Noname manuscript No.
(will be inserted by the editor)

A Constrained MDP-based Vertical Handoff Decision Algorithm for 4G Heterogeneous Wireless Networks

Chi Sun · Enrique Stevens-Navarro · Vahid Shah-Mansouri · Vincent W.S. Wong

Received: 22nd December 2008 / Accepted: 10th January 2011

Abstract The 4th generation wireless communication systems aim to provide users with the convenience of seamless roaming among heterogeneous wireless access networks. To achieve this goal, the support of vertical handoff is important in mobility management. This paper focuses on the vertical handoff decision algorithm, which determines the criteria under which vertical handoff should be performed. The problem is formulated as a constrained Markov decision process. The objective is to maximize the expected total reward of a connection subject to the expected total access cost constraint. In our model, a benefit function is used to assess the quality of the connection, and a penalty function is used to model the signaling incurred and call dropping. The user's velocity and location information are also considered when making handoff decisions. The policy iteration and Q-learning algorithms are employed to determine the optimal policy. Structural results on the optimal vertical handoff policy are derived by using the concept of supermodularity. We show that the optimal policy is a threshold policy in bandwidth, delay, and velocity. Numerical results show that our proposed vertical handoff decision algorithm outperforms other decision schemes in a wide range of conditions such as variations on connection duration, user's velocity, user's budget, traffic type, signaling cost, and monetary access cost.

Keywords Vertical handoff · constrained Markov decision processes · heterogeneous wireless networks.

Department of Electrical and Computer Engineering
The University of British Columbia
2332 Main Mall, Vancouver, BC, Canada, V6T 1Z4
Tel.: +1-604-827-5135
Fax: +1-604-822-5949
E-mail: vincentw@ece.ubc.ca

1 Introduction

The goal of the 4th Generation (4G) wireless communication systems is to utilize the different wireless access technologies in order to provide multimedia services to users on an *anytime, anywhere* basis. Currently, the standardization bodies such as the 3rd Generation Partnership Project (3GPP) [1], 3GPP2 [2], and the IEEE 802.21 Media Independent Handover (MIH) working group [3] are working towards this vision. In 4G communication systems, users will have a variety of choices on the selection of wireless networks to send and/or receive their data. They can either choose to use Long Term Evolution (LTE) to benefit from good quality of service (QoS), Worldwide Interoperability for Microwave Access (WiMAX) to achieve a high data rate, or wireless local area network (WLAN) to enjoy a moderate access cost. As a result, the users in the 4G communication systems should be able to switch to whichever wireless network they want to use at any time, in a seamless manner. In other words, seamless mobility must be properly managed to achieve the goal of the 4G wireless systems.

Vertical handoff is responsible for service continuity when a connection needs to migrate across heterogeneous wireless access networks. It generally involves three phases [4], [5]: *system discovery*, *vertical handoff decision*, and *vertical handoff execution*. During the system discovery phase, the mobile terminal (MT) with multiple radio interfaces receives advertised information from different wireless access networks. The information may include the access costs and current QoS parameters for different services. In vertical handoff decision phase, the MT determines whether the current connection should continue to use the same network or be switched to another one. The decision is based on the information that the MT received during the system discovery phase, as well as the conditions of its current state (e.g., MT's current location, velocity, and battery status). In the

vertical handoff execution phase, the connections are seamlessly migrated from the existing network to another. This process involves authentication, authorization, and also the transfer of context information.

We now summarize some of the recent work on vertical handoff decision algorithms in heterogeneous wireless networks. In [6], a fuzzy logic-based vertical handoff decision algorithm is proposed. Network parameters such as the current received signal strength (RSS), predicted RSS, user's velocity, and the available bandwidth are considered when making the decision. In [7], a middleware solution called vertical handoff manager is implemented to address the vertical handoff decision problem. The architecture of the vertical handoff manager consists of three components: network handling manager, feature collector, and artificial neural networks selector. In [8], the vertical handoff decision is based on the cost function of each candidate network. A cost function consists of three aspects such as the access network capacity, signalling cost, and the load balancing factor. The chosen network is the one with the least cost function value. In [9], the WLAN is selected as the preferred network for the MT. The objective of the handoff decision algorithm is to maximize the time during which the MT is served by the WLAN, while satisfying the QoS requirements as well as the call dropping probability and the average number of ping-pong events constraints.

In [10], the vertical handoff decision is formulated as a fuzzy multiple attribute decision making (MADM) problem. Two MADM methods are proposed: SAW (Simple Additive Weighting) and TOPSIS (Technique for Order Preference by Similarity to Ideal Solution). In SAW, the overall score of a candidate network is determined by the weighted sum of all attribute values. In TOPSIS, the selected candidate network is the one which is the closest to the ideal network, where the property of the ideal network is obtained by using the best values for each metric considered. In [11], a vertical handoff decision scheme based on ELECTRE is proposed. ELECTRE is an MADM algorithm, which performs pair-wise comparisons among the alternatives. The attributes considered in [11] include the bandwidth, delay, packet jitter, packet loss, utilization, and network cost.

In [12], a vertical handoff decision algorithm which uses the received signal to interference and noise ratio (SINR) from various access networks as the handoff criterion is proposed. It has the ability to make handoff decision with multimedia QoS consideration, such as to offer the user maximum downlink throughput from the integrated network, or to guarantee the minimum user required data rate during vertical handoff. In [13], a handoff management system which includes several modules and procedures is proposed. It determines the destination network based on the sojourn time of the MT in the candidate networks and the QoS estimation of these networks, including RSS, channel utilization, and link delay/jitter. Based on the output of the handoff decision module, the system will choose to either enter a handoff routine or keep the current connection. In [14], an RSS-based handoff decision scheme is implemented. By applying the auto-regressive integrated moving average model, the future RSS values can be predicted. The handoff decision can then be made according to these RSS predictions. In [15], a utility-based network selection strategy is presented. Several utility functions are examined which explore different users' preferences on their current applications.

In [16], a vertical handoff decision algorithm based on dynamic programming is proposed. Since the enhancement of a user's satisfaction by a vertical handoff depends on the user's sojourn time in the wireless network (e.g., WLAN), the algorithm takes the user's location and mobility information into consideration. The user's velocity and moving patterns are also considered in the vertical handoff decision algorithms in [17] and [18]. Moreover, in [19], a framework is proposed to evaluate different vertical handoff decision algorithms, in which the MT's mobility is modeled by a Markov chain.

In [20], a Markov decision process (MDP) approach for vertical handoff decision making problem is proposed. This MDP approach takes into account multiple factors such as user's preference, network conditions, and device capability. In [21], the vertical handoff decision problem is formulated as an MDP model. The model considers the available bandwidth and delay of the candidate networks. The model in this paper is an extension of the one proposed in [21], such that it not only considers the QoS of the candidate networks, but also takes the user's mobility and location information into account. Moreover, this model addresses a practical issue that users have monetary budgets for their connections.

Although there have been various vertical handoff decision algorithms proposed in the literature, most of them only make decisions based on the current system state (e.g., current QoS of the networks and current MT's conditions). Handoff decision should also consider the probabilistic outcomes of the future system states as a result of the current decision. Some work (e.g., [16], [20], and [21]) follows this approach; however those algorithms do not take the user's monetary budget into consideration. In our work, the vertical handoff decision algorithm considers the following aspects:

1. The state of the wireless access networks. This includes the available bandwidth, delay, switching cost, and access cost information of the overlaying networks.
2. The state of the user and MT. This includes the user's velocity and location information.
3. The preference of the user.
4. The current condition of the system as well as its future possible evolutions.

5. User's monetary budget. For example, a user may agree to spend at most $3 for a multimedia session with an average duration of 30 minutes.

Considering these aspects, we propose a vertical handoff decision algorithm for 4G wireless networks based on the constrained MDP (CMDP) model. The objective of the problem is to determine the policy which maximizes the expected total reward per connection, while the expected total access cost associated with this connection does not exceed the user's budget for it. The main contributions of this paper are as follows [22] [23]:

- Our CMDP-based vertical handoff decision algorithm takes into account the resources available in different networks (e.g., QoS, switching costs, access costs), and the MT's information (e.g., location, velocity). A benefit function is used to model the available bandwidth and delay of the connection. A penalty function is used to model the signaling incurred and the call dropping probability. For each connection, an access cost function is used to capture the access cost of using a specific network.
- We determine the optimal vertical handoff policy for decision making via the use of policy iteration and Q-learning algorithms.
- We derive structural results regarding the optimal vertical handoff policy, and show that the optimal policy is a threshold policy in available bandwidth, delay, and velocity.
- We evaluate the performance of our proposed algorithm under different criteria. Numerical results show that our vertical handoff decision algorithm outperforms other decision schemes (e.g., SAW [10], ELECTRE [11]) in a wide range of conditions such as variations on connection duration, user's velocity, user's budget, traffic type, signaling cost, and monetary access cost.

The rest of the paper is organized as follows. The system model is described in Section 2. The CMDP formulation and optimality equations are presented in Section 3. Section 4 investigates the structure of the optimal vertical handoff policy. Section 5 presents the numerical results and discussions. Conclusions are given in Section 6.

2 System Model

In this section, we describe how the vertical handoff decision problem can be formulated as a finite state, infinite horizon CMDP. A CMDP model can be characterized by six elements: *decision epochs, states, actions, transition probabilities, rewards,* and *costs* [24]. At each decision epoch, the MT has to choose an action (i.e., select a network) based on the current system state (e.g., QoS that can be provided by each candidate network, velocity and location of the MT). With this state and action, the system then evolves to a new state according to a transition probability function. This new state lasts for a period of time until the next decision epoch comes, and then the MT makes a new decision again (i.e., selects a network again). For any action that the MT chooses at each state, there is a reward and a cost associated with it. The goal of each MT is to maximize the expected total reward that it can obtain during the connection, subject to a constraint on its expected total access cost.

