Mobile Networks and Applications 11, 101–110, 2006

DOI: 10.1007/s11036-005-4464-2

# Efficient QoS Provisioning for Adaptive Multimedia in Mobile Communication Networks by Reinforcement Learning*

YU FEI†, VINCENT W.S. WONG and VICTOR C.M. LEUNG
*Department of Electrical and Computer Engineering, The University of British Columbia, 2356 Main Mall, Vancouver, BC, Canada, V6T 1Z4*

**Published online:** 9 December 2005

**Abstract.** The scarcity and large fluctuations of link bandwidth in wireless networks have motivated the development of adaptive multimedia services in mobile communication networks, where it is possible to increase or decrease the bandwidth of individual ongoing flows. This paper studies the issues of quality of service (QoS) provisioning in such systems. In particular, call admission control and bandwidth adaptation are formulated as a constrained Markov decision problem. The rapid growth in the number of states and the difficulty in estimating state transition probabilities in practical systems make it very difficult to employ classical methods to find the optimal policy. We present a novel approach that uses a form of discounted reward reinforcement learning known as $Q$-learning to solve QoS provisioning for wireless adaptive multimedia. $Q$-learning does not require the explicit state transition model to solve the Markov decision problem; therefore more general and realistic assumptions can be applied to the underlying system model for this approach than in previous schemes. Moreover, the proposed scheme can efficiently handle the large state space and action set of the wireless adaptive multimedia QoS provisioning problem. Handoff dropping probability and average allocated bandwidth are considered as QoS constraints in our model and can be guaranteed simultaneously. Simulation results demonstrate the effectiveness of the proposed scheme in adaptive multimedia mobile communication networks.

**Keywords:** QoS, adaptive multimedia, mobile communication networks, reinforcement learning

## 1. Introduction

Recent years have witnessed a tremendous growth in the use of mobile communications around the world. With the growing demand for integrated services supporting multimedia such as video and audio in mobile communication systems, quality of service (QoS) provisioning in mobile multimedia networks is becoming increasingly important. Since radio bandwidth is one of the most precious resources in wireless systems, an efficient call admission control (CAC) scheme is very important to guarantee QoS and to maximize radio resource utilization simultaneously. Numerous CAC schemes in mobile communication systems have been proposed in the literature (e.g., [11,25]). Most strategies proposed previously in the literature only consider non-adaptive traffic and non-adaptive networks, which cannot change the bandwidth of ongoing calls. However, in recent years, the scarcity and large fluctuations of link bandwidth in wireless networks have motivated the development of adaptive multimedia services where the bandwidth of a connection can be dynamically adjusted to adapt to the highly variable communication environments. Some examples of these adaptive multimedia services include the International Organization for Standardization's (ISO's) Motion Picture Experts Group (MPEG)-4 [12] and the International Telecommunication Union's (ITU's) H.263 [13], which are expected to be used extensively in future mobile communication networks. Accordingly, future mobile communication networks, e.g., third generation (3G) universal mobile telecommunication systems (UMTS), can provide flexible radio resource management functions. The bandwidth of an ongoing call in these networks can be changed dynamically [9].

QoS provisioning for adaptive multimedia in mobile communication networks requires the use of a bandwidth adaptation (BA) algorithm in conjunction with the CAC algorithm. BA reallocates the bandwidth of ongoing calls, whereas CAC decides whether to admit or reject new and handoff calls.

There are several schemes in the literature addressing CAC and BA for adaptive multimedia services [8,14,15,20,28]. Authors in [20] study the tradeoffs between network overload and fairness in bandwidth adaptation. However, the proposed scheme in [20] does not consider maximizing wireless network utilization and may result in sub-optimal solutions. A near optimal scheme is proposed in [14]. Zaruba et al. [28] use simulated annealing algorithm to find the optimal call-mix selection to maximize the total network revenue under the assumption that future arrivals and departures are known, which may not be realistic in practice. Only one class of adaptive traffic is studied in [8] and [15], and the extension of these schemes to the case of multiple classes may not be an easy task.

In this paper, we formulate the QoS provisioning for adaptive multimedia as a Markov decision process (MDP) to find the optimal CAC and BA algorithms that can maximize

*This work is based in part on a paper presented at BroadNet's 04, San Jose, CA, Oct. 2004.

†Corresponding author.

network revenue and guarantee QoS constraints. Moreover, we propose a scheme using a form of real-time reinforcement learning known as $Q$-learning [23] to solve the MDP. Distinct features of the proposed scheme include the following:

It does not require a priori knowledge of the state transition probabilities associated with the mobile communication networks, which are very difficult to estimate in practice due to the irregular network topology, different propagation environment and random user mobility. Therefore, the assumptions behind the underlying system model can be made more general and realistic than those in previous schemes.

It can efficiently handle problems with large state spaces and action sets. Since there will be several classes of traffic in future mobile multimedia networks and each class of traffic has several bandwidth levels, the state spaces and action sets are very large in the QoS provisioning problem. The proposed scheme can use stochastic approximation to eliminate the need to compute the state transition probabilities and the complex optimization algorithms.

Handoff dropping probability and average allocated bandwidth are considered as QoS constraints in our scheme and can be guaranteed simultaneously. This is in contrast with our previous study [26], which does not take QoS constraint into consideration.

Reinforcement learning has previously been used to solve CAC and routing problems in wireline networks [7,16,21] and channel allocation problem in wireless networks [17,19]. Moreover, there are some attempts to use model-independent and self-learning approaches to solve mobility management problems in mobile networks [6]. However, the primary focuses of [6,7,16,17,19,21] are different from the QoS provisioning problem for adaptive multimedia considered in this paper. Average reward reinforcement learning is used in [27] to solve a similar problem as that in this paper. However, average reward problems are generally more difficult to solve than discounted reward ones. In this paper, we formulate the QoS provisioning problem using the classical discounted reward $Q$-learning algorithm [23], which is simpler in formulation than that in [27].

The rest of this paper is organized as follows. Section 2 describes the QoS provisioning problems. Section 3 gives the problem formulation and our new approach to solve this problem. Section 4 presents and discusses the simulation results to demonstrate the effectiveness of our approach. Finally, we conclude this study in Section 5.

## 2. QoS provisioning for adaptive multimedia

### *2.1. Adaptive multimedia and adaptive mobile communication systems*

Originally, adaptive multimedia applications are introduced in wireline networks. Congestion in wireline broadband networks can cause fluctuations in the availability of network resources, thereby resulting in severe degradation of QoS. To overcome this problem, many techniques are proposed such as the adaptation of compression parameters [22] and layered coding [24]. The much more severe bandwidth fluctuations in mobile wireless networks make it interesting to consider the use of adaptive multimedia in future mobile communication systems.

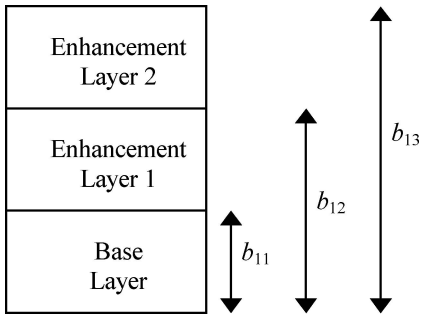



Figure 1. Bandwidth usage of adaptive multimedia.

