probability requirement) does not exceed a threshold value.

This paper models the CAC problem in a discrete time control based fashion. In particular, the plant models the connection and mobility dynamics in a given cell (reference...
cell), while the controller models the CAC mechanism. This one, basing on the plant variables which can be measured by the Base Station, computes the control variables which serve as plant inputs.

The CAC problem will be formulated as an optimal control problem subject to a set of constraints. As a matter of fact, the proposed controller - modelling the CAC mechanism - computes the control variables so that (i) a set of proper constraints, which model the QoS requirements (link availability, blocking probability and dropping probability), are respected and (ii) a proper performance index, which models the exploitation degree of the available bandwidth, is maximized.

II. A DISCRETE TIME DYNAMICAL MODEL FOR TRAFFIC IN A GIVEN CELL IN ITS ENVIRONMENT

A stochastic continuous time model for the in progress connections in a given cell has already been provided in [4] within the context of birth and death processes. Here the attention is focused on a discrete time model. Moreover, being the admissible control at each time based on all available traffic information up to that time, the model will be directly given in terms of mean values of in progress connections conditioned on the same information.

Let \( t_0 < t_1 < t_2 < \ldots < t_f \) be a sequence of discrete times. With reference to a given cell, let us denote by

- \( M_k(t_j | t_i) \) the mean value of the number of in progress connections \( M_k(t_j) \) in class \( k \), \( k = 1,2, \ldots, C \) at time \( t_f \), given all information about traffic available up to time \( t_1, \ldots, t_j \). Clearly, we have \( M_k(t_0 | t_i) = M_k(t_i) \).
- \( N(t_i) \) the number of stand-by mobile users at time \( t_i \).
- \( \lambda_k(t_i) \) the connection request rate of any mobile user relevant to class \( k \) at time \( t_i \).
- \( \mu_k(t_i) \) the sum of departure and termination rates of any mobile user with a connection in progress relevant to class \( k \) at time \( t_i \).
- \( \lambda_k^{(e)}(t_i) \) the arrival rate of mobile users with a call in progress relevant to the class \( k \) from neighbouring cells at time \( t_i \).
- \( u_k(t_i) \) the binary acceptance control in the Service Class \( k \) at time \( t_i \). If \( u_k(t_i) \) equals to one, then a new connection relevant to the Service Class \( k \) at time \( t_i \) is accepted; if \( u_k(t_i) \) equals to zero the same is rejected.
- \( v_k(t_i) \) the nonnegative integer valued dropping control in the Service Class \( k \). Specifically, \( v_k(t_i) \) denotes the number of connections relevant to the Service Class \( k \) which have to be forcedly dropped at time \( t_{i+1} \). Of course, \( v_k(t_i) \) cannot exceed \( M_k(t_i) \).

As previously mentioned, we now propose a dynamical model for \( M_k(t_{i+1} | t_i) \). This may be given by the following recursive relationship which stems from the usual discrete time approximation of birth and death processes:

\[
M_k(t_{i+1} | t_i) = M_k(t_i) - \mu_k(t_i)M_k(t_i) + \lambda_k^{(e)}(t_i)M_k(t_i) + \lambda_k(\eta(t_i) - M_k(t_i))
\]

(1)

In (1) the conditional mean increase of \( M_k \) over \( [t_i, t_{i+1}] \) is clearly proportional to the arrival rate \( \lambda_k^{(e)}(t_i) \) and the (accepted) birth rate \( \lambda_k(\eta(t_i) - M_k(t_i)) \), while the conditional mean decrease is proportional to the departure and termination rate \( M_k(t_i) \mu_k(t_i) \). The number \( v_k(t_i) \) of forcibly dropped connections also contributes to the decrease of \( M_k \).

III. QUALITY OF SERVICE CONSTRAINTS AND PERFORMANCE EVALUATION

In order to define the admissible control set, we now specify the quality of service constraints that the controls must manage to satisfy.

A. Link Availability Constraint

The requirement of link availability [5],[6] accounts for the fact that, in order to consider the link to be available, the ratio \( A(t) \) between the total power density received by the base station and the useful energy per bit received by the base station itself has to be lower than a given threshold over a sufficiently large percentage of time in the interval \( [t_i, t_{i+1}] \). Being this ratio a stochastic process, the formal constraint requires that the probability of \( A(t) \) falling below the threshold in \( [t_i, t_{i+1}] \), conditioned on all available traffic information up to \( t_j \), be sufficiently high.