2.1 States, Actions, and Transition Probabilities

We represent the decision epochs by $T = \{1, 2, \ldots, N\}$, where the random variable N indicates the time that the connection terminates. We denote the state space of the MT by $\mathbf{S}$, and we only consider a finite number of states that the system can possibly be in. The state of the system contains information such as the current network that the MT connects to, the available bandwidth and delay that the candidate networks can offer, and the velocity and location information of the MT. Specifically, the state space can be expressed as

$$\mathbf{S} = \mathbf{M} \times \mathbf{B}^1 \times \mathbf{D}^1 \times \cdots \times \mathbf{B}^{|\mathbf{M}|} \times \mathbf{D}^{|\mathbf{M}|} \times \mathbf{V} \times \mathbf{L},$$

where $\times$ denotes the Cartesian product, $\mathbf{M}$ represents the set of available network IDs that the MT can connect to. $\mathbf{B}^m$ and $\mathbf{D}^m$ denote the set of available bandwidth and delay of network $m \in \mathbf{M}$, respectively. $\mathbf{V}$ denotes the set of possible velocity values of the MT, and $\mathbf{L}$ denotes the set of location types that the MT can possibly reside in.

In order to reduce the size of the state space, we consider a *finite countable* state space in this paper. The bandwidth and delay can be quantized into multiples of unit bandwidth and unit delay, respectively [24]. Specifically, the set of available bandwidth of network m is

$$\mathbf{B}^m = \{1, 2, \ldots, b^m_{max}\}, \qquad m \in \mathbf{M},$$

where b^m_{max} denotes the maximum bandwidth available to a connection from network m. For example, the unit bandwidth of WLAN and the LTE network can be 500 *kbps* and 16 *kbps*, respectively.

Similarly, the set of packet delay of network m is

$$\mathbf{D}^m = \{1, 2, \ldots, d^m_{max}\}, \qquad m \in \mathbf{M},$$

where d^m_{max} denotes the maximum delay provided to a connection by network m. For example, the unit delay of WLAN and the LTE network can be 50 *ms* and 20 *ms*, respectively.

The velocity of the MT is also quantized as multiples of unit velocity. The set of possible velocity values is

$$\mathbf{V} = \{1, 2, \ldots, v_{max}\},$$

where v_{max} denotes the maximum velocity that an MT can travel at. For example, the unit of velocity can be 10 *km/h*.

For the set of location types that the MT can possibly reside in, we have

$$\mathbf{L} = \{1, 2, \ldots, l_{max}\},$$

where l_{max} denotes the total number of different location types in the area of interest. Location types are differentiated by the number of networks they are covered by.

Let vector $\mathbf{s} = [i, b_1, d_1, \ldots, b_{|\mathbf{M}|}, d_{|\mathbf{M}|}, v, l]$ denote the current state of the MT, where i denotes the current network used by the connection, b_m and d_m denote the current bandwidth and delay of network m, respectively, v denotes the current velocity of the MT, and l denotes the current location type that the MT resides in. At each decision epoch, based on the current state $\mathbf{s}$, the action of the MT is to decide whether to remain connected to the existing network or to switch to another network. Let $\mathbf{A_s} \subset \mathbf{M}$ denote the action set, which consists of the ID of the networks that the MT can potentially switch to given the current state $\mathbf{s}$. Thus, the action $a \in \mathbf{A_s}$ is to select one of the available networks from the set $\mathbf{A_s}$. In other words, the chosen action a corresponds to the selected network. Given the current state is $\mathbf{s}$ and the chosen action is a, the probability that the next state becomes $\mathbf{s}' = [j, b'_1, d'_1, \ldots, b'_{|\mathbf{M}|}, d'_{|\mathbf{M}|}, v', l']$ is

$$P[\mathbf{s}' \mid \mathbf{s}, a] = \begin{cases} P[v' \mid v] P[l' \mid l] \prod_{m \in \mathbf{M}} P[b'_m, d'_m \mid b_m, d_m], & j = a, \\ 0, & j \neq a, \end{cases} \tag{1}$$

where $P[v' \mid v]$ is the transition probability of the MT's velocity, $P[l' \mid l]$ is the transition probability of the MT's location type, and $P[b'_m, d'_m \mid b_m, d_m]$ is the joint transition probability of the bandwidth and delay of network m. We now explain how we obtain these transition probabilities. The transition probability of the MT's velocity is obtained based on the Gauss-Markov mobility model from [25]. In this mobility model, an MT's velocity is assumed to be correlated in time and can be modeled by a discrete Gauss-Markov random process. The following recursive realization is used to calculate the transition probability of the MT's velocity

$$v' = \alpha v + (1 - \alpha)\mu + \sigma\sqrt{1 - \alpha^2}\zeta, \tag{2}$$

where v is the MT's velocity at the current decision epoch, v' is the MT's velocity at the next decision epoch, α is the memory level (i.e., $0 \leq \alpha \leq 1$), μ and σ are the mean and standard deviation of the velocity, respectively, and ζ is an uncorrelated Gaussian process with zero mean and unit variance. By varying v and counting the number of different outcomes of v' according to (2), the MT's velocity transition probability function (i.e., $P[v' \mid v]$) can be determined.

For the transition probability of the MT's location type, we assume that a wireless access network which has a smaller coverage area (e.g., WLAN) always lies within another access network which has a larger coverage area (e.g., WiMAX). Although this assumption may not hold for the cases when $|\mathbf{M}|$ is large, it is reasonable as long as the number of different network types does not exceed three, which is a typical case in today's wireless communication systems.

We define location type $l \in \mathbf{L}$ as follows. Location type 1 is the area covered only by the LTE network. Location type 2 is the area covered by LTE and WiMAX, but not WLAN. Location type 3 is the area covered by all three networks (i.e., LTE, WiMAX, and WLAN). Let Θ_l denote the total area of location type l and ρ_l denote the user density of location type l. The effective area of location type l is defined as

$$\hat{\Theta}_l = \Theta_l \, \rho_l, \qquad l \in \mathbf{L}. \tag{3}$$

In practice, the user density in areas covered by different access networks (e.g., WLAN and the LTE network) is usually not the same [26], [27]. For example, the area covered by both WLAN and the LTE network usually has more active connections than the area only covered by the LTE network. As a result, the density index of each location type is considered in order to achieve a more realistic model.

We assume that an MT currently at location type l can only move to its neighboring location types (i.e., either $l+1$ or $l-1$) or stay at l at the next decision epoch. This is because the duration of each decision epoch is too short for the MT to traverse more than one location type area[1]. Thus, the probability that an MT's next location type is l' given its current location type is l is assumed to be proportional to the effective area of l'. Specifically, the transition probability of an MT's location type, denoted as $P[l' \mid l]$

$$= \begin{cases} \dfrac{\hat{\Theta}_{l'}}{\sum_{\xi=l,l+1} \hat{\Theta}_\xi}, & l = 1, \;\; l' = 1, 2, \\ \dfrac{\hat{\Theta}_{l'}}{\sum_{\xi=l-1,l,l+1} \hat{\Theta}_\xi}, & l = 2, \ldots, l_{max} - 1, \;\; l' = l-1, l, l+1, \\ \dfrac{\hat{\Theta}_{l'}}{\sum_{\xi=l-1,l} \hat{\Theta}_\xi}, & l = l_{max}, \;\; l' = l_{max} - 1, l_{max}. \end{cases} \tag{4}$$

For the joint transition probabilities of the bandwidth and delay of each network, we use the following approach to estimate them. For the cellular network, the values of bandwidth and delay are assumed to be guaranteed for the duration of the connection (i.e., $P[b'_1 = b_1, d'_1 = d_1 \mid b_1, d_1] = 1$).

[1] The time between two successive decision epochs is on the order of seconds.

For WiMAX and WLAN, we estimate such probabilities in a simulation-based manner. In ns-2 simulator [28], typical IEEE 802.16 WiMAX [29] and IEEE 802.11*b* WLAN are simulated in which the users arrive and depart from the networks according to Poisson processes. The resulting available bandwidth and delay are rounded according to the predefined units, and then the counting of transitions among states is performed to estimate the state transition probabilities of WiMAX and WLAN (i.e., $P[b_2', d_2' \mid b_2, d_2]$ and $P[b_3', d_3' \mid b_3, d_3]$, respectively).

2.2 Rewards

When an MT in state $\mathbf{s} = [i, b_1, d_1, \ldots, b_{|\mathbf{M}|}, d_{|\mathbf{M}|}, v, l]$ chooses an action a, it receives an immediate reward $r(\mathbf{s}, a)$. This reward function is composed of a benefit function and a penalty function, which are explained in detail below.

For the benefit function of the MT, two aspects are considered: bandwidth and delay. Let the *bandwidth benefit function* represent the benefit that an MT can gain (in terms of bandwidth) by selecting action a in state $\mathbf{s}$ (recall that i denotes the ID of the current network)

$$f_b(\mathbf{s}, a) = \begin{cases} \frac{b_a - b_i}{\max\limits_{k \in \mathbf{M}} \{b_k - b_i\}}, & \text{if } b_a > b_i, \\ 0, & \text{if } b_a = b_i, \\ -\frac{b_a - b_i}{\min\limits_{k \in \mathbf{M}} \{b_k - b_i\}}, & \text{if } b_a < b_i. \end{cases} \tag{5}$$

The benefit is being assessed as follows. Given that the MT is currently connecting to network i, if the action a leads to a network with a higher bandwidth, then the benefit function value is represented by a fraction, in which the numerator is the MT's actual increase of bandwidth by choosing action a in state $\mathbf{s}$, and the denominator is the MT's maximum possible increase of bandwidth. As a result, the benefit function value is a positive number between 0 and 1. Similarly, if the action leads to a network with a lower bandwidth, the benefit function value becomes a negative number between -1 and 0. Finally, if the MT chooses to remain at the same network, then the benefit function value is 0.