In our adaptive multimedia framework, a multimedia call can dynamically change its bandwidth to adapt to the fluctuating communication environment throughout its call duration. Assume that there are $K$ classes of services in the network. A class $i$ call uses bandwidth among, $\{b_{i1}, b_{i2}, \ldots, b_{ij}, \ldots, b_{iN_i}\}$ where $b_{ij} < b_{i(j+1)}$ for $i = 1, 2, \ldots, K$, $j = 1, 2, \ldots, N_i$, and $N_i$ is the highest bandwidth level that can be used by a class $i$ call. For example, using the layered coding technique, a raw video sequence is compressed into several layers [24]; say, three layers consisting of a base layer and two enhancement layers. The base layer can be independently decoded and it provides basic video quality; the enhancement layers can only be decoded together with the base layer and they further refine the quality of the base layer. Therefore, a video stream compressed into three layers can adapt to three levels of bandwidth usage. Assume this video stream is class 1 traffic in a mobile communication system. The bandwidth usage of this video stream is shown in figure 1.

Compared to wired networks, the fluctuations in resource availability in mobile communication systems are much more severe. This stems from two inherent characteristics of mobile wireless networks: fading and mobility. The fading in a wireless channel is highly variable with time and spatial dependencies that result in a transmission link with highly varying bandwidth. Moreover, mobile users are free to move from one cell to another one. The availability of resources in the original cell does not necessarily guarantee that the resources are available in new cells. The change in network resources can result in a major fluctuation in the availability of network resources needed to service a call.

Due to the severe fluctuation of resources in wireless mobile networks, the ability of adapting to the communication environment is very important in future mobile communication systems. For example, in UMTS system, a radio bearer established for a call can be dynamically reconfigured during the call session. Figure 2 shows the signaling procedure between user terminal (UE) and universal terrestrial radio access

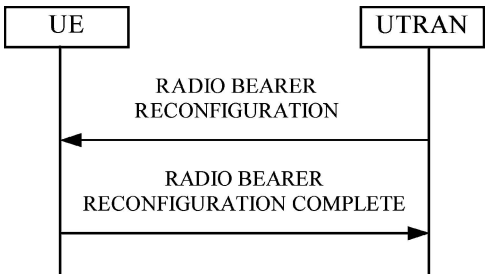



Figure 2. Radio bearer reconfiguration in UMTS [5].

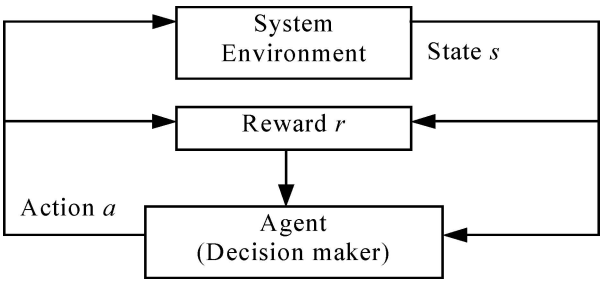



Figure 3. A reinforcement learning model.

network (UTRAN) in radio bearer reconfiguration during a call. Radio bearer information in UMTS includes most of the information in layer 2 and layer 1 for that call, e.g., radio link control (RLC), power control, spreading factor, diversity, etc. By reconfiguring the radio bearer, the bandwidth of a call can be changed dynamically during a call session.

### *2.2. QoS provisioning in adaptive multimedia framework*

The goal of QoS provisioning in the adaptive multimedia framework is to maximize the long-term revenue and guarantee the QoS constraints. We consider two important modules for QoS provisioning: CAC and BA, in this study. When a cell is in an under-loaded condition, CAC tries to accept every call and BA tries to allocate as much bandwidth as possible to each call. However, when network congestion occurs, QoS constraints may be violated. In this case, arriving calls could be rejected by CAC and arriving/existing calls could be degraded to a lower bandwidth by BA. On the other hand, if a call releases its allocated bandwidth due to either call completion or handoff to another cell, some of the calls left in that cell may increase their bandwidth. To decide which call to accept and which call(s) to change the bandwidth are the roles of CAC and BA in the adaptive multimedia framework.

Two QoS constraints are considered in this paper. Since it is generally believed that forced call terminations due to handoff dropping are more objectionable to users than new call blocking, the first QoS constraint in mobile communication networks is to keep $P_{\text{hd}}$, the probability of handoff dropping, below a target level. In addition, although adaptive applications can tolerate decreased bandwidth, it is desirable for some applications to have a bound on the average allocated bandwidth. Therefore, we need another QoS parameter to quantify the average allocated bandwidth. The normalized average allocated bandwidth of class $i$ calls, denoted as $AB^i$, is the ratio of the average bandwidth received by class $i$ calls to the bandwidth with un-degraded service. In order to guarantee the QoS of adaptive multimedia, $AB^i$ should be kept above a target value. These two constraints are formulated in Section 3.

We formulate the QoS provisioning problem as a Markov decision process (MDP) [18]. There are several well-known algorithms, such as policy iteration, value iteration and linear programming [18], which find the optimal solution of a MDP. However, traditional model-based solutions require explicit state transition probabilities and suffer from two "curses": the "curse of dimensionality" and the "curse of modeling". The curse of dimensionality is caused by the fact that the algorithms require computation time that is polynomial in the number of states. QoS provisioning in mobile multimedia networks involve very large state spaces that make traditional solutions infeasible. The curse of modeling occurs in that in order to apply traditional methods, it is first necessary to express state transition probabilities explicitly; however, they are very difficult to estimate in real networks due to the irregular network topology, different propagation environment and random user mobility. Therefore, we choose to solve the problem using a form of real-time reinforcement learning known as $Q$-learning [23]. This method does not require the explicit expression of the state transition probabilities and can handle MDP problems with large state spaces efficiently. The formulation of this method in solving QoS provisioning for adaptive multimedia is presented in Section 3.

## 3. Solving QoS provisioning for adaptive multimedia by $Q$-learning

### *3.1. Q-learning algorithm*

In recent years, reinforcement learning (RL) has become a topic of intensive research. Reinforcement learning is a way of teaching agents (decision makers) optimal policies by assigning rewards and punishments for their actions based on the temporal feedback obtained during active interactions of the agent with the system environment. In the RL model depicted in figure 3, a learning agent selects an action for the system that leads the system along a unique path till another decision-making state is encountered. At this time, the system needs to consult with the learning agent for the next state. During a state transition, the agent gathers information about the new state, immediate reward and the time spent during the state-transition, based on which the agent updates its knowledge base using an algorithm and selects the next action. The process is repeated and the learning agent continues to improve its performance.

