WIRELESS COMMUNICATIONS AND MOBILE COMPUTING
Wirel. Commun. Mob. Comput. 2008; **8**:501–511
Published online 22 November 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/wcm.467

Optimized bandwidth allocation with fairness and service differentiation in multimedia wireless networks

Nidal Nasser[1*,†] and Hossam Hassanein[2]
[1]*Department of Computing and Information Science, University of Guelph, Guelph, Ontario, N1G 2W1, Canada*
[2]*Telecommunications Research Laboratory, School of Computing, Queen's University, Kingston, Ontario, K7L 3N6, Canada*

Summary

In this article we present an optimal Markov Decision-based Call Admission Control (MD-CAC) policy for the multimedia services that characterize the next generation of wireless cellular networks. A Markov decision process (MDP) is used to represent the CAC policy. The MD-CAC is formulated as a linear programming problem with the objectives of maximizing the system utilization while ensuring class differentiation and providing quantitative fairness guarantees among different classes of users. Through simulation, we show that the MD-CAC policy potentially achieves the optimal decisions. Hence our proposed MD-CAC policy satisfies its design goals in terms of call-class-differentiation, fairness and system utilization. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: optimization; Markov decision process; call admission control; wireless cellular networks; multimedia handoff; QoS provisioning; radio resource management

1. Introduction

Future broadband wireless communication systems, such as the third Generation (3G) Universal Mobile Telecommunications System (UMTS) will extend current 2G voice-based wireless services to broadband multimedia services through enabling packet-switched technologies. Broadband multimedia services necessitate support of different classes of service with widely different bandwidth and Quality of Service (QoS), which need to be guaranteed by wireless cellular networks. To achieve this goal, QoS provisioning is critical.

In wireless cellular networks, two types of QoS parameters have been introduced at the connection level. These are: New Call Blocking Probability (NCBP) and Handoff Call Dropping Probability (HCDP) [1]. The NCBP is the percentage of new calls blocked. New call blocking occurs when all the resources of the wireless medium are used upon a new call request. The HCDP is the percentage of blocked handoff calls. When a user moves from one cell to a neighboring cell, a handoff takes place. The user requests resources from the new base station, and if all the resources are used dropping takes place. Therefore, one of the most important QoS issues in wireless cellular networks is how to

*Correspondence to: Nidal Nasser, Department of Computing and Information Science, University of Guelph, Guelph, Ontario, N1G 2W1, Canada.
†E-mail: nasser@cis.uoguelph.ca

reduce/control handoff drops due to lack of available resources in the new cell, since mobile users should be able to continue with their ongoing connections.

In UMTS and other 3G and beyond wireless cellular networks, QoS provisioning for emerging broadband multimedia services is a major challenge due to (i) the limited radio link bandwidth and (ii) the high rate of handoff events as the next generation of wireless cellular networks will use micro/pico cellular architectures in order to provide higher capacity. In earlier studies References [2,3], we attempted to overcome these challenges through an adaptive multimedia framework. In this article, we present an optimal and fair radio resource management policy for wireless cellular networks with non-adaptive multimedia services where the bandwidth of a call is fixed throughout the call lifetime.

The Radio Resource Management (RRM) module in the cellular network system is responsible for the management of air interface resources, for example, radio link bandwidth. Call Admission Control (CAC) is one of the most important components of RRM that affects the system utilization performance and QoS guarantees provided to users [4]. The function of the CAC policy is defined as a set of actions to determine if an arriving call request (new and handoff) can be accepted or rejected. One of the major challenges in the design of CAC policy is to provide preferential treatment among users while utilizing the system resources efficiently in congestion state of the system. Here, we mean by preferential treatment is to differentiate between users who want, for example, to pay more and hence receive a better QoS guarantee. Gibbens and Kelly in Reference [5] introduce pricing mechanisms to prevent congestion in future wireless networks. All users belonging to the same class are charged identical amounts. Clearly, users who pay more should be treated accordingly and belong to a higher priority class. The NCBP and the HCDP of higher priority classes should be lower. Such a requirement on NCBP and HCDP for different classes necessitates the use of a CAC policy.

Several CAC policies have been proposed to guarantee the QoS for handoff calls like the Guard Channel (GC) policy [6]. This policy reserves exclusively a portion of bandwidth (guard channels) allocated to a given cell for handoff calls. Clearly, increasing the number of guard channels will reduce the HCDP and, at the same time, it may increase the NCBP. Ramjee *et al*. [7] proved the GC scheme to be optimal for an objective function formed by a linear weighting of the new call blocking and the handoff dropping probabilities. However, both models assume that the traffic of all calls is identical. This assumption, however, is not valid if multimedia services are to be supported, since multimedia connections may differ in the amount of bandwidth they need to meet their QoS requirements. In order to achieve these requirements, many CAC policies have been presented to support the multimedia services in wireless cellular networks [8–11]. Such attempts utilized a stochastic control technique known as the Markov Decision Processes (MDPs) to construct optimal CAC policies that satisfy users QoS requirements and utilize the system resources efficiently. Yoon and Lee [8] proposed a CAC policy using MDP and used an approximation technique to solve the MDP-formulated problem that maximizes the system utilization. Xiao *et al.* [9] applied the MDP technique to rate-adaptive multimedia traffic to reallocate the system bandwidth to different service classes to maximize the revenue. Bartolini *et al.* [10] applied a MDP formulation to encompass the computational difficulties of the optimal solution when several classes of multimedia traffic are considered in content delivery networks. In Reference [11], Choi *et al*. studied the highway traffic control system with multiple traffic classes to maximize the revenue. We remark though that none of the schemes consider the relation between different service classes that we address in this article.