This in turn may be approximated by the constraint that the conditioned mean value of \( A(t) \) be sufficiently small over a sufficiently large fraction of time in \( [t_i, t_{i+1}] \). In our discrete time context, this leads to a constraint of the type:

\[
A(t_{i+1} | t_i) \leq \eta(t_i)
\]

(2)

where \( A(t_{i+1} | t_i) \) denotes the mean value of \( A(t_{i+1}) \) conditioned on all available information up to \( t_i \), while \( \eta(t_i) \) represents a threshold which possibly accounts for the statistics of external and thermal power densities, as well as the last known value of the same ratio, that is \( A(t_i | t_i) = A(t_i) \). Constraint (2) must hold for a fraction of the total discrete time instants not lower than a fixed value \( l \).
Now, we obviously have:

\[
A(t_{i+1}) = \sum_{k=1}^{C} M_k(t_{i+1}) p_k(t_{i+1})
\]  

(3)

where \( p_k(t_{i+1}) \) denotes the ratio between the power density received by the base station from a class \( k \) connection at time \( t_{i+1} \), and the useful energy per bit received by the base station itself.

Accounting for the independence of \( u_k(t_j) \), \( v_k(t_j) \) (and therefore of \( M_k(t_{i+1}) \)) with respect to \( p_k(t_{i+1}) \) conditioned on all traffic information available up to \( t_j \), we have:

\[
A(t_{i+1} | t_j) = \sum_{k=1}^{C} M_k(t_{i+1} | t_j) p_k(t_{i+1} | t_j)
\]

(4)

where \( p_k(t_{i+1} | t_j) \) denotes the mean value of \( p_k(t_{i+1}) \) conditioned on the same information.

A simple forecast model provides the following relationship:

\[
p_k(t_{i+1} | t_j) = \sum_{j=0}^{\infty} e^{-\theta_k(t_{i+1}-t_j)} p_k(t_j)
\]

(5)

where \( p_k(t_{i+1} | t_j) \) appears as a convex combination of the known past values \( p_k(t_j) \), \( j \leq i \), via exponentially growing coefficients with a suitably chosen discounting rate \( \theta_k \).

Using (4) and (1), the link availability constraint (2) may be put in the form:

\[
\sum_{j=0}^{\infty} e^{-\theta_k(t_{i+1}-t_j)} p_k(t_j) \geq \rho_k
\]

\[+ \sum_{j=0}^{\infty} e^{-\theta_k(t_{i+1}-t_j)} p_k(t_j) \]

(10)

This is expressed by the constraint:

\[
\sum_{j=0}^{\infty} u_k(t_j) [n_k(t_{j+1}) - n_k(t_j)] + u_k(t_j) N(t_j) \lambda_k(t_j) (t_{i+1} - t_j) \geq \rho_k [n_k(t_j) + N(t_j) \lambda_k(t_j) (t_{i+1} - t_j)]
\]

where \( n_k(t_j), j = 0, 1, 2, \ldots, i \), is the known number of requested class \( k \) connections up to time \( t_j \).

Indeed, the two terms in the square brackets at the right hand side of (10) respectively account for requested calls up to \( t_j \) and mean value of requested calls in \( t_{j+1} \). The same meaning have the two terms at the left hand side, with reference to accepted calls.

From (10) we get:

\[ u_k(t_j) \geq \delta_k(t_j) \]

(11)

with:

\[ \delta_k(t_j) = \rho_k + \frac{\sum_{j=0}^{\infty} (\rho_k - u_k(t_j)) [n_k(t_{j+1}) - n_k(t_j)]}{N(t_j) \lambda_k(t_j) (t_{i+1} - t_j)} \]

(12)

\[ C. \text{ Dropped Calls Constraint} \]

This accounts for the need of keeping for each class the number of forcibly dropped calls up to time \( t_{i+1} \) not greater than a fixed fraction \( \rho_{dk} \) of the mean value of the total number of terminated calls up to \( t_{i+1} \) conditioned on all traffic information available up to \( t_j \).