Reinforcement learning combines concepts from dynamic programming, stochastic approximation via simulation and

function approximation. This method has two distinct advantages over model-based methods. The first is that it can handle problems with complex transitions by making use of stochastic approximation, thereby eliminating the need to compute the transition probabilities and the complex optimization algorithms. Secondly, RL can integrate within it various function approximation methods (e.g., neural networks), which can be used to approximate the value function even when the size of the state space is gargantuan. $Q$-learning [23] is one of the most popular reinforcement learning algorithms. We use this algorithm to solve the QoS provisioning problem in this paper.

Assume that the environment is a finite-state discrete-time stochastic dynamic system. Let $S = \{s_1, s_2, \ldots, s_n\}$ be the set of possible states, and $A = \{a_1, a_2, \ldots, a_m\}$ be a set of possible actions, Based on the state $s_t \in S$, the agent interacting with the environment chooses an action $a_t \in A$ to perform. Then the environment makes a transition to the new state $s_{t+1} = s' \in S$ according to probability $P_{ss'}(a)$ and gives a reward $r_t$ to the agent. The process is repeated.

The goal of the agent is to find an optimal policy $\pi^*(s) \in A$ for each $s$, which maximizes some cumulative measure of the rewards received over time. The total expected discounted reward over an infinite time horizon is:

$$V^{\pi}(s) = E\left\{\sum_{t=0}^{\infty} \gamma^t r(s_t, \pi(s_t)) | s_0 = s\right\}, \tag{1}$$

where $0 \leq \gamma < 1$ is a discount factor and $E$ denotes the expectation. Equation (1) can be rewritten as:

$$V^{\pi}(s) = R(s, \pi(s)) + \gamma \sum_{s' \in S} P_{ss'}(\pi(s)) V^{\pi}(s'), \tag{2}$$

where $R(s, \pi(s)) = E\{r(s, \pi(s))\}$ is the mean value of reward $r(s, \pi(s))$. The optimal policy $\pi^*$ satisfies the optimality criterion:

$$V^*(s) = V^{\pi^*}(s) = \max_{a \in A}\left(R(s, a) + \gamma \sum_{s' \in S} P_{ss'}(a) V^*(s')\right). \tag{3}$$

However, it is difficult to get $R(s, a)$ and $P_{ss'}(a)$ in many practical situations such as the QoS provisioning problem in this paper. $Q$-learning is one of the most popular and effective algorithms for learning from delayed reinforcement to determine the optimal policy. For a policy $\pi$, define a $Q$ value as:

$$Q^{\pi}(s, a) = R(s, a) + \gamma \sum_{s' \in S} P_{ss'}(a) V^{\pi}(s') \tag{4}$$

which is the expected discounted reward for executing action $a$ at state $s$ and then following policy $\pi$ thereafter. Let

$$Q^*(s, a) = Q^{\pi^*}(s, a) = R(s, a) + \gamma \sum_{s' \in S} P_{ss'}(a) V^{\pi^*}(s') \tag{5}$$

So $V^*(s) = \max_{a \in A} Q^*(s, a)$. $Q^*(s, a)$ can therefore be written recursively as:

$$Q^*(s, a) = R(s, a) + \gamma \sum_{s' \in S} P_{ss'}(a) \max_{a' \in A}(Q^*(s', a')) \tag{6}$$

Then, we have $\pi^*(s) = \arg\max_{a \in A}(Q^*(s, a))$ as an optimal policy. The $Q$-learning process tries to find $Q^*(s, a)$ in a recursive manner using available information. The $Q$-learning rule is

$$Q_{t+1}(s, a) = \begin{cases} Q_t(s, a) + \alpha \Delta Q_t(s, a) & \text{if } s = s_t \text{ and } a = a_t \\ Q_t(s, a) & \text{otherwise} \end{cases} \tag{7}$$

where $\Delta Q_t(s, a) = r_t + \gamma \max_{a' \in A} Q_t(s', a') - Q_t(s, a)$ and $\alpha$ is the learning rate.

### 3.2. Q-learning formulation to solve QoS provisioning

In solving the QoS provisioning problem, the mobile communication system can be considered as a discrete-time event system. These events are modeled as stochastic variables with appropriate probability distributions. We assume that call arrivals including new call arrival events and handoff events follow Poisson distributions. Call holding time is assumed to be exponentially distributed. The call arrival distribution and the service time distribution are independent of each other. In order to utilize the $Q$-learning algorithm, we need to identify the system states, actions, rewards, and constraints.

#### 3.2.1. System states

An event $e$ can occur in a cell $c$, where $e$ is either a new call arrival, a handoff call arrival, a call termination, or a call handoff to a neighboring cell. At this time, cell $c$ is in a particular configuration $x$ defined by the number of each type of ongoing calls in the cell; $x = (x_{11}, x_{12}, \ldots x_{ij}, \ldots x_{KN_K})$, where $x_{ij}$ denotes the number of ongoing calls of class $i$ using bandwidth $b_{ij}$ in cell c for $1 \leq i \leq K$ and $1 \leq j \leq N_i$. Recall that $K$ is the number of service classes in the system and $N_i$ is the highest bandwidth level of class $i$ defined in Section 2. The configuration and event together determine the state, $s = (x, e)$, of cell $c$.

We assume that each cell has a fixed channel capacity $C$. The state space is defined as:

$$S = \left\{s = (\mathbf{x}, e) : \sum_{i=1}^{K} \sum_{j=1}^{N_i} x_{ij} b_{ij} \leq C\right\}. \tag{8}$$

*3.2.2. Actions*

When an event occurs, the agent must choose an action according to the state. An action can be denoted as: $a = (a_a, a_d, a_u)$, where $a_a$ stands for the admission decision, i.e., admit ($a_a = 1$), reject ($a_a = 0$) or no action due to call departure ($a_a = -1$), ad stands for the action of bandwidth degradation when a call is accepted and $a_u$ stands for the action of bandwidth upgrade when there is a departure (call termination or handoff to a neighboring cell) from cell $c$. $a_d$ has the form

$$a_d = \left\{\left(d_{12}^1, \ldots, d_{ij}^n, \ldots, d_{KN_k}^{N_{k-1}}\right), 1 \le i \le K, \right. \\ \left. 1 < j \le N_i, 1 \le n < j\right\},$$

where $d_{ij}^n$ denotes the number of ongoing class $i$ calls using bandwidth $b_{ij}$ that are degraded to bandwidth $b_{in}$. $a_u$ has the form

$$a_u = \left\{\left(u_{11}^2, \ldots, u_{ij}^n, \ldots, u_{KN_k-1}^{N_k}\right), 1 \le i \le K, \right. \\ \left. 1 \le j < N_i, j < n \le N_i\right\},$$

where $u_{ij}^n$ denotes the number of ongoing class $i$ calls using bandwidth $b_{ij}$ that are upgraded to bandwidth $b_{in}$.