Similarly, a *delay benefit function* is used to represent the benefit that an MT can gain (in terms of delay) by choosing action a in state $\mathbf{s}$:

$$f_d(\mathbf{s}, a) = \begin{cases} \frac{d_i - d_a}{\max\limits_{k \in \mathbf{M}} \{d_i - d_k\}}, & \text{if } d_a < d_i, \\ 0, & \text{if } d_a = d_i, \\ -\frac{d_i - d_a}{\min\limits_{k \in \mathbf{M}} \{d_i - d_k\}}, & \text{if } d_a > d_i. \end{cases} \tag{6}$$

As a result, the total *benefit function* is given by

$$f(\mathbf{s}, a) = \omega f_b(\mathbf{s}, a) + (1 - \omega) f_d(\mathbf{s}, a), \qquad \mathbf{s} \in \mathbf{S},\ a \in \mathbf{A_s}, \tag{7}$$

where $f_b(\mathbf{s}, a)$ and $f_d(\mathbf{s}, a)$ are both normalized (i.e., between 0 and 1), and ω is the weight given to the bandwidth aspect with $0 \le \omega \le 1$. This weight can be set differently for different types of applications (e.g., constant bit rate (CBR) voice traffic, file transfer protocol (FTP) data traffic).

We consider two factors for the penalty of the MT. First, the *switching cost penalty function* is represented by

$$g_{switch}(\mathbf{s}, a) = \begin{cases} K_{i,a}, & \text{if } i \neq a, \\ 0, & \text{if } i = a, \end{cases} \tag{8}$$

where $K_{i,a}$ is the normalized switching cost from network i to network a. This penalty function captures the processing and signaling load incurred when the connection is migrated from one network to another. Second, we define the *call dropping penalty function* as

$$g_{drop}(\mathbf{s}, a) = \begin{cases} 0, & \text{if } i = a, \\ 0, & \text{if } i \neq a,\ 0 \le v \le V_{min}, \\ \frac{v - V_{min}}{V_{max} - V_{min}}, & \text{if } i \neq a,\ V_{min} < v < V_{max}, \\ 1, & \text{if } i \neq a,\ v \ge V_{max}, \end{cases} \tag{9}$$

where V_{max} and V_{min} denote the maximum and minimum velocity thresholds, respectively. When the MT moves faster, the probability that the connection will be dropped during the vertical handoff process increases. For example, if an MT moves out of a WLAN with a high speed, it may enter the area covered only by the LTE network (hence lose the WLAN signal) before the WLAN-to-LTE vertical handoff procedure is completed. As a result, the MT's connection may be dropped.

Consequently, the total *penalty function* of an MT is given by

$$g(\mathbf{s}, a) = \phi g_{switch}(\mathbf{s}, a) + (1 - \phi) \kappa g_{drop}(\mathbf{s}, a), \quad \mathbf{s} \in \mathbf{S},\ a \in \mathbf{A_s}, \tag{10}$$

where ϕ is the weight given to the switching cost factor with $0 \le \phi \le 1$, and $\kappa \in [0, 1]$ is the user's preference on vertical handoff. Some users would allow vertical handoffs in order to obtain better QoS although there is a risk that the connection may be dropped during handoff. Other users may refrain from switching whenever a risk is present.

Finally, the *reward function* between two successive vertical handoff decision epochs is

$$r(\mathbf{s}, a) = f(\mathbf{s}, a) - g(\mathbf{s}, a), \qquad \mathbf{s} \in \mathbf{S},\ a \in \mathbf{A_s}, \tag{11}$$

and is normalized within the range from 0 to 1. Recall this reward function considers the bandwidth and delay of all candidate networks, the signaling cost incurred when switching occurs, the call dropping probability, and the user's preference.

2.3 Costs

For each period of time that the MT uses a network, it will incur the following access cost (in monetary units per unit time):

$$c(\mathbf{s},a) = \frac{b_a\, C_a}{\max\limits_{m\in\mathbf{M}}\{b_m\, C_m\}}, \qquad \mathbf{s}\in\mathbf{S},\ a\in\mathbf{A_s}, \tag{12}$$

where b_m is the available bandwidth in *bps* and C_m is the access cost of network m in monetary units per bit. This access cost is normalized such that its value is between 0 and 1. The user has a budget such that it is willing to spend up to C_{max} monetary units per connection.

3 CMDP Formulation and Optimality Equations

In this section, we present the problem formulation and describe how to obtain the optimal policy. First, some concepts need to be clarified. A *decision rule* specifies the action selection for each state at a particular decision epoch. It can be expressed as $\delta_t : \mathbf{S}\to\mathbf{A}$, where δ_t represents the decision rule at decision epoch t. A *policy* $\pi = (\delta_1,\delta_2,\ldots,\delta_N)$ is a set of sequential decision rules to be used at all N decision epochs.

Let $v^\pi(\mathbf{s})$ denote the *expected total reward* between the first decision epoch and the connection termination, given that policy π is used with initial state $\mathbf{s}$. We can represent $v^\pi(\mathbf{s})$ as

$$v^\pi(\mathbf{s}) = E_{\mathbf{s}}^\pi\left[E_N\left\{\sum_{t=1}^{N} r(\mathbf{s}_t,a_t)\right\}\right], \qquad \mathbf{s}\in\mathbf{S}, \tag{13}$$

where $E_{\mathbf{s}}^\pi$ denotes the expectation with respect to the variables given policy π and initial state $\mathbf{s}$, and E_N denotes the expectation with respect to the random variable N. The random variable N, which denotes the *connection termination time*, is assumed to be geometrically distributed with mean $1/(1-\lambda)$. Equation (13) can be re-written as

$$v^\pi(\mathbf{s}) = E_{\mathbf{s}}^\pi\left\{\sum_{t=1}^{\infty}\lambda^{t-1}\, r(\mathbf{s}_t,a_t)\right\}, \qquad \mathbf{s}\in\mathbf{S}, \tag{14}$$

where λ can also be interpreted as the *discount factor* of the model (i.e., $0\le\lambda<1$). We define a policy $\pi^* = (\delta_1^*,\delta_2^*,\cdots)$ to be *optimal* in Π if $v^{\pi^*}(\mathbf{s})\ge v^\pi(\mathbf{s})$ for all $\pi\in\Pi$, where Π is the set of all possible policies. A policy is said to be *stationary* if $\delta_t=\delta$ for all t. A stationary policy has the form $\pi=(\delta,\delta,\cdots)$, and for convenience we denote π simply by δ. A policy is said to be *deterministic* if it chooses an action with certainty at each decision epoch. We refer to stationary deterministic policies as *pure* policies [30, pp. 22].

Since our objective is to maximize the *expected discounted total reward* (i.e., $v^\pi(\mathbf{s})$) subject to an access cost constraint, we can state the CMDP optimization problem as

$$\begin{aligned}&\text{maximize}\ \ v^\pi(\mathbf{s}) = E_{\mathbf{s}}^\pi\left\{\sum_{t=1}^{\infty}\lambda^{t-1}\, r(\mathbf{s}_t,a_t)\right\}\\ &\text{subject to}\ \ C^\pi(\mathbf{s}) = E_{\mathbf{s}}^\pi\left\{\sum_{t=1}^{\infty}\lambda^{t-1}\, c(\mathbf{s}_t,a_t)\right\}\le C_{max},\end{aligned} \tag{15}$$

where $C^\pi(\mathbf{s})$ denotes the *expected discounted total access cost* with respect to the variables given policy π and initial state $\mathbf{s}\in\mathbf{S}$.

To solve (15) without the constraint on the expected discounted total access cost, we can use the Policy Iteration Algorithm (PIA) [30, 31] to solve the following *optimality equations*

$$v(\mathbf{s}) = \max_{a\in\mathbf{A_s}}\left\{r(\mathbf{s},a) + \sum_{\mathbf{s}'\in\mathbf{S}}\lambda\ P[\mathbf{s}'\mid\mathbf{s},a]\ v(\mathbf{s}')\right\}, \qquad \mathbf{s}\in\mathbf{S}, \tag{16}$$

and the corresponding optimal policy is given as

$$\delta^*(\mathbf{s}) = \arg\max_{a\in\mathbf{A_s}}\left\{r(\mathbf{s},a) + \sum_{\mathbf{s}'\in\mathbf{S}}\lambda\ P[\mathbf{s}'\mid\mathbf{s},a]\ v(\mathbf{s}')\right\}, \quad \mathbf{s}\in\mathbf{S}. \tag{17}$$

However, since (15) has a constraint in it, we cannot use the PIA directly. We need to first use the *Lagrangian approach* [24, 32] to convert the CMDP problem to an unconstrained MDP problem. By including the Lagrange multiplier β with $\beta>0$, we have

$$r(\mathbf{s},a;\beta) = r(\mathbf{s},a) - \beta c(\mathbf{s},a), \qquad \mathbf{s}\in\mathbf{S},\ a\in\mathbf{A_s}, \tag{18}$$

where $r(\mathbf{s},a;\beta)$ is the *Lagrangian reward function*.