In this article, we use the stochastic control technique MDP to design an optimal CAC policy, which we call Markov Decision-based Call Admission Control (MD-CAC) for wireless cellular networks that support multimedia services, viz. UMTS. Unlike previous work, we consider three important factors in designing the CAC policy. These are service differentiation, fairness across call classes, and efficiency. Service differentiation requires establishing preferential treatment between different service classes in order to provide better QoS for higher priority classes. Defining a restrictive relation between QoS requirements of different classes provides a fair access to the system resources. Efficient utilization of system resources is essential due to the scarcity of bandwidth in wireless networks. The MD-CAC is formulated as a linear programming (LP) problem, which aims to maximize the system utilization while ensuring class differentiation and providing quantitative fairness guarantees for calls of the different classes. The optimal decisions of the MD-CAC policy are obtained by solving a set of MDP linear programming equations.

The rest of this article is organized as follows. The system model is described in Section 2. The design objectives of the CAC policy are discussed in Section 3.

DOI: 10.1002/wcm

MDP formulation and the optimal CAC policy are presented in Section 4. Simulation results are shown in Section 5. Finally, conclusions drawn from the article are discussed in Section 6.

2. System Model

We consider a multimedia wireless network with a cellular infrastructure, comprising a wired backbone and a number of base stations (BSs). The geographical area controlled by a BS is called a cell. The BS in each cell implements admission controller module that operates a call admission policy to utilize the system resources efficiently. A mobile communicates with others via a BS while residing in the cell of that BS. When a mobile leaves a cell, it could be either successfully handoff or dropped in case of shortage of resources in the new cell.

In this article, we are concerned with CAC in each cell. Therefore, we decompose the cellular network into individual sub-systems, each corresponding to a single cell. The correlation between these sub-systems, results from handoff calls between the corresponding cells, which is re-introduced as an input to each sub-model. Under this assumption, each cell can be modeled and analyzed individually. Thus, a same model can be used for all cells in the network, but the model parameters may be different, reflecting the mobility and traffic conditions in individual cells. Therefore, we can model the system at a single-cell level.

We assume that the system uses Fixed Channel Allocation (FCA), which means the cell has a fixed amount of capacity. No matter which multiple access technology (FDMA, TDMA, or CDMA) is used, we could interpret system capacity in terms of effective or equivalent bandwidth [12]. Hereafter, whenever we refer to the bandwidth of a call, we mean the number of Basic Bandwidth Units (bbu) that is adequate for guaranteeing desired QoS for this call with certain traffic characteristics.

Consider a cell that has a total capacity of B bbu. Traffic from new and handoff calls compete for the bandwidth (B). Traffic arriving at the cell is partitioned into K separate classes based on bandwidth requirements. The bandwidth of a class-i, that is, the number of basic bandwidth units required to accommodate the call, is given by b_i. The classes are indexed in an increasing order according to their bandwidth requirements, such that:

$$b_1 \leq \cdots \leq b_i \leq b_{i+1} \leq \cdots b_K$$

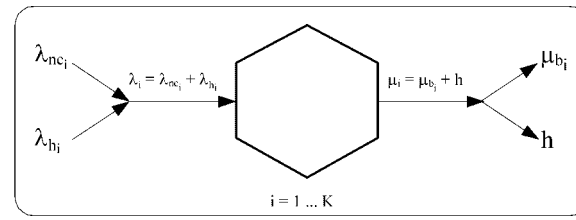

Fig. 1. Traffic model.

Here, the priority of a class-i call is proportional to i, with 1 indicating the lowest priority and with K indicating the highest priority.

For traffic characterization, we assume a simple model from the cell's perspective. New call arrivals and handoff call arrivals of class-i ($i = 1, 2, \ldots, K$) are assumed to follow a Poisson process with rates λ_{nc_i} and λ_{h_i}, respectively. The total arrival rate of class-i calls is $\lambda_i = \lambda_{nc_i} + \lambda_{h_i}$. The Call Holding Time (CHT) of a class-i call is assumed to follow an exponential distribution with mean $1/\mu_{b_i}$. For mobility characterization, we assume the Cell Residence Time (CRT), that is, the amount of time during which a mobile terminal stays in a cell during a single visit, is exponentially distributed with mean $1/h$ [13]. We assume that the CRT is independent of the service class; hence, calls in any class follow the same CRT distribution. Note that the parameter h represents the call handoff rate.

The channel holding time is the minimum of the CHT and the CRT. As the minimum of two exponentially distributed random variables is also exponentially distributed, then the channel holding time for class-i call is, therefore, assumed to be exponentially distributed with mean $1/\mu_i$ where $\mu_i = \mu_{b_i} + h$. Our traffic model is illustrated in Figure 1.

3. Design Objectives of the Call Admission Control Policy

When a class-i call (new or handoff) arrives, a call admission decision occurs in which the admission controller decides whether the call can be admitted or not based on the current state of the system and the overall objectives of the network. When the system is underutilized, all arriving calls (new and handoff) are admitted. However, if the system is fully utilized, preferential treatment should be given to high priority call classes to provide them with better QoS guarantees. In this work, the MD-CAC policy meets this objective by establishing preferential treatment among traffic classes based on the upper bound QoS requirements of each

DOI: 10.1002/wcm

class as follows:

$$QoS_K < QoS_{K-1} < \cdots < QoS_1$$

where QoS_i denotes the maximum acceptable QoS satisfaction level of a class-i, that is, the MD-CAC policy should guarantee that class-$i+1$ will have higher priority access to the system than class-i, for $i = 1, 2, \ldots, K-1$. For example, QoS_i may refer to the maximum allowable dropping probability of a class-i.

At the class-level, fairness across call classes is considered as another objective of the MD-CAC policy. Fairness is obtained by defining a pair-wise relation between traffic classes, such that the absolute difference between the rejection (blocking or dropping) probabilities of consecutive classes does not exceed a deviation value ε:

$$\left|P_{r_i} - P_{r_{i-1}}\right| < \varepsilon, \quad i = 2, \ldots, K; \quad 0 < \varepsilon < 1$$

where P_{r_i} is the rejection probability of class-i. The value of ε can be set to be the maximum tolerable deviation from the desired relation between the rejection probabilities of classes.

While providing preferential treatment and fairness which is important from the user point of view, the network service provider also has to maintain high system utilization. Therefore, another objective of our CAC policy is to maximize the system utilization which we have elaborated in the following section.