After the action of degradation, the configuration $(x_{11}, x_{12}, \ldots, x_{ij}, \ldots, x_{KN_K})$ becomes

$$\left(x_{11} + \sum_{m=2}^{N_1} d_{1m}^1, x_{12} + \sum_{m=3}^{N_1} d_{1m}^2 - d_{12}^1, \ldots, x_{ij} \right. \\ \left. + \sum_{m=j+1}^{N_i} d_{im}^j - \sum_{m=1}^{j-1} d_{ij}^m, \ldots, x_{KN_K} - \sum_{m=1}^{N_K-1} d_{KN_K}^m\right).$$

Similarly, after the action of upgrading bandwidths, the configuration $(x_{11}, x_{12}, \ldots, x_{ij}, \ldots, x_{KN_K})$ becomes

$$\left(x_{11} - \sum_{m=2}^{N_1} u_{11}^m, x_{12} + u_{11}^2 - \sum_{m=3}^{N_1} u_{12}^m, \ldots, x_{ij} \right. \\ \left. + \sum_{m=1}^{j-1} u_{im}^j - \sum_{m=j+1}^{N_i} u_{ij}^m, \ldots, x_{KN_K} + \sum_{m=1}^{N_K-1} u_{Km}^{N_K}\right)$$

*3.2.3. Rewards*

Based on the action taken in a state, the system earns deterministic revenue due to the carried traffic in the cell. Let $r_{ij}$ be the reward rate of class $i$ call using bandwidth $b_{ij}$. The reward rate, $r(s,a)$ , can be calculated as:

$$r(s, a) = \sum_{i=1}^{K} \sum_{j=1}^{N_i} x_{ij} r_{ij}. \tag{9}$$

Note that all ongoing calls in the cell, including those that have been degraded and upgraded, contribute to the reward. Therefore, we do not need an extra term to formulate the penalty related to the bandwidth degradation.

*3.2.4. Constraints*

For a general MDP with $L$ constraints, the optimal policy for at most $L$ of the states is randomized [1]. Since $L$ is much smaller than the total number of states in the QoS provisioning problem considered in this paper, the non-randomized stationary policy learned by reinforcement learning is often a good approximation to the optimal policy [10]. To avoid the complications of randomization, we concentrate on non-randomized policies in this study.

The first QoS constraint is related to the handoff dropping probability. Upon the $n$th decision epoch, the measured handoff dropping ratio, $P_{\text{hd}}(s_n)$, should be kept below a target value. Let $TP_{\text{hd}}$ denote the target maximum allowed handoff dropping probability. The constraint associated with $P_{\text{hd}}$ can be formulated as:

$$\lim_{N\to\infty} \frac{\sum_{n=0}^{N} P_{\text{hd}}(s_n)\tau_n}{\sum_{n=0}^{N} \tau_n} \le TP_{\text{hd}}, \tag{10}$$

where $\tau_{\text{n}}$ is the time interval between decision epochs.

The Lagrange multiplier formulation relating the constrained optimization to an unconstrained optimization [4,5] is used in this paper to deal with the handoff dropping constraint. To fit into this formulation, we need to include the history information in our state descriptor. The new state descriptor is $\bar{s} = (N_{\text{hr}}, N_{\text{hd}}, \tau, s)$, where $N_{\text{hr}}$ and $N_{\text{hd}}$ are the total number of handoff call requests and handoff call drops, respectively, from each class, $\tau$ is the time interval between the last and the current decision epochs, and $s$ is the original state descriptor. In order to make the state space finite, quantified values of $P_{\text{hd}} = N_{\text{hd}}/N_{\text{hr}}$ and $\tau$ are used in the state aggregation approximation in the following subsection.

A Lagrange multiplier $\omega$ is used for the parameterized reward $\bar{r}(\bar{s}', \bar{s}, a) = r(\bar{s}', \bar{s}, a) - \omega\, z(\bar{s}', \bar{s}, a)$, where $r(\bar{s}', \bar{s}, a)$ is the original reward function and $z(\bar{s}', \bar{s}, a) = P_{\text{hd}}(\bar{s})\tau(\bar{s}', \bar{s}, a)$ is the cost function associated with the constraint. The multiplier $\omega$ is chosen so that the constraint is met in a fashion consistent with the desired optimization. A nice monotonicity property associated with $\omega$ shown in [4] facilitates the search for a suitable $\omega$.

The second QoS constraint is related to $AB^i$, the normalized average allocated bandwidth of class $i$ calls. Let $B_i$ denote the bandwidth allocated to class $i$ calls. $AB^i$ can be defined as the mean of $B^i/b_{iN_i}$ over all class $i$ calls in the current cell. Recall that $b_{iN_i}$ is the bandwidth of a class $i$ call with un-degraded service.

$$AB^i = E\left\{\frac{B^i}{b_{iN_i}}\right\} = \frac{E\{B^i\}}{b_{iN_i}} = \frac{\sum_{j=1}^{N_i} x_{ij} b_{ij}}{b_{iN_i} \sum_{j=1}^{N_i} x_{ij}}, \\ i = 1, \ldots, K.$$

$AB^i$ should be kept larger than the target value $TAB^i$:

$$AB^i \ge TAB^i, \quad i = 1, \ldots, K. \tag{11}$$

$AB^i$ is an intrinsic property of a state. With the current state and action information $(\bar{s}, a)$ we can forecast $AB^i$ in the next state $\bar{s}'$, $AB^i(\bar{s}')$. If $AB^i(\bar{s}') \le TAB^i$, $i = 1, \ldots, K$, the action

is feasible. Otherwise, this action should be eliminated from the feasible action set $A(\bar{s})$.

It is also interesting to consider the constraint of time-averaged allocated bandwidth of a call throughout its lifetime. Assume that the call holding time is $H$ and $B^i(t)$ is the bandwidth allocated to a class $i$ call at time $t$. The normalized mean bandwidth allocated to this call while in the system is:

$$\hat{B}^i = \frac{\int_0^H \frac{B^i(t)}{b_{iN_i}} dt}{H}.$$

Then $\bar{B}^i = E\{\hat{B}^i\}$ is a measure of the performance as far as time-averaged allocated bandwidth is concerned.

A very nice result in [2] shows that the average bandwidth allocated to a call during its lifetime is equal to the expected bandwidth allocated to the call at the moment it arrives to the system and independent of its holding time, i.e., $AB = \bar{B}$, in a system with one class of traffic. For the above reasons and to avoid any complications, we use equation (11) as the normalized average allocated bandwidth constraint.

#### 3.2.5. *Trading off action space complexity with state space complexity*

We can see that the action space in our formulation is quite large. It will be time-consuming to find the suitable action given a specific state using RL. We propose a method to trade off action space complexity with state space complexity in the QoS provisioning scheme using a scheme described in [3]. The advantages of doing this are that the action space will be reduced and the extra state space complexity may still be dealt with by using the function approximation.