After the Lagrangian approach, the new *optimality equations* are given by

$$v_\beta(\mathbf{s}) = \max_{a\in\mathbf{A_s}}\left\{r(\mathbf{s},a;\beta) + \sum_{\mathbf{s}'\in\mathbf{S}}\lambda\ P[\mathbf{s}'\mid\mathbf{s},a]\ v_\beta(\mathbf{s}')\right\}, \quad \mathbf{s}\in\mathbf{S}, \tag{19}$$

which can be solved by using the PIA with a fixed value of β. The procedures of the PIA algorithm are described in Algorithm 1. We denote the vectors $\mathbf{v}_\beta^k = \left(v_\beta^k(\mathbf{s}),\ \mathbf{s}\in\mathbf{S}\right)$, $\mathbf{r}_\beta = \left(r\left(\mathbf{s},\delta_\beta^k(\mathbf{s});\beta\right),\ \mathbf{s}\in\mathbf{S}\right)$. We denote the matrix $\mathbb{P} = \left(P[\mathbf{s}'\mid\mathbf{s},\delta_\beta^k(\mathbf{s})],\ \mathbf{s}\in\mathbf{S},\mathbf{s}'\in\mathbf{S}\right)$.

Algorithm 1 - The Policy Iteration Algorithm (PIA)

1: Set $k=0$, and select an arbitrary decision rule $\delta_\beta^0(\mathbf{s}) \in \mathbf{A_s}$ for all $\mathbf{s} \in \mathbf{S}$.
2: (Policy evaluation) Obtain $v_\beta^k(\mathbf{s})$ for all $\mathbf{s} \in \mathbf{S}$ by solving

$$(I - \lambda\ \mathbb{P})\ \mathbf{v}_\beta^k = \mathbf{r}_\beta,$$

where I is an identity matrix of size $|\mathbf{S}|$.
3: (Policy improvement) Choose $\delta_\beta^{k+1}(\mathbf{s})$ which satisfy

$$\delta_\beta^{k+1}(\mathbf{s}) = \arg\max_{a \in \mathbf{A_s}} \left\{ r(\mathbf{s},a;\beta) + \sum_{\mathbf{s}' \in \mathbf{S}} \lambda\ P[\mathbf{s}' \mid \mathbf{s},a]\ v_\beta^k(\mathbf{s}') \right\},$$

for all $\mathbf{s} \in \mathbf{S}$.
4: If $\delta_\beta^{k+1}(\mathbf{s}) = \delta_\beta^k(\mathbf{s})$ for all $\mathbf{s} \in \mathbf{S}$, stop and set $\delta_\beta^*(\mathbf{s}) = \delta_\beta^k(\mathbf{s})$. Otherwise, increment k by 1 and return to step 2.

The solutions of (19) correspond to the maximum expected discounted total reward $v_\beta(\mathbf{s})$ and the pure policy $\delta_\beta^*(\mathbf{s})$ which satisfies

$$\delta_\beta^*(\mathbf{s}) = \arg\max_{a \in \mathbf{A_s}} \left\{ r(\mathbf{s},a;\beta) + \sum_{\mathbf{s}' \in \mathbf{S}} \lambda P[\mathbf{s}' \mid \mathbf{s},a] v_\beta(\mathbf{s}') \right\},\ \mathbf{s} \in \mathbf{S}. \tag{20}$$

Note the pure policy $\delta_\beta^*(\mathbf{s})$ specifies the network to choose in each state $\mathbf{s}$ such that the expected discounted total reward is maximized.

When the CMDP problem is converted to an unconstrained MDP problem by a Lagrange multiplier β in (18), there is a relationship between the constraint (i.e., the user's budget C_{max}) and the Lagrange multiplier (i.e., β). In this paper, we use the Q-learning algorithm (see Algorithm 2) proposed in [33] to determine the proper β (i.e., β^*) for a feasible C_{max}.

Once β^* has been obtained, we follow the procedures in [32] [33] to find the optimal policy for the CMDP problem. As discussed in [33], the optimal policy for a CMDP with single constraint is a mixed policy of two pure policies. First, we perturb β^* by some $\Delta\beta$ to obtain $\beta^- = \beta^* - \Delta\beta$ and $\beta^+ = \beta^* + \Delta\beta$. Then, we calculate the pure policies δ^- and δ^+ in (20) (using β^- and β^+, respectively) via PIA and their corresponding average expected discounted total access costs $\bar{C}^-$ and $\bar{C}^+$ in (24) (using δ^- and δ^+, respectively). Next, we define a parameter q such that $q\bar{C}^- + (1-q)\bar{C}^+ = C_{max}$. Thus, the *optimal policy* δ^* of the CMDP problem is a *randomized* mixture of two policies (i.e., δ^- and δ^+), such that at each decision epoch, the first policy δ^- is chosen with probability q and the second policy δ^+ is chosen with probability $1-q$. In other words, the optimal policy can be obtained as follows:

$$\delta^*(\mathbf{s}) = q\delta^-(\mathbf{s}) + (1-q)\delta^+(\mathbf{s}), \qquad \mathbf{s} \in \mathbf{S}. \tag{21}$$

Algorithm 2 - The Q-learning Algorithm

1: Set β_1 to an arbitrary number greater than zero, and set $n=1$.
2: Solve for $\delta_{\beta_n}^*(\mathbf{s})$ in (20) via PIA for all $\mathbf{s} \in \mathbf{S}$.
3: Determine $C^{\delta_{\beta_n}^*}(\mathbf{s})$ from the following equation

$$C^{\delta_{\beta_n}^*}(\mathbf{s}) = c(\mathbf{s}, \delta_{\beta_n}^*(\mathbf{s})) + \sum_{\mathbf{s}' \in \mathbf{S}} \lambda\ P[\mathbf{s}' \mid \mathbf{s}, \delta_{\beta_n}^*(\mathbf{s})]\ C^{\delta_{\beta_n}^*}(\mathbf{s}'), \tag{22}$$

for all $\mathbf{s} \in \mathbf{S}$.
4: Update the Lagrange multiplier by

$$\beta_{n+1} = \beta_n + \frac{1}{n}\left(\bar{C}^{\delta_{\beta_n}^*} - C_{max}\right), \tag{23}$$

where

$$\bar{C}^{\delta_{\beta_n}^*} = \frac{1}{|\mathbf{S}|} \sum_{\mathbf{s} \in \mathbf{S}} C^{\delta_{\beta_n}^*}(\mathbf{s}). \tag{24}$$

5: If $|\beta_{n+1} - \beta_n| \le \varepsilon$, stop and set $\beta^* = \beta_{n+1}$. Otherwise, increment n by 1 and return to step 2.

4 Monotone Optimal Vertical Handoff Policy

Given the proper selection of the Lagrange multiplier β (i.e., β^*), the unconstrained MDP in (19) has a stationary optimal policy. It can be shown that for a scenario with two wireless access networks, the unconstrained MDP and the CMDP optimal vertical handoff policies are monotone in the available bandwidth, delay, and velocity. Monotonicity and the existence of two actions $\mathbf{A_s} = \{1,2\}$ define a *threshold policy*. The threshold policies are optimal policies with a special structure that facilitates computation and implementation [30].

4.1 Threshold Structure of Unconstrained MDP

Since two wireless access networks are considered (i.e., $\mathbf{M} = \{1,2\}$), the current system state $\mathbf{s} = [i, b_1, d_1, b_2, d_2, v, l]$, where i denotes the current network used by the MT, b_1 and b_2 denote the current available bandwidth in network 1 and 2, respectively, d_1 and d_2 denote the current delay in network 1 and 2, respectively, v denotes the current velocity of the MT, and l denotes the current location type that the MT resides in.

Recall that the unconstrained MDP can be solved by using PIA. From Algorithm 1, in each iteration $k \in \{0,1,2,\ldots\}$, we have

$$v_\beta^{k+1}(\mathbf{s}) = \max_{a \in \mathbf{A_s}} \left\{ r(\mathbf{s},a;\beta) + \sum_{\mathbf{s}' \in \mathbf{S}} \lambda\ P[\mathbf{s}' \mid \mathbf{s},a]\ v_\beta^k(\mathbf{s}') \right\}, \tag{25}$$

and

$$Q_\beta^{k+1}(\mathbf{s},a) = r(\mathbf{s},a;\beta) + \sum_{\mathbf{s}' \in \mathbf{S}} \lambda\ P[\mathbf{s}' \mid \mathbf{s},a]\ v_\beta^k(\mathbf{s}'), \tag{26}$$

where we refer $v_\beta^k(\mathbf{s})$ as the *value function* and $Q_\beta^k(\mathbf{s},a)$ as the *state-action reward function*. For any initial state $\mathbf{s}$ of $v_\beta(\mathbf{s})$, the sequence $v_\beta^k(\mathbf{s})$ generated by the PIA converges to the optimal expected discounted total reward for all $\mathbf{s}\in\mathbf{S}$ [30]. To establish the monotone structure of the optimal policy for any discount factor $0\le\lambda<1$, the concept of *supermodularity* needs to be introduced.