4. Optimal Call Admission Control Policy

In this section, we represent our Call Admission Control (CAC) policy as a MDP. Then, the MD-CAC policy is formulated as a LP problem to obtain optimal decisions. Before we proceed an overview of MDP is given.

4.1. Overview of Markov Decision Process

A MDP is a stochastic process that describes the evolution of dynamic systems controlled by sequences of decisions or actions. The dynamic system at random discrete Time Points (epochs) is observed and classified into one of a finite number of states. After classification, one of a finite number of possible actions must be chosen and the corresponding revenue for each state is gained due to this decision [14]. For each state x, a set of actions is available. If the system is in state x and action a is chosen then:

(1) The next state, y, of the system is chosen according the transition probability p^a_{xy}

(2) The time until the transition from x to y occurs is $\tau(x, a)$.

After the transition occurs, an action is again chosen and (1) and (2) are repeated. Markovian properties are satisfied if at a decision epoch the action a is chosen in the current state x, and the state at the next decision epoch depends only on the current state x and the chosen action a. They are thus independent of the past history of the system.

4.2. Markov Decision Process Formulation

Whenever there is an arrival of a class-i call (new or handoff), CAC must make some decisions based on the current system state. The decisions include whether the system accepts the call or not. The CAC policy selects its decision from a finite decision (action) space. The possible decision for all traffic types (new or handoff) is {accept/reject}.

The MDP formulation is characterized by the following five components:

4.2.1. State space

We denote the current state of the system by a vector x of K components as follows:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_K) \tag{1}$$

The non-negative integer x_i denotes the number of ongoing class-i calls (new and handoff). Let Ω denote the state space of the system as it evolves over time. We define Ω of all admissible states as:

$$\Omega = \left\{ x = (x_1, x_2, \ldots, x_K) : \sum_{i=1}^{K} x_i b_i \leq B; x_i \geq 0 \right\} \tag{2}$$

At each state x, the CAC policy should select an accept/reject decision for new or handoff calls. The set of all the selected decisions are called the admission policy of the system.

4.2.2. Decision epochs and action space

When an arriving call (new or handoff) of class-i desires to be admitted into the system, the MD-CAC policy will make a decision as to whether to grant admission or not. Note that the decision is made only at the occurrence of a call arrival (new or handoff) of class-i. Events of

DOI: 10.1002/wcm

call completion or handoff to other cells do not require decisions to be made.

We define a decision epoch as the time immediately following an arrival event. A decision epoch is a vector $v = (x, e)$, where x is the vector of class-i calls as in (1) and $e \in \{a_{nc_1}, \ldots, a_{nc_K}, a_{h_1}, \ldots, a_{h_K}\}$. The variable e represents the event type of an arrival and the indicators a_{nc_i}, and $a_{h_i} \in \{0, 1\}$ for $i = 1, \ldots, K$, denote the origination of class-i new call within the cell and the arrival of a class-i call due to handoff from an adjacent cell, respectively.

When the system is in state x, an accept/reject decision must be made for each type of possible arrival, that is, an origination of a class-i new call, or the arrival of a class-i handoff call. Thus, the action space A can be expressed as follows:

$$A = \{a = (a_{nc_1}, \ldots, a_{nc_K}, a_{h_1}, \ldots, a_{h_K}) : a_{nc_i}, a_{h_i} \in (0, 1), i = 1, \ldots, K\} \quad (3)$$

where $a_{nc_i} = 0$ (or 1): reject (or accept) the new call of class-i, and $a_{h_i} = 0$ (or 1): reject (or accept) the handoff call of class-i. For example, when $K = 2$ and $(a_{nc_1}, a_{nc_2}, a_{h_1}, a_{h_2}) = (0, 1, 1, 1)$ this indicates that only new calls of class-1 will be rejected by the MD-CAC policy.

For a given state $x \in \Omega$, the state action space, $A_x \subset A$, is defined as follows:

$$A_x = \{a \in A : a_{nc_i} = 0 \quad \text{and} \quad a_{h_i} = 0 \quad \text{if} \quad y = x + e_i \notin \Omega, i = 1, \ldots, K\} \quad (4)$$

Under this action space, a class-i call is accepted at state x if and only if $y = x + e_i \in \Omega$. In other words, A_x is composed of all those actions in A that do not result in a transition to a state $y = x + e_i \notin \Omega$. Here e_i is a vector of zeros, except for the ith component which is equal to 1.

4.2.3. Expected sojourn time

If the system is in state $x \in \Omega$ and the action $a \in A_x$ is chosen, then the expected sojourn time until a new state is entered is given by $\tau(x, a)$. The value of the expected sojourn time can be expressed by:

$$\tau(x, a) = \left[\sum_{i=1}^{K} \left(x_i\mu_i + \lambda_{nc_i}a_{nc_i} + \lambda_{h_i}a_{h_i}\right)\right]^{-1} \quad (5)$$

where $x_i\mu_i$ represents the rate at which calls terminate within the cell, $\lambda_{nc_i}a_{nc_i}$ represents the rate at which new calls originate from the cell, and $\lambda_{h_i}a_{h_i}$ represents the rate of incoming handoffs from adjacent cells to this cell.

4.2.4. State dynamics

The state dynamics of a MDP is completely specified by the state transition probabilities. Let p^a_{xy} be the transition probability that at the next decision epoch the system will be in state y if the present state is x and action a is chosen, where $a \in A_x$. For $y \in \Omega$, we have the following cases:

$$\begin{aligned} \text{Class-}i \text{ arrival: } & y = x + e_i : p^a_{xy} \\ & = (\lambda_{nc_i}a_{nc_i} + \lambda_{h_i}a_{h_i})\tau(x, a) \\ \text{Class-}i \text{ departure: } & y = x - e_i : p^a_{xy} = (x_i\mu_i)\tau(x, a) \\ \text{Otherwise: } & p^a_{xy} = 0, \quad \forall i \in \{1, \ldots, K\}, \quad x \in \Omega \end{aligned} \quad (6)$$

4.2.5. A revenue function

Let $r(x, a)$ be the revenue rate when the call is in state x and action a is chosen. If r_i is the revenue rate of class-i, the total revenue rate for the system is calculated by:

$$r(x, a) = \sum_{i=1}^{K} r_i x_i \quad (7)$$

Assuming that revenue is given in terms of the number of basic bandwidth units assigned, the total revenue rate in state x is equal to the system utilization in state x, and is given by

$$r(x, a) = \sum_{i=1}^{K} b_i x_i \quad (8)$$

where r_i is replaced by the assigned bandwidth b_i.