Suppose that a call arrival event occurs in a cell with state $s$, the action that can be chosen from is $a = (a_a, d_{12}^1, \ldots, d_{ij}^n, \ldots, d_{KN_k}^{N_k-1})$, where there are at most $W = 1 + \sum_{i=1}^{K} \sum_{j=2}^{N_i} (j-1)$ components. We can break down the action $a$ into a sequence of $W$ controls $a_a, d_{12}^1, \ldots, d_{ij}^n, \ldots, d_{KN_k}^{N_k-1}$, and introduce some artificial intermediate "states" $(\bar{s}, a_a)$, $(\bar{s}, a_a, d_{12}^1), \ldots, (\bar{s}, a_a, d_{12}^1, \ldots, d_{ij}^n, \ldots, d_{KN_k}^{N_k-1})$, and the corresponding transitions to model the effect of these actions. In this way, the action space is simplified at the expense of introducing $W-1$ additional layers of states and $W-1$ additional $Q$ values $Q(\bar{s}, a_a)$, $Q(\bar{s}, a_a, d_{12}^1), \ldots, Q(\bar{s}, a_a, d_{12}^1, \ldots, d_{ij}^n, \ldots, d_{KN_k}^{N_k-2})$ in addition to $Q(\bar{s}, a_a, d_{12}^1, \ldots, d_{ij}^n, \ldots, d_{KN_k}^{N_k-1})$. Actually, we view the problem as a deterministic dynamic programming problem with $W$ stages. For $w = 1, \ldots, W$, we can have a $W$-solution (a partial solution involving just $w$ components) for the $w$th stage of the problem. The terminal state corresponds to the $W$-solution (a complete solution with $W$ components). Moreover, instead of selecting the controls in a fixed order, it is possible to leave the order subject to choice.

In the reformulated problem, at any given state $\bar{s} = (N_{\text{hr}}, N_{\text{hd}}, \tau x, e)$ where $e$ is a call arrival of class $i$, the control choices are:

(1) Reject the call, in which case the configuration $x$ does not evolve.

(2) Admit the call and no bandwidth adaptation is needed, in which case the configuration $x$ evolves to $(x_{11}, x_{12}, \ldots, x_{ij}, \ldots, x_{iN_i} + 1, \ldots, x_{KN_K})$.

(3) Admit the call and bandwidth adaptation is needed. In this case, the problem can be divided into $W$ stages. On the $w$th stage ($w = 1, \ldots, W$), one particular call type that has not been selected in previous stages, say the one using bandwidth $b_{ij}$ with $x_{ij} > 0$, can be selected and there are following options:

   (a) Degrade one call using bandwidth $b_{ij}$ one level, in which case the configuration $x$ evolves to $(x_{11}, x_{12}, \ldots, x_{ij-1} + 1, x_{ij} - 1, \ldots, x_{iN_i} + 1, \ldots, x_{KN_K})$.

   (b) Degrade two calls using bandwidth one level, in which case the configuration $x$ evolves to $(x_{11}, x_{12}, \ldots, x_{ij-2} + 2, x_{ij} - 2, \ldots, x_{iN_i} + 1, x_{KN_K})$.

   (c) Increase the number of calls being degraded until the call arrival can be accommodated. Please note that the number of options depends on specific selected call type and the class of call arrival.

The similar trade-off can be done when a call departure event occurs.

### 3.3. *Implementation considerations*

To guarantee the convergence of the $Q$-learning algorithm, each action should be executed in each state an infinite number of times. Therefore, with a small probability $p_n$, upon the $n$th decision-making epoch, a decision other than the highest $Q$ value is taken. This is called exploration [3].

In practice, an important issue is how to store the $Q$ values in the $Q$-learning algorithm. There are several approaches to representing the $Q$ values, among which the lookup table is the most straightforward method. A lookup table representation means that a separate variable $Q(s, a)$ is kept in memory for each state-action pair $(s, a)$. Obviously, when the number of state-action pairs becomes large, the lookup table representation will be infeasible, and some compact representation method is necessary. In this paper, we use the state aggregation approximation method [3]. In this method, the state space $S$ is partitioned into $G$ disjoint sub-states $S_0, S_1, \ldots, S_G$. The $Q$ value function for all state $s \in S_g$ under action a is a constant $\phi(g, a)$ such that

$$\tilde{Q}(s, a, \phi) = \phi(g, a), \quad \text{if } s \in S_g.$$

Then a lookup table can be used for the aggregated problem.

The process of the $Q$-learning-based QoS provisioning is shown in figure 4. First, when an event (either a call arrival or a call departure) occurs, a state $s$ can be identified by getting the status of the local cell. Then, a set of feasible actions $\{a\}$ can be found. Second, look up the aggregated

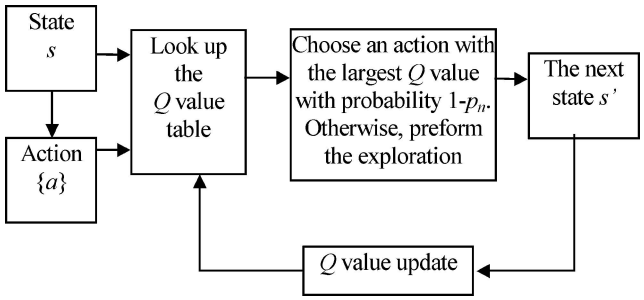



Figure 4. The process of $Q$-learning-based QoS provisioning scheme.

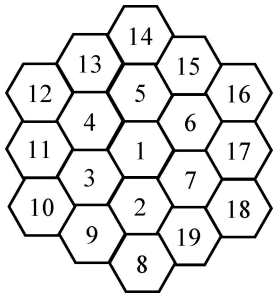



Figure 5. The cellular network used in simulations.

$Q$ value table and find a set of $Q$ values corresponding to the state $s$ and action set $\{a\}$. Third, with probability $1 - p_n$, choose one action from set $\{a\}$ with the maximum $Q$ value; otherwise, perform the exploration. According to the chosen action, the network makes the admission decision and resource adaptation. Fourth, another event occurs and the system reaches another state $s'$. Finally, the $Q$ value is updated according to equation (7).

## 4. Simulation results and discussions

We use a cellular network of 19 cells in simulations, as shown in figure 5. To avoid the edge effect of the finite network size, wrap-around is applied to the edge cells so that each cell has six neighbors. Each cell has a fixed bandwidth of 2 Mbps. Two classes of flows are considered (see Table 1). Class 1 traffic has three different bandwidth levels, 128, 192 and 256 Kbps. 64, 96 and 128 Kbps are the three possible bandwidth levels of class 2 traffic. The reward generated by a call is a linear growing function of the bandwidth assigned to the call. Specifically, $r_{ij} = b_{ij}$. We assume that the highest possible bandwidth level is requested at the time of call arrival. That is, call arrivals of class 1 always request 256 Kbps and call arrivals of class 2 always request 128 Kbps. Then the network will make the CAC decision and decide which bandwidth level the call can use if it is admitted. 30% of the offered traffic is from class 1. Moreover, call holding time and cell residence time are assumed to follow exponential distributions with mean 180 seconds and 150 seconds, respectively.