Definition 1. *A function $\mathbf{F}(x,y):\mathbf{X}\times\mathbf{Y}\to\mathbf{R}$ is supermodular in (x,y) if $\mathbf{F}(x_1,y_1)+\mathbf{F}(x_2,y_2)\ge\mathbf{F}(x_1,y_2)+\mathbf{F}(x_2,y_1)$ for all $x_1,x_2\in\mathbf{X}$, $y_1,y_2\in\mathbf{Y}$, $x_1>x_2$, $y_1>y_2$. If the inequality is reversed, then the function $\mathbf{F}(x,y)$ is submodular.*

If the state-action reward function $Q_\beta(\mathbf{s},a)$ is supermodular (or submodular) in action a and another variable in the state $\mathbf{s}$ (e.g., b_2), then the optimal vertical handoff policy is monotone in that variable (i.e., b_2). In fact, supermodularity is a sufficient condition for optimality of monotone policies [30], [34]. Based on Definition 1, if $\mathbf{F}(x,y)$ is supermodular (submodular) in (x,y), then $y(x)=\arg\max_y\mathbf{F}(x,y)$ is monotonically non-decreasing (non-increasing) in variable x [34].

By supermodularity, we can see that the state-action reward function $Q_\beta(\mathbf{s},a)$ being submodular in (b_2,a), and supermodular in (d_2,a) and (v,a) implies that the optimal vertical handoff policy $\delta_\beta^*(\mathbf{s})$ is monotonically non-increasing in the available bandwidth, and non-decreasing in the delay and velocity, respectively.

The methodology of proving the threshold structure of the optimal policy consists of the following steps:

1. Proof on the monotonicity of the value function (Lemma 1);
2. Proof on the supermodularity/submodularity of the state-action reward function (Theorems 1, 2, and 3).

Then, the threshold structure of the optimal vertical handoff policy follows.

Lemma 1. *For any discount factor $0\le\lambda<1$, the optimal expected discounted total reward (i.e., the value function $v_\beta(\mathbf{s})$) is monotonically non-decreasing in the available bandwidth, and non-increasing in the delay and velocity.*

The proof of Lemma 1 is given in Appendix A.

In the following theorems, with loss of generality, we assume the current network in use is network 2 (i.e., $i=2$).

Theorem 1. *For any discount factor $0\le\lambda<1$, if the value function $v_\beta(\mathbf{s})$ is a monotonically non-decreasing function of the available bandwidth, and the state-action reward function $Q_\beta(\mathbf{s},a)$ is submodular in (b_2,a), that is,*

$$\begin{aligned}&Q_{\beta_c}(i,b_1,d_1,b_2,d_2,v,l,2)-Q_{\beta_c}(i,b_1,d_1,b_2,d_2,v,l,1)\\ \ge\ &Q_{\beta_c}(i,b_1,d_1,b_2+1,d_2,v,l,2)\\ &-Q_{\beta_c}(i,b_1,d_1,b_2+1,d_2,v,l,1),\end{aligned}\tag{27}$$

then the optimal policy is deterministic and monotonically non-increasing in the available bandwidth component of the state $\mathbf{s}$. *Consequently,* $\delta_\beta^*(i,b_1,d_1,b_2,d_2,v,l)$

$$=\begin{cases}a\ne i, & \text{if } 0\le b_2\le\tau_b(i,b_1,d_1,d_2,v,l),\\ a=i, & \text{if } b_2>\tau_b(i,b_1,d_1,d_2,v,l),\end{cases}\tag{28}$$

where $\tau_b(i,b_1,d_1,d_2,v,l)$ defines the threshold for the rest of the elements of the state $\mathbf{s}$ *for a given Lagrange multiplier* β.

The proof of Theorem 1 is given in Appendix B.

Theorem 2. *For any discount factor $0\le\lambda<1$, if the value function $v_\beta(\mathbf{s})$ is a monotonically non-increasing function of the delay, and the state-action reward function $Q_\beta(\mathbf{s},a)$ is supermodular in (d_2,a), that is,*

$$\begin{aligned}&Q_\beta(i,b_1,d_1,b_2,d_2,v,l,2)-Q_\beta(i,b_1,d_1,b_2,d_2,v,l,1)\\ \le\ &Q_\beta(i,b_1,d_1,b_2,d_2+1,v,l,2)\\ &-Q_\beta(i,b_1,d_1,b_2,d_2+1,v,l,1),\end{aligned}\tag{29}$$

then the optimal policy is deterministic and monotonically non-decreasing in the delay component of the state $\mathbf{s}$. *Consequently,* $\delta_\beta^*(i,b_1,d_1,b_2,d_2,v,l)$

$$=\begin{cases}a=i, & \text{if } 0\le d_2\le\tau_d(i,b_1,d_1,b_2,v,l),\\ a\ne i, & \text{if } d_2>\tau_d(i,b_1,d_1,b_2,v,l),\end{cases}\tag{30}$$

where $\tau_d(i,b_1,d_1,b_2,v,l)$ defines the threshold for the rest of the elements of the state $\mathbf{s}$ *for a given Lagrange multiplier* β.

The proof of Theorem 2 is given in Appendix C.

Theorem 3. *For any discount factor $0\le\lambda<1$, if the value function $v_\beta(\mathbf{s})$ is a monotonically non-increasing function of the velocity, and the state-action reward function $Q_\beta(\mathbf{s},a)$ is supermodular in (v,a), that is,*

$$\begin{aligned}&Q_\beta(i,b_1,d_1,b_2,d_2,v,l,2)-Q_\beta(i,b_1,d_1,b_2,d_2,v,l,1)\\ \le\ &Q_\beta(i,b_1,d_1,b_2,d_2,v+1,l,2)\\ &-Q_\beta(i,b_1,d_1,b_2,d_2,v+1,l,1),\end{aligned}\tag{31}$$

then the optimal policy is deterministic and monotonically non-decreasing in the velocity component of the state $\mathbf{s}$. *Consequently,* $\delta_\beta^*(i,b_1,d_1,b_2,d_2,v,l)$

$$=\begin{cases}a=i, & \text{if } 0\le v\le\tau_v(i,b_1,d_1,b_2,d_2,l),\\ a\ne i, & \text{if } v>\tau_v(i,b_1,d_1,b_2,d_2,l),\end{cases}\tag{32}$$

where $\tau_v(i,b_1,d_1,b_2,d_2,l)$ defines the threshold for the rest of the elements of the state $\mathbf{s}$ *for a given Lagrange multiplier* β.

The proof of Theorem 3 is given in Appendix D.

4.2 Threshold Structure of Constrained MDP

Having shown the threshold structure of the unconstrained MDP optimal policy and based on the results described in Section 3 and [33], we can state that the constrained optimal vertical handoff policy is a randomized mixture between two threshold policies.

Corollary 1. *There exists a stationary vertical handoff policy δ^* that is the optimal solution of the CMDP equivalent to the one given in (15) such that δ^* is a randomized mixture of two threshold vertical handoff policies as follows:*

$$\delta^*(\mathbf{s}) = q\delta^*_{\beta^-}(\mathbf{s}) + (1-q)\delta^*_{\beta^+}(\mathbf{s}), \qquad \mathbf{s} \in \mathbf{S}, \tag{33}$$

where $\delta^*_{\beta^-}$ and $\delta^*_{\beta^+}$ are two unconstrained optimal policies with Lagrange multipliers β^- and β^+ that are of the form (28) for the available bandwidth, (30) for the delay, or (32) for the velocity.

Proof. These results follow directly from Theorems 1, 2, 3, and [33, Theorem 4.3].

The existence of the monotone policy with a structure allows the use of more efficient algorithms which exploit features such as the structured policy iteration and the structured modified policy iteration (cf. [30]). These algorithms seek the optimal policy δ^* only in the subset of policies with certain structure (e.g., $\Pi_\delta \subset \Pi$). Consequently, computational effort will be considerably reduced by using structured algorithms.

Structured algorithms also facilitate implementation because they can be used to find the threshold values for the optimal policy. As an example, the monotone policy iteration algorithm [30, pp. 259] can be used. When an optimal policy is strictly increasing, the action set $\mathbf{A_s}$ decreases in size with increasing $\mathbf{s}$ (e.g., b_2) and hence reduces the number of actions which need to be evaluated by the algorithm. If at some state, say $\overline{b}_2 = \tau_b(i, b_1, d_1, d_2, v, l)$, the action set $\mathbf{A_s}$ contains only one element (i.e., a^*), then no further maximization is required because the action will be optimal for all $b_2 \geq \overline{b}_2$. Thus, the threshold value is $\overline{b}_2$ and the optimal action is a^* for all $b_2 \geq \overline{b}_2$.

5 Numerical Results and Discussions

We compare the performance of our proposed CMDP-based vertical handoff decision algorithm with two other schemes. The first one is the SAW algorithm [10]. The second one is ELECTRE [11], which is an MADM algorithm for network selection. The performance metric is the *expected total reward per connection*. Two applications are considered: constant bit rate (CBR) voice traffic over user datagram protocol (UDP), and file transfer protocol (FTP) data traffic over transmission control protocol (TCP).

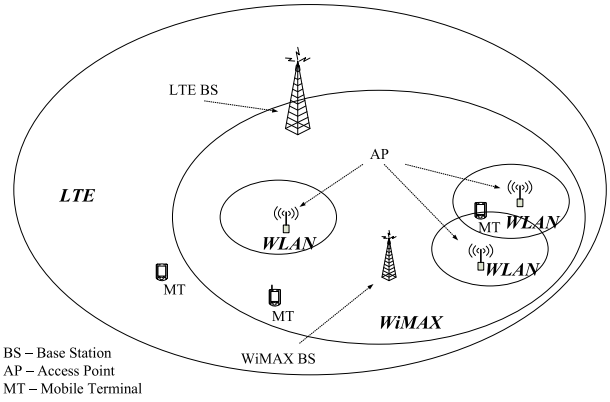


Fig. 1 Coverage areas of different networks.

Table 1 Summary of simulation parameters.