4.3. Constructing Optimal MD-CAC Policy

MDPs are usually analyzed and solved within the framework of discrete-time average cost Markov decision processes; see Reference [14] for a detailed discussion. We define the decision variable π_{xa}, $x \in \Omega$ $a \in A_x$, as the long-run fraction of time at which the state x chooses action a, and the set of π_{xa} collectively determines the MD-CAC policy. Searching for the optimal CAC policy is equivalent to finding the decision variables for all states. This can be achieved by solving the following MD-CAC LP formulation,

DOI: 10.1002/wcm

which aims to maximize long-run system utilization and to guarantee QoS.

$$\text{Maximize} \sum_{x\in\Omega}\sum_{a\in A_x} r(x,a)\tau(x,a)\pi_{xa} \quad (9)$$

$$\text{Subject to:} \sum_{x\in\Omega}\sum_{a\in A_x} \tau(x,a)\pi_{xa} = 1 \quad (10)$$

$$\sum_{a\in A_x} \pi_{ya} = \sum_{x\in\Omega}\sum_{a\in A_x} p^a_{xy}\pi_{xa},\ y\in\Omega \quad (11)$$

$$\pi_{xa} \geq 0, x\in\Omega, a\in A_x \quad (12)$$

The term $\tau(x,a)\pi_{xa}$ in Equation (9) can be interpreted as the long-run fraction of decision epochs when the system in state x and action a is chosen. Hence, the objective function (9) is the system utilization. The constraints are: the normalization condition (10), the stationary global balance Equation (11), and Equation (12) is the constraint on decision variables. The optimal feasible solutions π^*_{xa} to Equation (9) give the optimal CAC policy.

A nice feature of the LP formulation for solving MDP is that it permits optimization over additional constraints. Since dropping handoff calls is usually less desirable and less tolerable than blocking newly initiated calls, we focus on the handoff dropping probability as the main QoS requirement. Thus, we consider the QoS requirements of the upper bound on the handoff dropping probability of each class-i and the relation between the handoff call dropping probabilities of different call classes. Hence, throughout the rest of this article QoS_i will be used to refer to the maximum allowable dropping probability of a class-i. Let $d(x,a) = 1 - a_{h_i}$. The stationary handoff call dropping probability for class-i when action a is selected is given by Reference [15]:

$$P_{d_i} = \sum_{x\in\Omega}\sum_{a\in A_x} d(x,a)\tau(x,a)\pi_{xa}, i = 1,\ldots,K \quad (13)$$

We, therefore, add the QoS constraint of the handoff dropping probability to Equation (9) to limit the handoff call dropping probabilities for class-i below a target value QoS_i:

$$P_{d_i} \leq \text{QoS}_i, \quad i = 1,\ldots,K \quad (14)$$

As our MD-CAC objective also includes providing fairness across call classes. The system should satisfy the following constraint, discussed in Section 3:

$$\left|P_{d_i} - P_{d_{i-1}}\right| < \varepsilon, \quad i = 2,\ldots,K;\ 0 < \varepsilon < 1 \quad (15)$$

We, therefore, also add this fairness constraint to Equation (9).

We remark that, even though the complexity and solution space for the LP formulas (9)–(15) increases exponentially as B and K increase, there are techniques for solving large-scale LP problems such as interior-point method [16] can be applied to the cases of large cellular systems and/or cellular systems with larger capacity.

5. Simulation Results

In this section, the performance of the MD-CAC policy is evaluated using simulation. We first describe the simulation model, that is, used in this article. We then show that the MD-CAC policy potentially achieves the optimal solution which satisfies its design goals in terms of call class differentiation, fairness, and system utilization.

5.1. Simulation Model

We simulate one cell with a diameter of the cell is 1 km (i.e., micro cellular environment). The BS resides at the center of the cell. Part of the BS is the Admission Controller (AC) that operates the CAC policy. The AC components are shown in Figure 2 and operate as follows. Given the current state of the system and the traffic parameters, MDP components are calculated. Then, the AC uses the LP technique to compute the optimal decision value, π^*_{xa}, that aims at maximizing the system utilization function as given in Equation (9). Solving the LP problem to find the optimal call admission decisions is an offline procedure, that is, the decisions are obtained before invoking the CAC mechanism.

The following describe the assumptions used in our simulation:

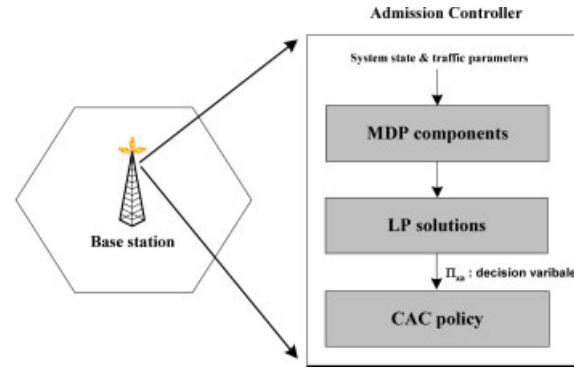

Fig. 2. Admission controller module.