Table I
Experimental parameters.

| Traffic class | Bandwidth level (Kbps) | Reward |
|---|---|---|
| Class 1 | $b_{11}$:128 | 128 |
| | $b_{12}$:192 | 192 |
| | $b_{13}$:256 | 256 |
| Class 2 | $b_{21}$:64 | 64 |
| | $b_{22}$:96 | 96 |
| | $b_{23}$:128 | 128 |

The mobile communication system can be simulated as a discrete-time event system. A list of future events is maintained dynamically and a simulation clock is advanced according to the future events. The proposed scheme is trained and the $Q$ values are learnt by running the simulation for 20 million steps with a constant arrival rate of 0.1 call/second. The discount factor $\gamma$ is chosen to be 0.5, and the learning rate $\gamma$ varies with the state-action over time as follows. Each state-action is associated with a learning rate that is inversely proportional to the frequency of the state-action being visited up to the present time. The probability of a user handing off to any one of the six adjacent cells is equally likely. The monotonicity property associated with $\omega$ is used to search for a suitable $\omega$, which is 157560 in the simulations. The target maximum allowed handoff dropping probability, $TP_{\text{hd}}$, is $1\%.P_{\text{hd}}$ is quantified into 100 levels. $\tau$ is quantized into 2 levels, i.e., $\tau$ less than or equal to the average inter-decision time and $\tau$ greater than the average inter-decision time. The average allocated bandwidth constraint is changed for evaluation purpose.

Two other schemes are used for comparisons, the guard channel (GC) scheme [11] and TBA98 [20]. In the GC scheme, a set of bandwidth is reserved permanently for handoff calls. In our simulations, 256 Kbps is reserved for handoff calls. In TBA98, the average bandwidth of currently active flows is used to determine which calls should be increased or decreased in the BA operation. In order to perform fair comparisons, new calls may be admitted to an overloaded cell in TBA98 scheme in our simulations.

Figure 6 shows the performance improvement during the training phase. The reward is normalized by the GC scheme

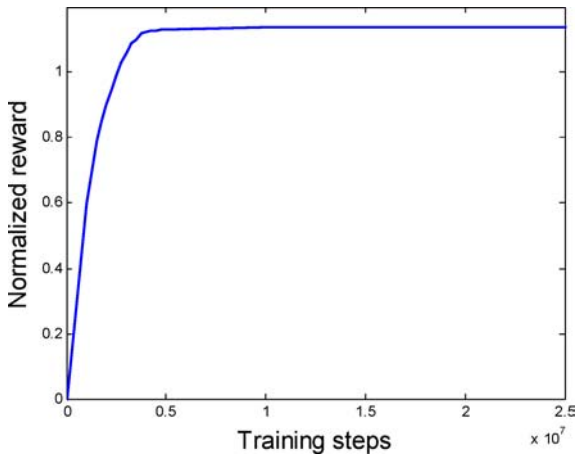



Figure 6. The learning curve.

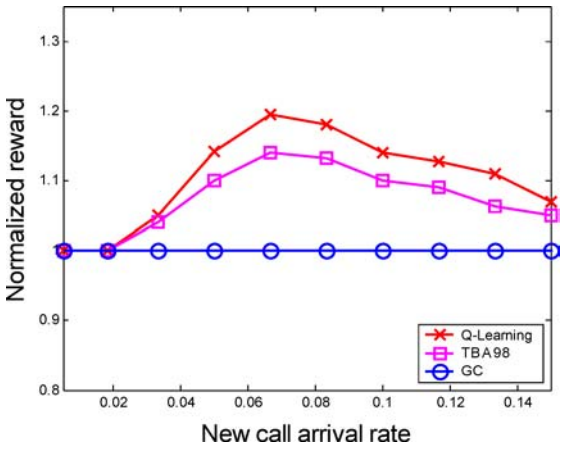



Figure 7. Normalized rewards.

with new call arrival rate of 0.1 calls/second. We obtain the final control policy after 20 million steps. Figure 7 shows the rewards of different schemes. Average allocated bandwidth constraints are not considered here. It is clear the $Q$-learning-based scheme yields more reward than the TBA98 or GC schemes. The traditional GC scheme does not use bandwidth adaptation and a call will be rejected if no free bandwidth is available. TBA98 has bandwidth adaptation function and therefore can gain more reward than GC. However, TBA98 does not consider the problem of maximizing the reward. That is why it receives less reward than the proposed scheme. We can also see from figure 7 that at low traffic loads, as the new call arrival rate increases, the gain becomes more significant. This is because the heavier the traffic load, the more the bandwidth adaptation is needed when the cell is not saturated. However, when the traffic is high and the cell is becoming saturated, the performance gain of the proposed scheme and TBA98 over GC is less significant.

Figure 8 plots the handoff dropping probability vs. new call arrival rate. We can see that the $Q$-learning-based scheme can keep the handoff dropping probability below the target value regardless of the offered load. Although GC and TBA98 can keep it below the target value when the traffic is low, they fail to do so when the traffic is high. We can reduce the handoff dropping probability in GC scheme by increasing the number of guard channel and in TBA98 scheme by rejecting new calls in an overloaded cell. However, this will further reduce the reward earned in these two schemes. Figures 9 and 10 show the new call blocking probabilities of class 1 and class 2 calls, respectively. Both TBA98 and the proposed scheme have less new call blocking probabilities compared with GC, because both of them have adaptation capabilities and can accept more new calls.

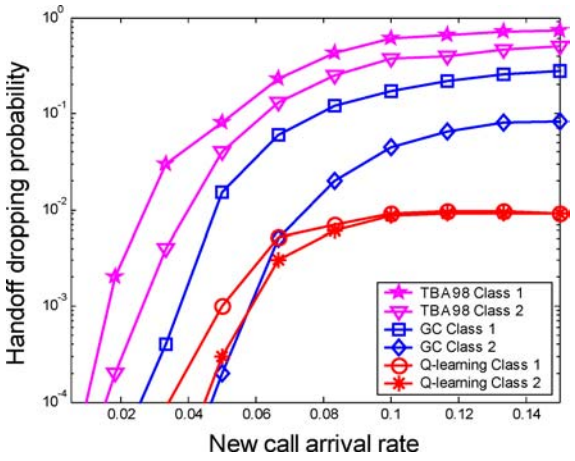



Figure 8. Handoff dropping probabilities.

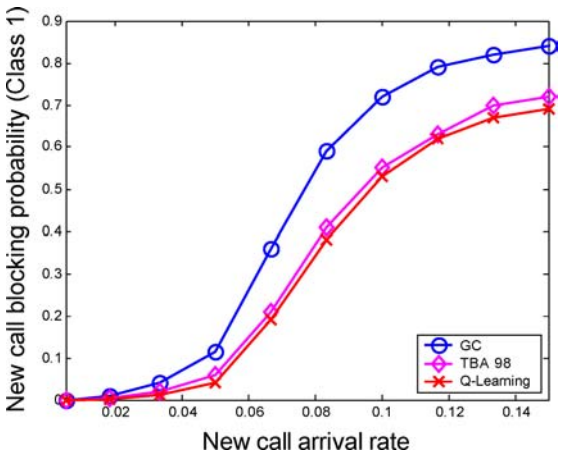



Figure 9. New call blocking probabilities of class 1 calls.