Parameter	Value
b^1_{max}, b^2_{max}, b^3_{max}	3, 4, 5 units
C_1, C_2, C_3	2, 1.5, 1.2
d^1_{max}, d^2_{max}, d^3_{max}	3, 3, 3 units
$K_{i,a}$	0.5
$\lvert\mathbf{M}\rvert$	3
V_{min}, V_{max}	1, 3 units
v_{max}	3 units
α	0.5
$\Theta_1, \Theta_2, \Theta_3$	50%, 25%, 25%
κ	0.5
μ	1 unit
$\rho_1 : \rho_2 : \rho_3$	1:1:8
σ	0.1 unit
ϕ	0.5

We consider the scenario depicted in Fig. 1. There are three networks in the system: network 1 is a cellular network, network 2 is a WiMAX network, and network 3 is a WLAN. For the simulation parameters, the unit of bandwidth is 16 kbps, the unit of delay is 60 *ms*, and the unit of the MT's velocity is 8 *km/hr*. The time between two successive decision epochs is 15 *sec*. The bandwidth importance weight ω is 0.25 for CBR traffic and 0.9 for FTP traffic. The reason is that CBR traffic is more sensitive to delay, while FTP traffic is elastic. Other simulation parameters are summarized in Table 1.

For the SAW and ELECTRE algorithms, the available bandwidth, delay, switching cost, and the MT's velocity are considered when calculating the policy. The importance weights for these parameters are consistent with those used in the CMDP model. Once the corresponding vertical handoff policies are calculated by the SAW and ELECTRE algorithms, the PIA is used to obtain the expected total reward achieved by each decision algorithm.

The probability q that determines the randomized optimal policy in (21) is calculated for different discount factors (i.e., different average connection durations). Specifically, for λ equal to [0.9, 0.95, 0.966, 0.975, 0.98], the correspond-

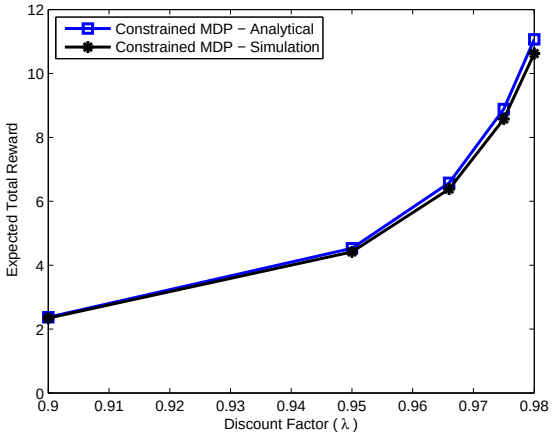


Fig. 2 Performance comparison between the results obtained from the analytical and simulation models.

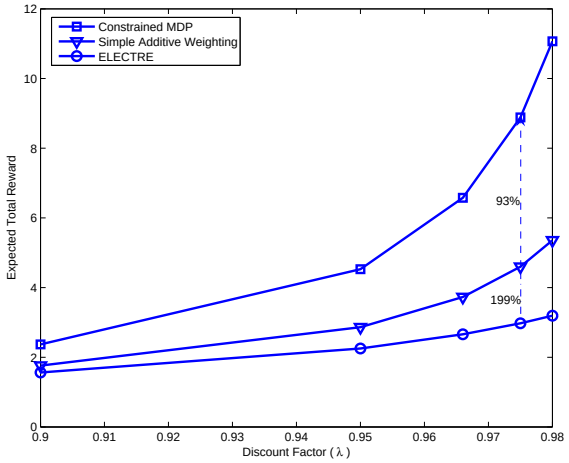


Fig. 3 Expected total reward under different discount factor λ for CBR traffic.

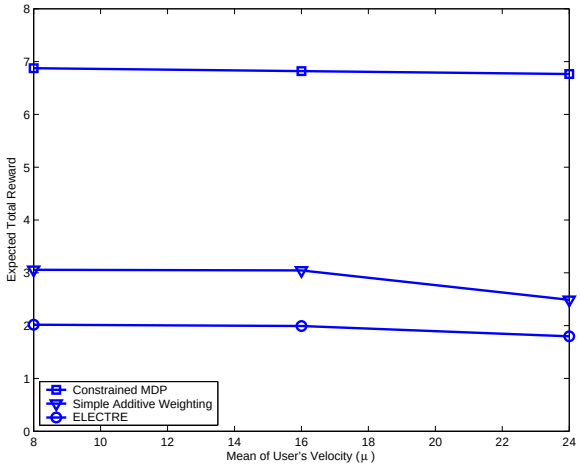


Fig. 4 Expected total reward under different mean of user's velocity μ for CBR traffic.

ing probabilities q are [0.61, 0.42, 0.53, 0.86, 0.46]. Moreover, the user's budget on the expected total access cost is also predefined for different discount factors. Specifically, for λ equal to [0.9, 0.95, 0.966, 0.975, 0.98], the predefined constraints C_{max} are [2, 4, 6, 8, 10].

5.1 Results for CBR Voice Traffic over UDP

For analytical model validation, a discrete event driven network simulator is created using C++. Simulations results are then compared with the results obtained from the analytical model. The simulation results are averaged over 100 simulation runs. The simulation time for each run is 250 *mins*. Fig. 2 compares the analytical and simulation results for the expected total reward obtained from the CMDP algorithm versus different discount factors (λ). Fig. 2 shows that the analytical results matched closely with the simulation results.

The expected total reward of a user under different discount factors for CBR traffic is shown in Fig. 3. For all the three schemes considered here, the expected total reward increases as λ becomes larger. This is because the larger λ is, the longer the average duration of the connection becomes. With the same constraint on the expected total access cost, the CMDP algorithm achieves the highest expected total reward among the four schemes. For example, when λ equals to 0.975, (i.e., the average duration of connection is 10 *mins*), for which the predefined constraint is 8 monetary units, the expected total reward from CMDP algorithm is 8.9. The expected total reward is obtained from the PIA and Q-learning algorithms (i.e., Algorithms 1 and 2), as well as equation (21). The CMDP algorithm achieves 93% higher expected total reward than the SAW scheme, and 199% higher expected total reward than the ELECTRE algorithm. SAW and ELECTRE algorithms achieve a lower reward than the CMDP scheme because they neither consider the user's budget for the connection, nor the long term effect of the action. In other words, SAW and ELECTRE choose actions only based on the instantaneous reward rather than the expected total reward.

Fig. 4 shows the expected total reward of a user versus the mean of its velocity under a budget of 8 monetary units for CBR traffic. As the user moves faster, the expected total reward of the CMDP algorithm decreases slightly. This is because the CMDP algorithm effectively avoids dropped calls by taking the user's velocity into consideration. For example, handoffs are only performed when the user's velocity is not likely to cause a dropped call. The SAW and ELECTRE algorithms still achieve a lower expected total reward.

The expected total reward a user can obtain versus its budget on the expected total access cost for CBR traffic is shown in Fig. 5. As the user's budget increases, the expected total reward becomes larger. This occurs because the more money that a user can spend on a connection, the more reward it will obtain. For the same budget, the CMDP al-

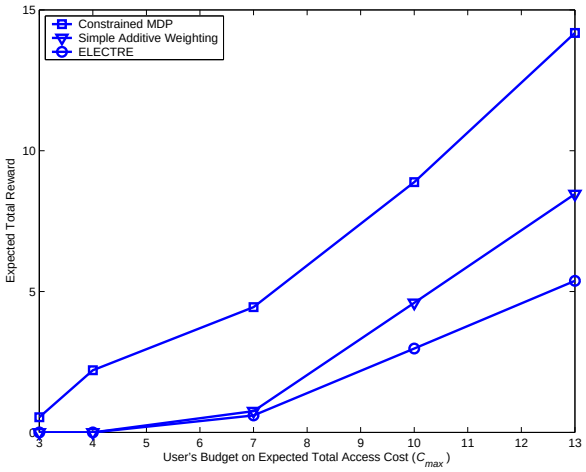


Fig. 5 Expected total reward under different user's budget on expected total access cost C_{max} for CBR traffic.

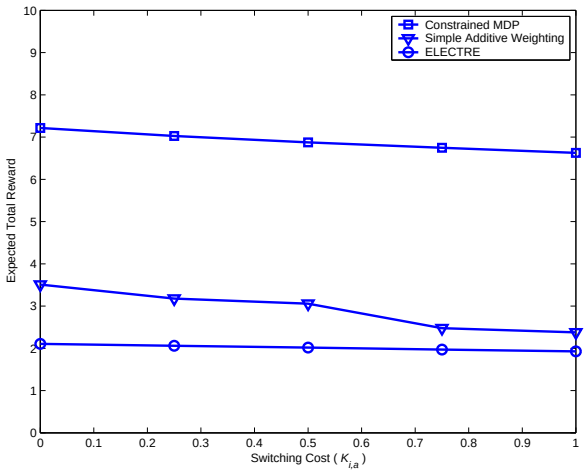


Fig. 6 Expected total reward under different switching cost $K_{i,a}$ for CBR traffic.

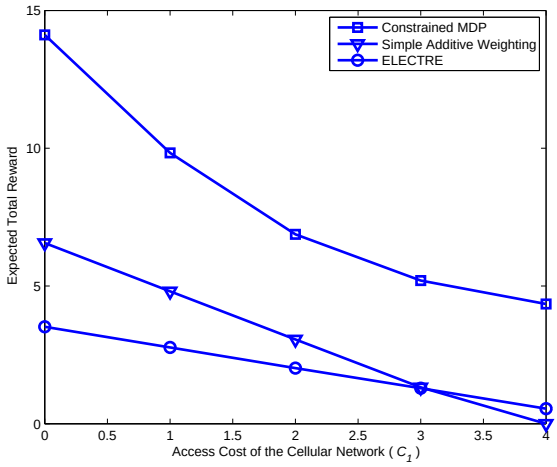


Fig. 7 Expected total reward under different access cost of the cellular network C_1 for CBR traffic.