DOI: 10.1002/wcm

- The total bandwidth capacity of the simulated cell is B basic bandwidth units.
- Two classes of multimedia service are considered. Bandwidth requirement of class-i connection is b_i ($i = 1, 2$). Class-2 has a higher priority than class-1.
- New call requests and handoff call requests of class-i are generated in the cell according to a Poisson process with rate λ_{nc_i} and λ_{h_i} (calls/s), respectively. The total arrival rate of class-i calls is $\lambda_i = \lambda_{nc_i} + \lambda_{h_i}$. A newly generated call can randomly appear in the cell with an equal probability. We assume the call holding time of a class-i is exponentially distributed with mean $\mu_{b_i}^{-1}$ (s). Also, we assume that the cell residence time is exponentially distributed with mean h^{-1} (s). The channel holding time of class-i calls is exponentially distributed with mean $\mu_i^{-1}(\mu_i = \mu_{b_i} + h)$. Also, the handoff call arrival rate of class-i is assumed to be proportional to the new call arrival rate of class-i by $\lambda_{h_i} = (h/\mu_{b_i})\lambda_{nc_i}$ for $i = 1, 2$.
- Mobiles can travel in one of eight directions with equal probability. A constant randomly selected speed is assigned to a mobile when it enters a cell either at call initiation or after handoff. The speed is obtained from a uniform probability distribution function ranging from $V_{\min}$ to $V_{\max}$.
- We assume that the maximum allowable handoff dropping probability for class-2 calls, QoS_2, always has a lower value than QoS_1 since class-2 traffic has higher priority than class-1.

The simulation model is very flexible and allows us to test the system under different scenarios. Here, we limit our experimental tests to the simulation parameters values that are shown in Table I. However, we believe that the higher the bandwidth capacity is, the more efficiency our policy can achieve.

5.2. Performance Evaluation

In this section, we show by simulation that our MD-CAC achieves the optimal solution and guarantees QoS parameters at the same time. We develop two case studies to explore the comprehensive effect of the policy on system behavior when traffic parameters vary. In the first case, we exclude the fairness constraint (Equation 15) from the LP optimization problem (Equation 9) to show that our policy meets the QoS dropping probability constraint for traffic classes. In the second case, we add the fairness constraint to LP optimization problem to provide fair access to the system resources.

Table I. Simulation parameters.

Parameter	Value	Unit
B	10	bbu
b_1	1	bbu
b_2	2	bbu
$1/\mu_{b_1} = 1/\mu_{b_2}$	500	s
$1/h$	100	s
$V_{\min}$	10	km/h
$V_{\max}$	60	km/h

5.2.1. *Case study one: excluding fairness constraint*

In this case, we evaluate the proposed CAC policy without the fairness constraint (Equation 15). For this purpose we develop two sets of experiments. In the first set, in order to simplify the problem, we only consider one class of traffic with different upper bound dropping probability values. In the second set, we increase number of classes to two and compare the proposed policy with upper limit (UL) policy [17]. In all experiments below we obtain the performance measures, HCDP and bandwidth utilization as the offer load to the system (cell) is varying. Note that the offered load to the system by each traffic type is ρ_i (Erlang load) and $\rho_i = \lambda_i/\mu_i$.

5.2.1.1. Single class case. Here, we consider only one class of traffic (class-1). Thus, Erlang load is defined as $\rho_1 = \lambda_1/\mu_1$. In our simulation, we evaluate the system with different *QoS* values as shown in Figures 3 and 4. Simulation parameters are the same as in Table I.

The effect of varying the load on the QoS parameter HCDP is illustrated in Figure 3. Obviously, the HCDP increases as the load and QoS increases. However, our proposed MD-CAC policy achieves an almost constant HCDP that is bounded by the maximum allowable HCDP, even under extremely high loading

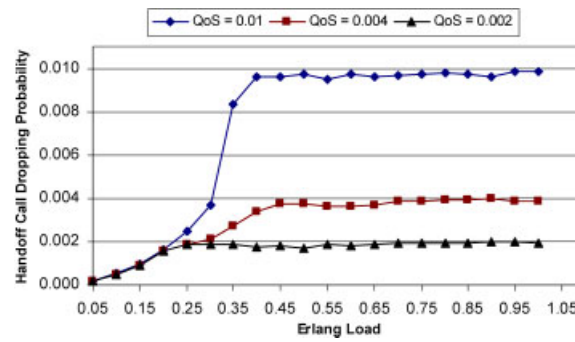

Fig. 3. The effect of QoS on handoff call dropping probability (single class).

DOI: 10.1002/wcm

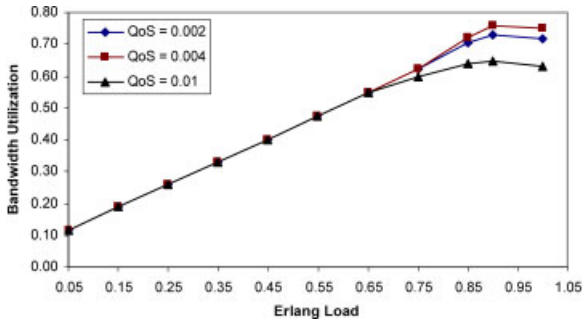

Fig. 4. The effect of QoS on bandwidth utilization.

conditions. Therefore, QoS requirements are satisfied. Thus, higher QoS requirements will reduce bandwidth utilization only on high traffic load condition.

Figure 4 shows the bandwidth utilization for different QoS as the load increases. As light and medium load levels, the bandwidth utilization has almost the same value irrespective of QoS and increases linearly with increase of load. However, at heavy load, the bandwidth utilization decreases in spite of the increase of the offered load. This phenomenon becomes more serious with stricter QoS requirements. This results from the fact that handoff calls are dropped with high probability at heavy load. Since most of handoff calls are dropped, the load decreases and it in turn results in the decrease of bandwidth utilization.

5.2.1.2. Two-class case. In this section, we design several experiments to compare our proposed policy to the UL CAC policy with respect to the HCDP and the bandwidth utilization. The UL CAC policy blocks a new call request of class-i if the number of the calls is greater or equal to an upper-limit value, that is, threshold t_i. The UL CAC policy used for comparison has a threshold $t_1 = 3$ and $t_2 = 5$. Simulation parameters are the same as in Table I.