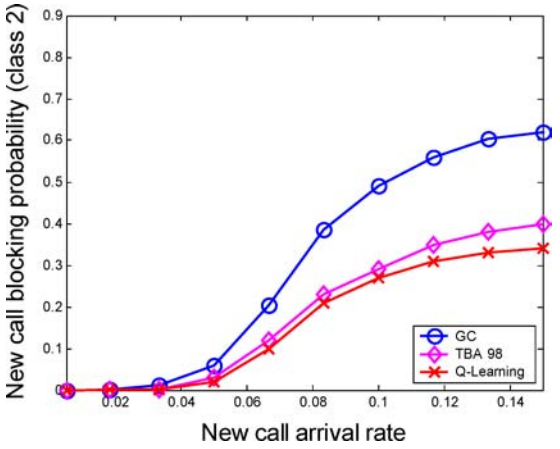



Figure 10. New call blocking probabilities of class 2 calls.

Figures 11 and 12 show the normalized average allocated bandwidth of class 1 and class 2 traffic, respectively, with a target normalized average bandwidth value of 0.7. We observe that as the new call arrival rate increases, the average bandwidths in both TBA98 and the proposed scheme decrease. This is the result of the bandwidth adaptation. It is shown that the normalized average allocated bandwidth can be bounded by the target value in the proposed scheme. In contrast, TBA98 cannot guarantee this average bandwidth QoS constraint. The average bandwidth of GC is always 1, because no adaptation operation is done in GC. The achieve-

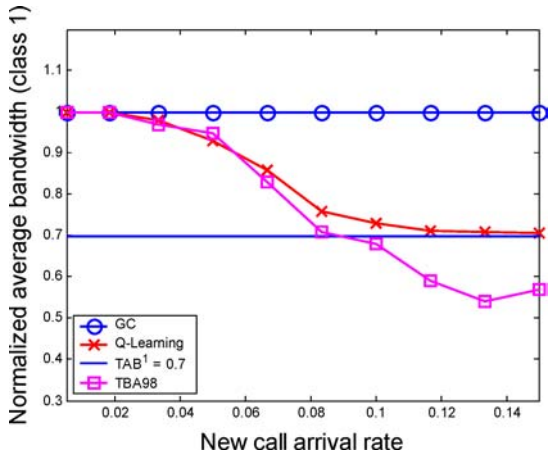



Figure 11. Normalized average bandwidths of class 1 calls.

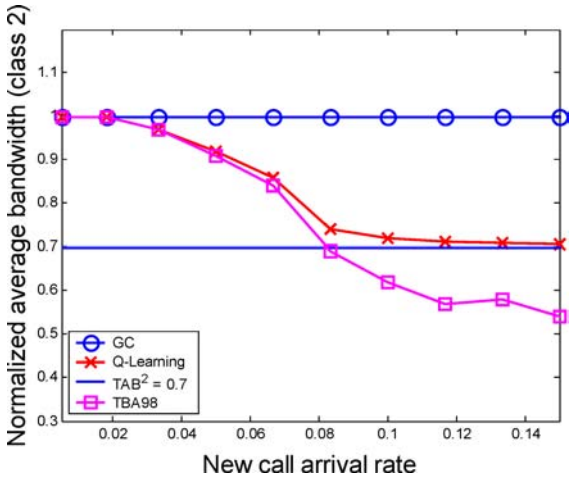



Figure 12. Normalized average bandwidths of class 2 calls.

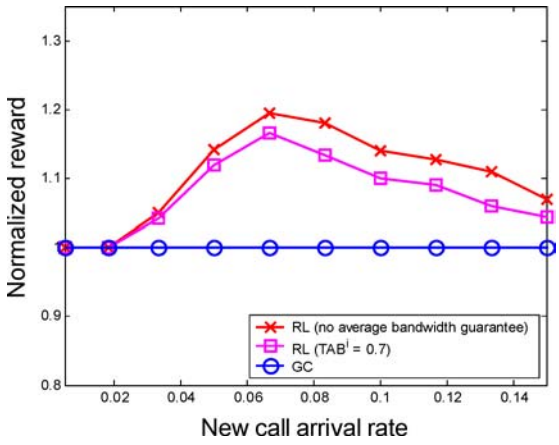



Figure 13. Normalized rewards for different average bandwidth requirements.

ments of QoS guarantee come at a cost to the system. The effects of different values of TAB on the reward are shown in figure 13. We can see that a higher TAB, which is preferred from users' point of view, will reduce the reward.

## 5. Conclusions

In this paper, we have formulated the QoS provisioning problem for adaptive multimedia in mobile communication systems as a constrained MDP to find the optimal CAC and BA policies that can maximize network revenue and guarantee QoS constraints. We have further proposed a scheme using $Q$-learning to solve the QoS provisioning problem. This scheme does not require a priori knowledge of the state transition probabilities associated with the mobile communication systems and can efficiently handle problems with large state spaces and action sets. Two important QoS constraints, i.e., handoff dropping probability constraint and average allocated bandwidth constraint, have been considered. The performance of the proposed scheme has been evaluated by simulations. We have presented numerical results to show that the proposed scheme employing $Q$-learning outperforms existing schemes [11,20].

Future work includes studying other function approximations, such as neural networks, to approximate $Q$ values. It is also very interesting to perform more experiments to compare our scheme with a new scheme recently proposed in [28].

## Acknowledgments

This work was supported in part by Canadian Natural Sciences and Engineering Research Council under grant RGPIN 262604-03 and 44286-00.

## References

[1] E. Altman, *Constrained Markov Decision Process* (Chapman and Hall, London, 1999).

[2] N. Argiriou and L. Georgiadis, Channel sharing by rate-adaptive streaming applications, in: *Proc. IEEE Infocom'02* (June 2002).

[3] D.P. Bertsekas and J.N. Tsitsiklis, *Neuro-Dynamic Programming* (Athena Scientific, 1996).

[4] F.J. Beutler and K.W. Ross, Optimal policies for controlled Markov chains with a constraint, J. Math. Anal. Appl. 112 (1985) 236–252.

[5] F.J. Beutler and K.W. Ross, Time-average optimal constrained semi-Markov decision processes, Adv. Appl. Prob. 18 (1986) 341–359.

[6] A. Bhattacharya and S.K. Das, LeZi-update: An information-theoretic framework for personal mobility tracking in PCS networks, Wireless Networks 8(2/3) (2002) 121–135.

[7] J.A. Boyan and M.L. Littman, Packet routing in dynamically changing networks: A reinforcement learning approach, in: *Advances in NIPS 6*, J.D. Cowan et al. (eds.) (1994) pp. 671–678.

[8] C. Chou and K.G. Shin, Analysis of combined adaptive bandwidth allocation and admission control in wireless networks, in: *Proc. IEEE Infocom'02* (June 2002).