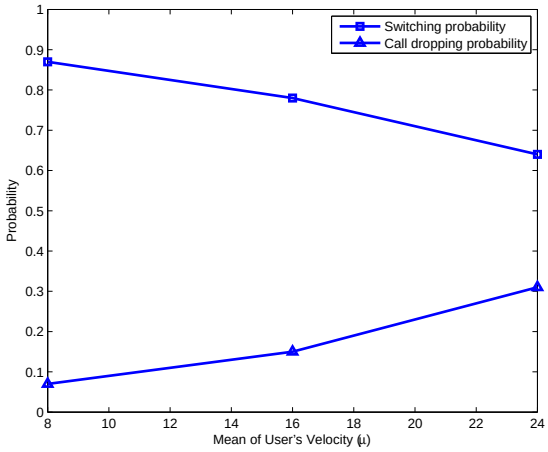


Fig. 8 Switching and dropping probabilities under different mean of user's velocity μ for CBR traffic.

gorithm always achieves higher reward than the SAW and ELECTRE schemes. The reason is that the CMDP algorithm can fully utilize the user's budget and avoid dropped calls to achieve the optimal reward, while the total reward obtained by the SAW and ELECTRE schemes are reduced because of the dropped connections.

Fig. 6 shows the expected total reward of users under different switching cost for CBR traffic. The budget used here is 8 monetary units. As we can see from the graph, when the switching cost (i.e, $K_{i,a}$) increases, the expected total reward of all three schemes decreases. For the same constraint on the expected total access cost, the CMDP scheme achieves better expected total reward than the SAW and ELECTRE schemes.

Fig. 7 shows the expected total reward of a user versus the access cost of the cellular network under a budget of 8 monetary units for CBR traffic. As C_1 increases (while C_2 and C_3 are fixed), the expected total reward becomes smaller for all three algorithms. The reason is that in order to take advantage of the cellular network, users need to pay more as the price of the cellular network increases. Thus, the expected total reward of the user decreases. For the same constraint on the expected total access cost, the CMDP scheme achieves higher expected total reward than the SAW and ELECTRE schemes.

In Fig. 8, we present the results for the switching and dropping probabilities of the MT in the WLAN. The switching probability is the probability that a user in the WLAN requests a vertical handoff at a decision epoch. The dropping probability is defined as the probability that the handoff request cannot be performed because of the high velocity of the MT. When the mean of user's velocity μ increases, the cost from the call dropping penalty function in (9) also increases. The MT will have a higher chance to remain in the existing network. Thus, the switching probability decreases. However, when μ increases and the MT chooses to perform vertical handoff, the probability that the vertical handoff can be completed before the MT losses the connection with the

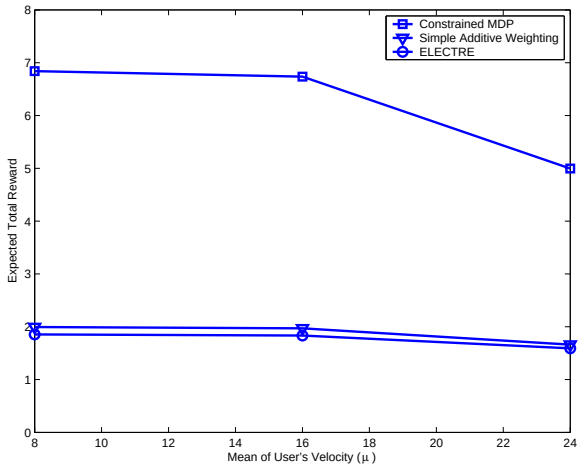


Fig. 9 Expected total reward under different mean of user's velocity μ for FTP traffic.

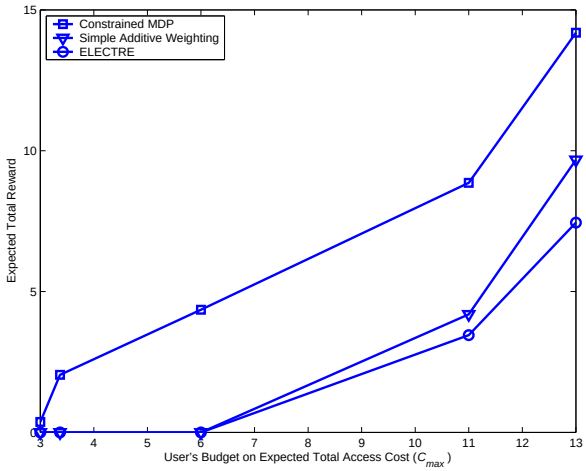


Fig. 10 Expected total reward under different user's budget on expected total access cost C_{max} for FTP traffic.

existing network decreases. Thus, the call dropping probability increases.

5.2 Results for FTP Data Traffic over TCP

Fig. 9 shows the expected total reward of a user versus the mean of its velocity under a budget of 8 monetary units for FTP traffic. Note when the MT's velocity is high (μ = 24 *km/h*), the expected total reward achieved by the CMDP algorithm with FTP traffic is smaller than that achieved with CBR traffic. The reason is when an MT moves faster, the probability that it can switch from the cellular network to WiMAX or WLAN is lower. As a result, it cannot take advantage of the high bandwidth in WiMAX and WLAN (which is crucial for FTP traffic that relies on high bandwidth); hence is not able to achieve a high expected total reward when it is moving fast.

The expected total reward a user can obtain versus its budget on the expected total access cost for FTP traffic is shown in Fig. 10. Similar to the CBR traffic case, we can see for the same budget, the CMDP algorithm always achieves higher expected total reward than the SAW and ELECTRE schemes. Note the expected total reward decreases dramatically when the user's budget is below 3.5 monetary units. This also happens in Fig. 5, however for CBR traffic this decrease is less noticeable.

5.3 Structure of the Optimal Policy

From Theorem 1, we showed that the optimal handoff policy is monotonically non-increasing in the bandwidth component of the current state **s**. Fig. 11 shows the structure of the unconstrained MDP optimal vertical handoff policy by varying the bandwidth of network 2 and network 3 (i.e., WiMAX and WLAN, respectively). The condition of the current state

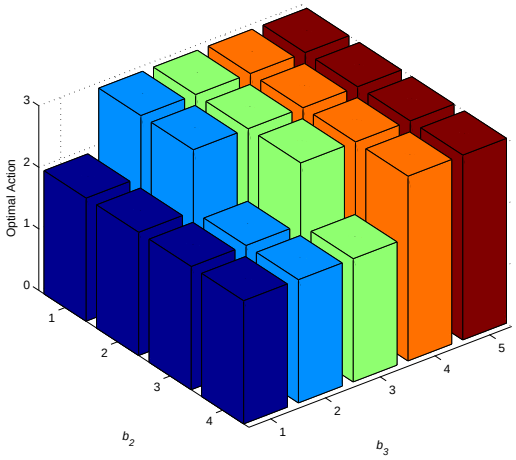


Fig. 11 Structure of the unconstrained MDP optimal vertical handoff policy on the available bandwidth when $\mathbf{s} = [i = 3, b_1 = 1, d_1 = 3, d_2 = 1, d_3 = 3, v = 2, l = 3]$.

s is $[i = 3, b_1 = 1, d_1 = 3, d_2 = 1, d_3 = 3, v = 2, l = 3]$. We can see for the threshold structure of the optimal policy, given a fixed value of b_2 (e.g., $b_2 = 3$), the optimal policy chooses network 3 when $b_3 \geq 3$, and selects network 2 when $b_3 < 3$. Thus, the threshold $\tau_b(i = 3, b_1 = 1, d_1 = 3, d_2 = 1, d_3 = 3, v = 2, l = 3)$ in (28) is equal 3. Similarly, for a fixed b_3 (e.g., $b_3 = 2$), the optimal policy chooses network 2 when $b_2 \geq 3$, and selects network 3 when $b_2 < 3$.

From Theorem 2, we showed that the optimal handoff policy is monotonically non-decreasing in the delay component of the current state **s**. Fig. 12 shows the structure of the unconstrained MDP optimal vertical handoff policy by varying the delay of networks 1 and 2 (i.e., cellular network and WiMAX, respectively). The condition of the current state **s** is $[i = 2, b_1 = 1, b_2 = 4, b_3 = 2, d_3 = 3, v = 1, l = 3]$. The threshold structure of the optimal policy shows that, for a fixed value of d_2 (e.g., $d_2 = 3$), the optimal policy chooses network 2 when $d_1 \geq 3$, and selects network 1 when $d_1 < 3$. Thus, the threshold $\tau_d(i = 2, b_1 = 1, b_2 = 4, b_3 = 2, d_3 = 3, v = 1, l = 3)$ in (30) is equal 3. Similarly, for a fixed d_1

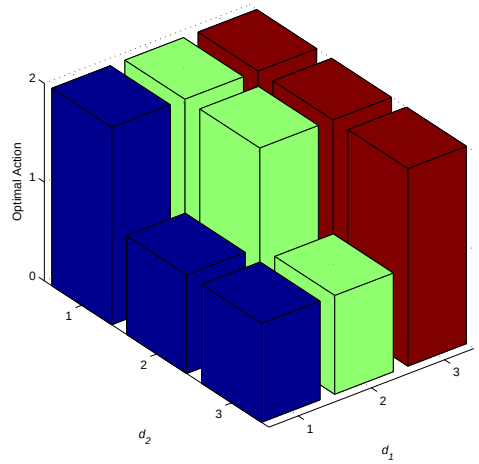


Fig. 12 Structure of the unconstrained MDP optimal vertical handoff policy on the packet delay when $\mathbf{s} = [i = 2, b_1 = 1, b_2 = 4, b_3 = 2, d_3 = 3, v = 1, l = 3]$.