Figure 5 shows the HCDP for both policies as the load increases. The maximum handoff call dropping probability for class-1 and class-2 are equal to $\mathrm{QoS}_1 = 0.04$ and $\mathrm{QoS}_2 = 0.02$, respectively. It is shown that the HCDP for the MD-CAC policy is bounded by 0.04 and 0.02 for class-1 and class-2 connections, respectively, and therefore, their QoS requirements are satisfied. On the other hand, the UL CAC policy cannot guarantee such bound, especially when the load increases.

Different cases were investigated as shown in Table II to compare the two policies in terms of the bandwidth utilization. The bandwidth utilization of both policies is obtained while the maximum allowable handoff call dropping probability, QoS_i, is varied. The last column of Table II shows the Utilization Improvement Ratio (UIR) of our policy over the UL CAC policy. The UIR is obtained as follows:

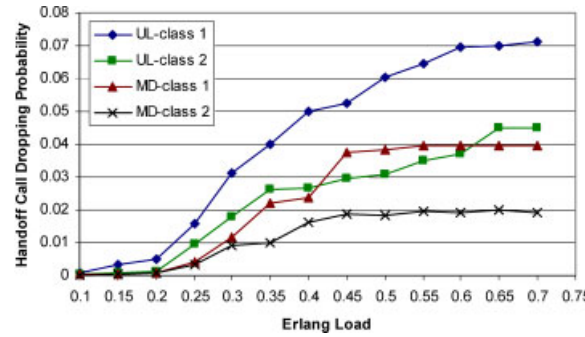

Fig. 5. The effect of QoS on handoff call dropping probability (two classes).

Table II. Utilization improvement results.

No.	QoS_1	QoS_2	U_{MD} %	U_{UL} %	UIR %
1	0.03	0.01	29.08	26.88	*8.18*
2	0.04	0.02	28.10	26.88	*4.52*
3	0.06	0.04	27.01	26.88	*0.48*

$$\mathrm{UIR} = \frac{U_{\mathrm{MD}} - U_{\mathrm{UL}}}{U_{\mathrm{UL}}} * 100,$$

where U_{MD} is the utilization of the Markov decision CAC policy and U_{UL} is the utilization of the upper-limit policy.

Figures 6–8 demonstrate the effect of varying the load on the bandwidth utilization for both policies considering all the cases of Table II. It is observed that the MD-CAC policy (labeled MD in the Figures) over MDP always achieves higher utilization than the UL CAC policy. Also, we observe that as the upper bound for the handoff dropping probabilities increases, the bandwidth utilization decreases. This clearly indicates

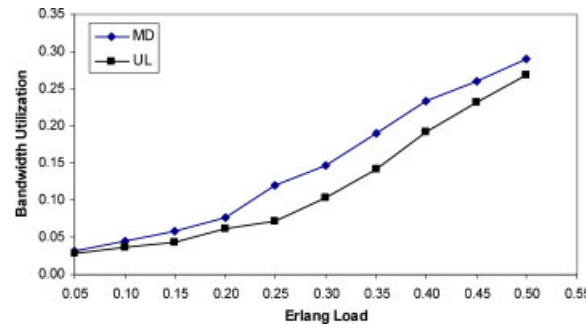

Fig. 6. Bandwidth utilization versus erlang load ($\mathrm{QoS}_1 = 3\%$, $\mathrm{QoS}_2 = 1\%$, $t_1 = 3$, $t_2 = 5$).

DOI: 10.1002/wcm

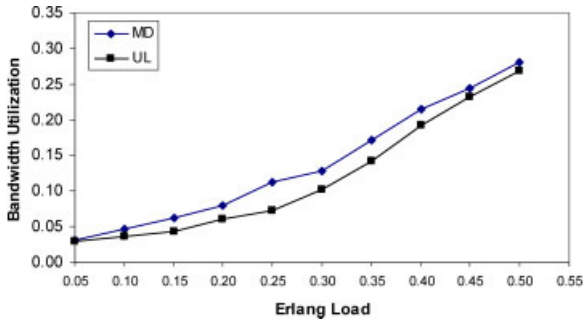

Fig. 7. Bandwidth utilization versus erlang load ($QoS_1 = 4\%$, $QoS_2 = 2\%$, $t_1 = 3$, $t_2 = 5$).

the conflicting relationship between the bandwidth utilization and QoS. Further results presenting the utilization improvement are summarized in Table II.

5.2.2. *Case study two: including fairness constraint*

In this case, we study the behavior of the MD-CAC policy when the fairness constraint is added to the formulated LP optimization problem. We consider two classes of traffic. The maximum handoff call dropping probability for class-1 and class-2 are fixed and equal to $QoS_1 = 0.03$ and $QoS_2 = 0.01$, respectively. The absolute difference between the handoff call dropping probabilities of the two call classes must be smaller than 10^{-2} (i.e., $\varepsilon = 10^{-2}$). In all experiments below we fix the call arrival rate of class-2 calls ($\lambda_2 = 1$ call/s) and treat the call arrival rate of class-1 calls (λ_1) as a parameter.

Figure 9 demonstrates the effect of varying the call arrival rate of class-1 (λ_1) on the behavior of the optimal CAC policy for two case studies described below. Note that each point (x_1, x_2) in the figures represents the optimal action decision that the admission controller

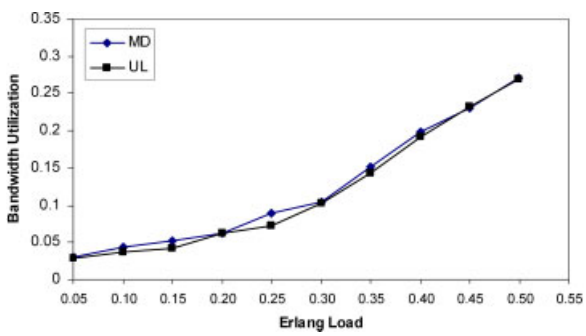

Fig. 8. Bandwidth utilization versus erlang load ($QoS_1 = 6\%$, $QoS_2 = 4\%$, $t_1 = 3$, $t_2 = 5$).