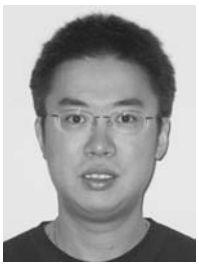
[9] 3GPP, RRC protocol specification, 3G TS25.331 version 3.20.0 (Sept. 2004).

[10] Z. Gabor, Z. Kalmar and C. Szepesvari, Multi-criteria reinforcement learning, in: *Proc. Int'l Conf. Machine Learning*, Madison, WI (July 1998).

[11] D. Hong and S.S. Rappaport, Traffic model and performance analysis for cellular mobile radio telephone systems with prioritised and non-prioritised handoff procedures, IEEE Trans. Veh. Technol. VT-35 (1986) 77–92.

[12] ISO/IEC 144962-2, *Information Technology Coding of Audio-Visual Objects: Visual* (Committee draft, Oct. 1997).
[13] ITU-T H. 263, *Video Coding for Low Bitrate Communication* (Jan. 1998).
[14] T. Kwon, J. Choi, Y. Choi and S.K. Das, Near optimal bandwidth adaptation algorithm for adaptive multimedia services in: Wireless/Mobile Networks, in: *Proc. IEEE VTC'99-Fall*, vol. 2, Amsterdam, The Netherland (Sept. 1999) pp. 874–878.
[15] T. Kwon, Y. Choi, C. Bisdikian and M. Naghshineh, QoS provisioning in wireless/mobile multimedia networks using an adaptive framework, Wireless Networks 9 (2003) 51–59.
[16] P. Marbach, O. Mihatsch and J.N. Tsitsiklis, Call admission control and routing in integrated services networks using neuro-dynamic programming, IEEE J. Select. Areas Commun. 18(2) (2000) 197–208.
[17] J. Nie and S. Haykin, A *Q*-learning based dynamic channel assignment technique for mobile communication systems, IEEE Trans. Veh. Technol. 48(5) (1999) 1676–1687.
[18] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming* (Wiley, New York, 1994).
[19] S.P. Singh and D.P. Bertsekas, Reinforcement learning for dynamic channel allocation in cellular telephone systems, in: *Advances in NIPS* Vol. 9, M. Mozer et al. (eds.) (1997) pp. 974–980.
[20] A.K. Talukdar, B.R. Badrinath and A. Acharya, Rate adaptation schemes in networks with mobile hosts, in: *Proc. ACM/IEEE MobiCom'98* (Oct. 1998).
[21] H. Tong and T.X. Brown, Adaptive call admission control under quality of service constraints: a reinforcement learning solution, IEEE J. Select. Areas Commun. 18(2) (2000) 209–221.
[22] D. Taubman and A. Zakhor, A common framework for rate and distortion based scaling of highly scalable compressed video, IEEE Trans. Circuits Syst. Video Technol. 6(4) (1996) 329–354.
[23] C.J.C.H. Watkins and P. Dayan, *Q*-learning, Machine Learning 8 (1992) 279–292.
[24] D. Wu, Y.T. Hou and Y.Q. Zhang, Scalable video coding and transport over broadband wireless networks, Proc. IEEE 89(1) (2001) 6–20.
[25] S. Wu et al., A dynamic call admission policy with precision QoS guarantee using stochastic control for mobile wireless networks, IEEE/ACM Trans. Networking 10(2) (2002) 257–271.
[26] F. Yu, V.W.S. Wong and V.C.M. Leung, Reinforcement learning for call admission control and bandwidth adaptation in mobile multimedia networks, in: *Proc. of ICICS-PCM'3*, Singapore (Dec. 2003).
[27] F. Yu, V.W.S. Wong and V.C.M. Leung, A new QoS provisioning method for adaptive multimedia in cellular wireless networks, in: *Proc. IEEE Infocom'04*, HongKong, China, (Apr. 2004).
[28] G.V. Zaruba, I. Chlamtac and S.K. Das, A prioritized real-time wireless call degradation framework for optimal call mix selection, Mobile Networks and Applications 7 (2002) 143–151.

**Fei Yu** received the M.S. degree in Computer Engineering from Beijing University of Posts and Telecommunications, P.R. China, in 1998, and the Ph.D. degree in Electrical Engineering from the University of British Columbia (UBC), Canada, in 2003. From 1998 to 1999, Dr. Yu was a system engineer at China Telecom, P.R. China, working on the planning, design and performance analysis of national SS7 and GSM networks. From 2002 to 2004, He was a research and development engineer at Ericsson Mobile Platforms, Sweden, where he worked on dual-mode UMTS/GPRS handsets. He is currently a postdoctoral research fellow at UBC. His research interests are quality of service, cross-layer design and mobility management in wireless networks.
E-mail: feiy@ece.ubc.ca



**Vincent W.S. Wong** (S'94-M'00) received the B.Sc. (with distinction) degree from the University of Manitoba, Winnipeg, MB, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2000, all in electrical engineering. From 2000 to 2001, he was a Systems Engineer at PMC-Sierra, Inc., Burnaby, BC. Since 2002, he has been with the Department of Electrical and Computer Engineering, UBC, where he is currently an Assistant Professor. His research interests are in wireless communications and networking. Dr. Wong received the Natural Science and Engineering Research Council (NSERC) postgraduate scholarship and the Fessenden Postgraduate Scholarship from Communications Research Centre, Industry Canada, during his graduate studies.
E-mail: vincentw@ece.ubc.ca

**Victor C.M. Leung** received the B.A.Sc. (Hons.) degree in electrical engineering from the University of British Columbia (U.B.C.) in 1977, and was awarded the APEBC Gold Medal as the head of the graduating class in the Faculty of Applied Science. He attended graduate school at U.B.C. on a Natural Sciences and Engineering Research Council Postgraduate Scholarship and obtained the Ph.D. degree in electrical engineering in 1981.

From 1981 to 1987, Dr. Leung was a Senior Member of Technical Staff at Microtel Pacific Research Ltd. (later renamed MPR Teltech Ltd.), specializing in the planning, design and analysis of satellite communication systems. He also held a part-time position as Visiting Assistant Professor at Simon Fraser University in 1986 and 1987. In 1988, he was a Lecturer in the Department of Electronics at the Chinese University of Hong Kong. He joined the Department of Electrical Engineering at U.B.C. in 1989, where he is a Professor, Associate Head of Graduate Affairs, holder of the TELUS Mobility Industrial Research Chair in Advanced Telecommunications Engineering, and a member of the Institute for Computing, Information and Cognitive Systems. His research interests are in the areas of architectural and protocol design and performance analysis for computer and telecommunication networks, with applications in satellite, mobile, personal communications and high speed networks.

Dr. Leung is a Fellow of IEEE and a voting member of ACM. He is an editor of the IEEE Transactions on Wireless Communications, and an associate editor of the IEEE Transactions on Vehicular Technology. He has served on the technical program committees of numerous conferences, and is serving as the Technical Program Vice-Chair of IEEE WCNC 2005.
E-mail: vleung@ece.ubc.ca