This is expressed by the constraint:

\[ \sum_{j=0}^{\infty} v_k(t_j) \leq \rho_{dk} \left[ m_k(t_j) + \sum_{j=0}^{\infty} v_k(t_j) + u_k(t_j) M_k(t_j) (t_{i+1} - t_j) \right] \]

(13)

where \( m_k(t_j) \) is the known number of departed and spontaneously terminated class \( k \) connections up to \( t_j \).

From (13) we get:

\[ v_k(t_j) \leq \zeta_k(t_j) \]

(14)

with:

\[ \zeta_k(t_j) = \frac{\rho_{dk} [m_k(t_j) + \sum_{j=0}^{\infty} v_k(t_j) + \sum_{j=0}^{\infty} u_k(t_j) M_k(t_j) (t_{i+1} - t_j)]}{N(t_j) \lambda_k(t_j) (t_{i+1} - t_j)} \]
\[ \zeta_k(t_i) = \frac{\rho_{zk}}{1 - \rho_{zk}} \left[ m_k(t_i) + \mu_k(t_i) M_k(t_i)(t_{i+1} - t_i) \right] + \sum_{j=0}^{i-1} v_j(t_i) \]  
\[ \delta_k(t_i) \leq 1 \quad k = 1, 2, ..., C \]  \[ \zeta_k(t_i) \geq 0 \quad k = 1, 2, ..., C \]

D. Performance Evaluation

At each time \( t_i \), the performance of the CAC procedure will be evaluated in terms of the following index representing the expected value of the total throughput at time \( t_{i+1} \), conditioned on all traffic information available up to time \( t_i \):

\[ J(t_i, u_k(t_i), v_k(t_i)) = \sum_{k=1}^{C} w_k A(t_{i+1} | t_i) + \sum_{k=1}^{C} w_k b_k(t_i) v_k(t_i) \]  
\[ J(t_i, u_k(t_i), v_k(t_i)) = \widetilde{J}(t_i) + N(t_i) \sum_{k=1}^{C} w_k a_k(t_i) u_k(t_i) + \sum_{k=1}^{C} w_k b_k(t_i) v_k(t_i) \]  
where \( \widetilde{J}(t_i) \) is a suitable constant.

IV. THE OPTIMAL CONTROL PROBLEM AND ITS SOLUTION

From the content of Section III, we see that at each \( t_i \) the CAC problem may be formulated as an optimal control problem, defined by the control variables \( u_k(t_i) \in [0,1], v_k(t_i) \in [0,1, ..., M_k(t_i)], k = 1, 2, ..., C \) satisfying the constraints (6), (11), (12) and by the performance index (17).

This appears as a classical integer valued linear programming problem.

Being the admissible control set finite, an optimal solution exists whenever the set itself is non empty. A necessary and sufficient set of conditions for that is easily obtained for each \( t_i \) as:

\[ \delta_k(t_i) \leq 1 \quad k = 1, 2, ..., C \]  
\[ \zeta_k(t_i) \geq 0 \quad k = 1, 2, ..., C \]

\[ N(t_i) \sum_{k=1}^{C} a_k(t_i) \left[ \delta_k(t_i) \right] + \sum_{k=1}^{C} b_k(t_i) \min \left\{ \zeta_k(t_i), M_k(t_i) \right\} \leq c(t_i) \]
the proposed optimal control strategy when applied to a mobile communication network. The relevant traffic data come from simulated experiments. We carry on some validation and comparison of our procedure with respect to other existing popular control policies. In order to test the robustness of the proposed procedure, we also analyzed the effects obtained when the parameters of the forecast model for \( p_k \) are modified over a rather wide range.

Traffic data were generated by a specific software hereafter referred to as “Traffic Simulator” designed to represent a mobile communication network under various situations of traffic classes and intensity, number and mobility of users, geometry of cells. This software was developed within the SAILOR (Satellite Integrated UMTS Emulator) Project, supported by the European Union within the fifth framework program (Information Society Technology (IST) program). It is able to describe user mobility, traffic evolution in a network with a given number of cells and traffic classes of given types. In particular, the traffic classes considered in the software are voice, interactive, and background. In order to run, the software must be provided with the cells size, the amount of traffic in the various classes, the control policy, the simulation time.