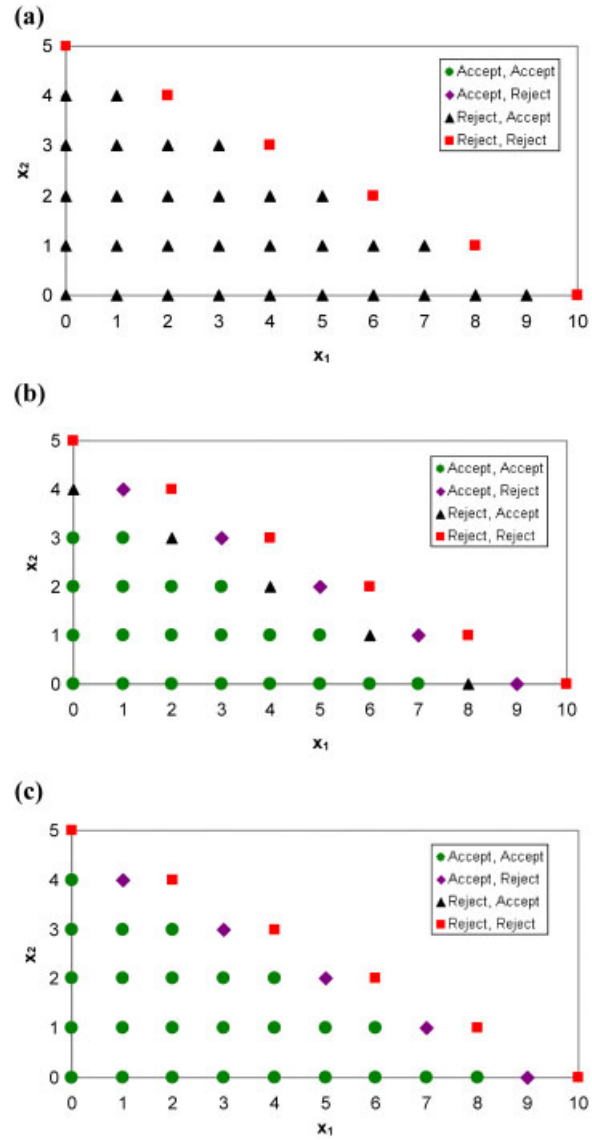

Fig. 9. The effect of fairness constraint (a) λ_1 =0.01 (call/s); (b) λ_1 =0.45 (call/s); (c) λ_1 =1 (call/s).

selects. To clarify this we introduce the following example. In Figure 9(a) when there is one call of class-1 and one call of class-2 (i.e., state (1, 1)), the system can apply one of the following actions: (Accept, Accept), (Accept, Reject), (Reject, Accept), or (Reject, Reject). The solution of the LP problem will determine which action the system should select at that state in order to maximize the system utilization. In this example, the optimal action decision solution is (Reject, Accept) which represents the point (1, 1) in Figure 9(a).

Figure 9 shows the results for three different values of λ_1. For the very low value of λ_1 (see Figure 9(a), $\lambda_1 = 0.01$), the optimal policy is accepting only class-2

DOI: 10.1002/wcm

calls. This is because the call arrival value of class-2 calls is very high compared to the call arrival rate of class-1 calls. Thus, bandwidth in the system is occupied by a class-2 calls which translates in high bandwidth utilization. Therefore, the policy will keep accepting class-2 calls rather than accepting class-1 calls in all system states to increase the system utilization and thus making the effect of service differentiation clear. However, the behavior of the policy will be different as the call arrival value of class-1 calls increase as shown in Figure 9(b,c). In Figure 9(b), since a class-2 call has higher priority than a class-1 call, the optimal policy rejects class-1 calls when the system state satisfies

$$x = \left\{ (x_1, x_2) \,|\, \sum_{i=1}^{2} x_i b_i = 8 \right\}.$$

Otherwise, the policy admits all call classes. The advantage of the fairness constraint becomes more evident in Figure 9(c) where both call arrival rates are equal ($\lambda_1 = \lambda_2 = 1$). This follows since a highly restrictive fairness constraint ensures that the absolute difference between the handoff call dropping probabilities of the two call classes does not exceed the value 10^{-2}. As a result, the policy admits all call classes with only a moderate sacrifice in the system utilization.

In Figure 10, we plot the bandwidth utilization obtained for the above cases (Figure 9) as a function of class-1 call arrival rate (λ_1). Another value of ε is also used to demonstrate the effect of the limit on fairness deviation. Figure 10 shows that for low values of λ_1, the bandwidth utilization linearly increases for both values of ε until λ_1 reaches a value of 0.5. However, as λ_1 increases (beyond 0.5), we observe that the bandwidth utilization when $\varepsilon = 10^{-5}$ is lower than in case $\varepsilon = 10^{-2}$ as expected. This is because by increasing the fairness deviation value, the system will have less dropping rates for all classes and this results in a more efficient usage of the system bandwidth. Therefore, the better utilization comes at the cost of a large difference in the handoff call dropping probabilities of the two call classes ($\varepsilon = 10^{-2}$).

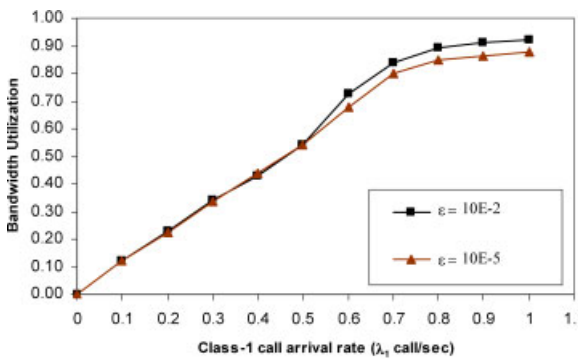

Fig. 10. Bandwidth utilization for different values of fairness deviation.