The traffic simulator has been interfaced with a so-called CAC module implementing either the proposed CAC policy, or one of the reference CAC policies.

In the proposed CAC policy, the solution of each integer linear programming problem defined by the constraints (6),(11),(14) and by the performance index (17), is provided by a specifically devoted software package called CPLEX. For each \( t_n \), at first it is checked whether the set of necessary and sufficient conditions (18)-(20) are satisfied. In the affirmative case, the simulator provides to CPLEX all the variables which are required to formulate the corresponding optimal control problem, with fixed weights \( w_k, k = 1,2,...,C \), namely

\[ N(t_i), a(t_i), b(t_i), c(t_i), \delta(t_i), M_k(t_i), \delta_k(t_i), \]

\[ k = 1,2,...,C. \]

Then the solution \( \{u^a(t_i), v^a(t_i)\} \) of the problem (one of the solutions in case of non uniqueness) is computed by CPLEX and fed back into the simulator, to simulate the behaviour of the network. This provides an online optimal control of the simulated network at time \( t_{i+1} \).

In case one or more of the conditions (18), (19) are violated, the corresponding constraints (10), (13) are waived and replaced by equality constraints of the type:

\[ u_k(t_i) = 1 \quad \text{and} \quad v_k(t_i) = 0 \quad (29) \]

After that, in case the condition (20), with the possible replacements (29) of the control variables as above, is violated, then the corresponding constraint (6) is simply waived. The implementation of the steps above leads to the formulation of a (relaxed) control problem, characterized by a new (non empty) admissible control set and the same performance index (17) as before. A solution of such problem, still denoted by an abuse of notation by \( \{u^v(t_i), v^v(t_i)\} \), therefore exists and may again be found by the same CPLEX package. The rationale behind this procedure is twofold. On one hand, we must remove those constraints which cannot be met; on the other hand, we must strive, as we do in (5.1), to go as close as possible to the satisfaction of the constraints themselves, whenever this might help in the satisfaction of the same constraints in the subsequent time interval.

In order to validate the proposed strategy, its behaviour has been tested against traffic dynamics generated by the above mentioned Traffic Simulator.

The following performance parameters computed over the whole simulation time set \( \{t_0, t_1, ..., t_f\} \) have been considered:

- The a posteriori over all performance evaluation:

\[ J_p = \sum_{i=0}^{t_f} \sum_{k=1}^C w_k M_k(t_i)p_k(t_i) \quad (30) \]

- The fraction of time in which the rate \( A(t) \) did not exceed its threshold, hereafter referred as “a posteriori link availability”:

\[ L_a = \frac{1}{f+1} \sum_{i=0}^{f} I_{A(t_i) \leq \delta(t_i)} \quad (31) \]

- The blocked calls frequency for \( k = 1,2,...,C \):

\[ \Pi_{b,k} = 1 - \frac{\sum_{i=0}^{f} u_k(t_i) [n_k(t_{i+1}) - n_k(t_i)]]}{n_k(t_f)} \quad (32) \]

- The dropped calls frequency for \( k = 1,2,...,C \):

\[ \Pi_{d,k} = \frac{\sum_{i=0}^{f} v_k(t_i)}{m_k(t_f) + \sum_{i=0}^{f} v_k(t_i)} \quad (33) \]

In addition, to assess the optimality default introduced by the above mentioned relaxed conditions, we also computed the number of subintervals in which the optimal control problem did not admit a solution, hereafter referred to as “optimality default”.

The proposed strategy has been compared to other existing policies such as the Arrows and the Interference CAC . These CACs, like other similar ones described in literature, base their control policy only on the present traffic situation (ARROWS) [2] or take in account also the near past in order to compute a mean value of interference (Interference CAC)[1]. Instead, our proposed CAC policy employs a future traffic forecast, based on the present and past traffic, to estimate the traffic trend and decide on this basis the best
control values. In this way it is possible to better handle the traffic and by consequence, to improve the network efficiency.

In particular, in our simulations, we consider the following two classes ($C=2$):

- Conversational: voice call ($k=1$)
- Background Data: e-mail and ftp ($k=2$)

The Traffic Simulator was tuned so as to generate traffic according to a standard profile [8] [9]. The weights $w_k$, $k=1,2$, in (16) were set equal to 1. Thus the a posteriori overall performance evaluation $J_p$ (29) coincides with the total throughput. The thresholds have been set according to the following choices:

In Eq. (2): \[ \eta(t_i) = \eta = 4.07; \quad i = 0,1,..,f; \]

In Eq. (5): \[ \theta_1 = 60 \text{ sec}; \quad \theta_2 = 2.4 \text{ sec} \]

In Eq. (10): \[ \rho_{11} = \rho_{12} = \rho_1 = 0.95 \]

In Eq. (13): \[ \rho_{21} = \rho_{22} = \rho_2 = 0.02 \]

The duration of time interval $t_f - t_0$ was set equal to 1 hour, sufficiently long with respect to call and mobility dynamics. Time increments $t_{i+1} - t_i$ were set to 20 ms, sufficiently small with respect to $t_f - t_0$.

The following diagrams illustrate the simulation results for background data bit rate set equal to 150 kbps and increasing conversational traffic. In particular, Figs. 1, 2 and 3 illustrate blocked and dropped calls frequencies with their selected thresholds, for the proposed Optimal strategy, the ARROWS strategy and the Interference strategy. For figure clarity reason, the background dropped frequency ($\Pi_{b,2}$) is not shown: it actually is zero for ARROWS and Interference CAC’s, while it is negligible for Optimal CAC. Fig. 4 represents the a posteriori Link Availability with its selected threshold for three CAC’s. Fig. 5 shows the a posteriori overall performance evaluation for the three CAC’s up to their maximum allowable voice traffic. Finally, Fig. 6 describes the optimality default for the optimal CAC.
It appears from Fig.s 1-4 that for the fixed background data traffic (150 kbps) the ARROWS CAC and the Interference CAC allow the satisfaction of a posteriori QoS constraints, assessed as frequency constraints, up to a conversational traffic of 30 Erlang and respectively 34 Erlang, while the Optimal CAC achieves the limit of about 60 Erlang. From Fig. 5 it appears that, as far as the total throughput is concerned, the three CAC’s behave in a similar way as long as the QoS constraints are respected (∼30-35 Erlang of conversational traffic). It is of paramount importance that the Optimal CAC achieves a significant further improvement in the range up to 60 Erlang.

It is important to note that, differently from ARROWS and Interference (Fig.s 2-3), in our procedure (Fig. 1) the drop and block thresholds are attained at roughly the same conversational traffic (∼60 Erlang).

The good results achieved with respect to the other considered CAC procedures depends on the fact that the proposed CAC policy exploits at the best the degree of freedom offered by the QoS constraints, as already performed on a heuristical basis in a previous work of one of the authors [7]. In addition, in order to satisfy the link availability constraint, the available information over the forthcoming interference power density is exploited.

We then changed the block threshold \( 1 - \rho_1 \) and the drop threshold \( \rho_2 \). Fig.s 7-9 refer to the case \( 1 - \rho_1 = 0.005 \) and \( \rho_2 = 0.01 \), while Fig.s 10-12 refer to the case \( 1 - \rho_1 = 0.1 \) and \( \rho_2 = 0.05 \).
While, as expected, the throughput changes according to the variation of the thresholds, it is to be remarked that again, as already in Fig. 1, the drop and block thresholds in Fig.s 7-10 are attained at roughly the same conversational traffic. Finally, in order to test the robustness of the procedure, we iterated the experiments by changing the values of $\theta_1$ and $\theta_2$, which influence the long term memory behaviour of interference power density. See Fig.s 13-15 for the case when $\theta_1$ and $\theta_2$ increase by a factor of 5, and Fig.s 16-18 for the case when they are decreased by the same factor.
The fact that we obtained substantially similar results is evidence of low sensitivity of the procedure itself against possible uncertainties in the interference power density dynamical model.

A last remark concerns the computational time for optimal CAC: our data confirm the possibility that this CAC may be implemented on line. Over the time interval of 1 hour the number of subintervals, and therefore the number of optimisation problems, was 180000. Their solution required a total time less than 30 seconds, out of 10 minutes of total simulation time on a standard Pentium III Personal Computer.

REFERENCES