6. Conclusions

Providing multimedia services with QoS guarantees in next generation wireless cellular networks poses great challenges due to the scarce radio bandwidth. Effective CAC is important for the efficient utilization of the limited bandwidth. In this article, we proposed an optimal CAC policy that aims at maximizing the system utilization while stratifying the QoS constraints bounding the handoff dropping probability of each traffic class and achieving quantitative fairness among different call classes. We used the MDP technique to represent the proposed CAC. The optimal CAC decisions for each state are found by solving the linear programming formulation problem.

Simulation results show that the system upholds the handoff call dropping probability required by each traffic class while providing fairness for all classes. The requirements of the mobile users are hence satisfied in periods with different loads, including overload. Moreover, the implemented policy ensures efficient utilization of resources. This latter facet is highly desirable by service providers.

To the best of our knowledge, our work presents the first attempt to design an optimal CAC policy that satisfies the above goals.

References

1. Rappaport SS. The multiple-call hand-off problem in high-capacity cellular communications systems. *IEEE Transactions on Vehicular Technology* August 1991; **40**(3): 546–557.
2. Nasser N, Hassanein H. Connection-level Performance Analysis for Adaptive Bandwidth allocation in multimedia wireless cellular networks. *Proceedings of the IEEE International Performance Computing and Communications Conference* (IPCCC), Phoenix, Arizona, April 2004; 61–68.
3. Nasser N, Hassanein H. Prioritized multi-class adaptive framework for multimedia wireless networks. *Proceedings of the IEEE 2004 International Conference on Communications* (ICC), Paris, France, June 2004; 917–922.
4. Zander J, Kim S-L, Almgren M, Queseth O. *Radio Resource Management for Wireless Networks*. Artech House Publishers: Norwood, MA, 2001.
5. Gibbens RJ, Kelly FP. Resource pricing and the evolution of congestion control. *Automatica* 1999; **35**(12): 546–557.
6. Hong D, Rappaport SS. Traffic model and performance analysis for cellular mobile radio telephone systems with prioritized

DOI: 10.1002/wcm

and non-prioritized handoffs procedures. *IEEE Transactions on Vehicular Technology* 1986; **35**(3): 77–92.
7. Ramjee R, Nagarajan R, Towsley D. On optimal call admission control in cellular networks. In *IEEE INFOCOM'96*, 1996; 43–50.
8. Yoon I, Lee B. A Distributed dynamic call admission control that supports mobility of wireless multimedia users, *ICC'99*, 1999; 1442–1446.
9. Xiao Y, Chen CLP, Wang Y. Optimal admission control for multi-class of wireless adaptive multimedia services. *IEICE Transactions on Communications* April 2001; **E84-B**(4): 795–804.
10. Bartolini N, Casalicchio E, Tucci S. Optimal mobility-aware admission control in content delivery networks. *IEEE MASCOTS 2003*; 234–237.
11. Choi J, Kown T, Choi Y, Naghshineh M. Call admission control for multimedia services in mobile cellular networks: A Markov decision approach, *IEEE Symposium on Computers and Communications (ISCC)*, 2000; 549–599.
12. Kesidis G, Warlrand J, Chang C. Effective bandwidths for multi-class Markov fluides and other ATM sources. *IEEE Transactions on Networking* 1993; **1**(4): 424–428.
13. Yeung K, Nanda S. Channel management in microcell/macrocell cellular radio systems. *IEEE Transactions on Vehicular Technology* 1996; **45**: 601–612.
14. Tijms H. Stochastic Modeling and Analysis: A Computational Approach, Wiley, New York, 1986.
15. Hyman J, Lazar A, Pacifici G. A separation principle between scheduling and admission control for broadband switching. *IEEE Journal on Selected Areas in Communication* May 1993; **11**(4): 605–616.
16. Zhao GY. Interior-point methods with decomposition for solving large-scale linear programs. *Journal of Optimization Theory and Applications* 1999; **102**(1): 169–192.
17. Biswas S, Sengupta B. Call admissibility for multirate traffic in wireless ATM networks. In *Proceedings of IEEE 1NFOCOM*, Apr 1997; Vol. 2, Kobe, Japan, 649–657.

Authors' Biographies

Nidal Nasser received his B.Sc. and M.Sc. degrees with Honors in Computer Engineering from Kuwait University, Kuwait, in 1996 and 1999, respectively. He obtained his Ph.D. in the School of Computing of Queen's University, Canada, in 2004. In December 2004, he joined the Department of Computing and Information Science at University of Guelph, Ontario, Canada, where he is an Assistant Professor. He has authored several journal publications, refereed conference publications and four book chapters. He has been a member of the technical programme and organizing committees of several international IEEE conferences and workshops. His current research interests include, multimedia wireless cellular networks, wireless and mobile sensor networks and heterogeneous wireless data networks interconnection, with special emphasis on the following topics: radio resource management techniques, performance modelling and analysis and provisioning QoS at connection level, class level, and packet level. Dr. Nasser is a member of the IEEE (Communications Society and Computer Society). He received Fund for Scholarly and Professional Development Award in 2004 from Queen's University.

Hossam Hassanein is a leading researcher in the School of Computing at Queen's University in the areas of broadband, wireless and variable topology networks architecture, protocols, control and performance evaluation. Before joining Queen's University in 1999, he worked at the department of Mathematics and Computer Science at Kuwait University (1993–1999) and the department of Electrical and Computer Engineering at the University of Waterloo (1991–1993). Dr. Hassanein obtained his Ph.D. in Computing Science from the University of Alberta in 1990. He is the founder and director of the Telecommunication Research (TR) Lab http://www.cs.queensu.ca/∼trl in the School of Computing at Queen's. Dr. Hassanein has more than 250 publications in reputable journals, conferences and workshops in the areas of computer networks and performance evaluation. Dr. Hassanein has organized and served on the programme committee of a number of international conferences and workshops. He is a senior member of the IEEE and serves as the Secretary of the IEEE Communication Society Technical Committee on Ad hoc and Sensor Networks (TC AHSN). Dr. Hassanein is the recipient of Communications and Information Technology Ontario (CITO) Champions of Innovation Research award in 2003.

DOI: 10.1002/wcm