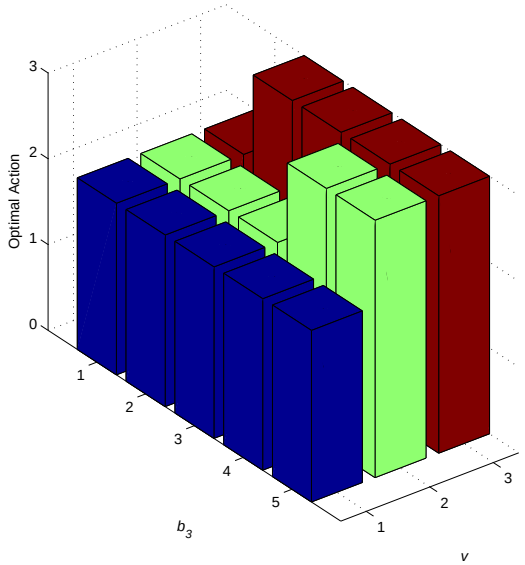


Fig. 13 Structure of the unconstrained MDP optimal vertical handoff policy on the mean of user's velocity when $\mathbf{s} = [i = 3, b_1 = 3, d_1 = 3, b_2 = 4, d_2 = 1, d_3 = 3, l = 3]$.

(e.g., $d_1 = 2$), the optimal policy chooses network 1 when $d_2 \geq 3$, and selects network 2 when $d_2 < 3$. The network and the MT only need to store the threshold values.

From Theorem 3, we showed that the optimal handoff policy is monotonically non-decreasing in the velocity component of the current state $\mathbf{s}$. Fig. 13 shows the structure of the unconstrained MDP optimal vertical handoff policy by varying the bandwidth of network 3 (i.e., WLAN) and the mean value of the MT's velocity. The current state $\mathbf{s}$ is $[i = 3, b_1 = 3, d_1 = 3, b_2 = 4, d_2 = 1, d_3 = 3, l = 3]$. We can see for the threshold structure of the optimal policy, given a fixed value of b_3 (e.g., $b_3 = 4$), the optimal policy chooses not to perform handoff when $v \geq 2$, and chooses to perform handoff when $v < 2$. Thus, the threshold $\tau_v(i = 3, b_1 = 3, d_1 = 3, b_2 = 4, d_2 = 1, d_3 = 3, l = 3)$ in (32) is equal 2. If $b_3 = 2$, the optimal policy does not perform vertical handoff when $v \geq 3$, and chooses to perform handoff when $v < 3$. The threshold policy simplifies the implementation of the proposed algorithm. Instead of storing the optimal policy of all possible states, only the threshold values need to be stored. The vertical handoff decision can be performed by a simple lookup table.

6 Conclusions

In this paper, we proposed a CMDP-based vertical handoff decision algorithm for 4G heterogeneous wireless networks. Our work considered the connection duration, the available bandwidth and delay of the candidate networks, MT's velocity and location information, signaling load incurred on the network, network access cost, user's preference, and user's monetary budget for the vertical handoff decision. The algorithm is based on the CMDP formulation with the objective of maximizing the expected total reward of a connection. The constraint of the problem is that users have monetary budgets for their connections. By using the PIA and Q-learning algorithm, a stationary randomized policy is obtained when the connection termination time is geometrically distributed. Structural results on the optimal vertical handoff policy are derived by using the concept of supermodularity. We showed that the optimal policy is a threshold policy in the available bandwidth, delay, and velocity. Numerical results showed that the proposed CMDP-based vertical handoff decision algorithm outperforms other decision schemes in a wide range of conditions such as variations on connection duration, user's velocity, user's budget, traffic type, signaling cost, and monetary access cost. For future work, we plan to consider other constraints in the problem formulation.

Acknowledgment

This work was supported by Bell Canada and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Appendix A. Proof of Lemma 1

We will prove $v_\beta(i, b_1, d_1, b_2, d_2, v, l)$ is monotone in available bandwidth, delay, and velocity. This consists of two steps:

1. To prove the reward function $r(\mathbf{s}, a)$ is monotone in available bandwidth, delay, and velocity;
2. To prove the sum of transition probabilities $\sum_{\mathbf{s}' \in \mathbf{S}} P[\mathbf{s}' \mid \mathbf{s}, a]$ is monotone in available bandwidth, delay, and velocity.

We first note that the only part that relates to the bandwidth in the reward function (i.e., $r(\mathbf{s}, a)$) is $f_b(\mathbf{s}, a)$. Let b_a^1 and b_a^2 be two possible bandwidth values, and $b_a^1 \geq b_a^2$. We denote $f_b(\mathbf{s}^1, a)$ as the value of $f_b(\mathbf{s}, a)$ when $b_a = b_a^1$, and $f_b(\mathbf{s}^2, a)$ as the value of $f_b(\mathbf{s}, a)$ when $b_a = b_a^2$. Clearly from the definition of $f_b(\mathbf{s}, a)$, $f_b(\mathbf{s}^1, a)$ is greater than (or equal to) $f_b(\mathbf{s}^2, a)$, since $f_b(\mathbf{s}, a)$ is linearly proportional to b_a. As a result, the reward function $r(\mathbf{s}, a)$ is monotonically non-decreasing in the available bandwidth.

Similarly, the only part that relates to the delay in the reward function is $f_d(\mathbf{s},a)$. Let d_a^1 and d_a^2 be two possible delay values, and $d_a^1 \geq d_a^2$. We denote $f_d(\mathbf{s}^1,a)$ as the value of $f_d(\mathbf{s},a)$ when $d_a = d_a^1$, and $f_d(\mathbf{s}^2,a)$ as the value of $f_d(\mathbf{s},a)$ when $d_a = d_a^2$. Clearly from the definition of $f_d(\mathbf{s},a)$, $f_d(\mathbf{s}^1,a)$ is smaller than (or equal to) $f_d(\mathbf{s}^2,a)$, since $f_d(\mathbf{s},a)$ is linear inverse proportional to d_a. As a result, the reward function $r(\mathbf{s},a)$ is monotonically non-increasing in the delay.

For the velocity, the only part that relates to it in the reward function is $-q(\mathbf{s},a)$. From the definition of $q(\mathbf{s},a)$ in (9), when the velocity v becomes larger, the value of $q(\mathbf{s},a)$ becomes larger or remains the same, which means that $-q(\mathbf{s},a)$ becomes smaller or stays the same. Consequently, the reward function $r(\mathbf{s},a)$ is monotonically non-increasing in velocity.

We assume that the transition probability function $P[\mathbf{s}' \mid \mathbf{s},a]$ satisfies the *first order stochastic dominance condition*. This implies when the system is in a better state (e.g., larger bandwidth, lower delay), its evolution will be in the region of better states with a higher probability. When the available bandwidth is considered, it implies that the sum of transition probabilities (i.e., $\sum_{\mathbf{s}'\in\mathbf{S}} P[\mathbf{s}' \mid \mathbf{s},a]$) is monotonically non-decreasing in the available bandwidth. Similarly, the delay (velocity) in the next decision epoch is stochastically decreasing with respect to the delay (velocity) in the current decision epoch is the condition under which the sum of transition probabilities (i.e., $\sum_{\mathbf{s}'\in\mathbf{S}} P[\mathbf{s}' \mid \mathbf{s},a]$) is monotonically non-increasing in the delay (velocity). ■

Appendix B. Proof of Theorem 1

To show that the optimal policy is monotonically non-increasing in the available bandwidth, we need to prove that $Q_\beta(\mathbf{s},a)$ is submodular in (b_2,a). We will prove via mathematical induction that for a suitable initialization,

$$\begin{aligned}
&Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],2) - Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],1)\\
&= r([i,b_1,d_1,b_2,d_2,v,l],2;\beta) - r([i,b_1,d_1,b_2,d_2,v,l],1;\beta)\\
&\quad + \sum_{\mathbf{s}'\in\mathbf{S}} \lambda P[b_1',d_1' \mid b_1,d_1]P[b_2',d_2' \mid b_2,d_2]P[v' \mid v]P[l' \mid l]\\
&\qquad \times \left(v_\beta^k(j,b_1',d_1',b_2',d_2',v',l') - v_\beta^k(j,b_1',d_1',b_2',d_2',v',l')\right), \quad (34)
\end{aligned}$$

is monotonically non-increasing in the available bandwidth b_2. It holds if $v_\beta(j,b_1',d_1',b_2',d_2',v',l')$ has non-increasing difference in b_2. Select $v_\beta^0(j,b_1',d_1',b_2',d_2',v',l')$ with non-increasing difference in b_2. Assume that $v_\beta^k(j,b_1',d_1',b_2',d_2',v',l')$ has non-increasing difference in b_2, which implies that $Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a)$ is submodular in (b_2,a). We will now prove that $v_\beta^{k+1}(j,b_1',d_1',b_2',d_2',v',l')$ also has non-increasing difference in b_2. That is,

$$\begin{aligned}
&v_\beta^{k+1}(i,b_1,d_1,b_2+1,d_2,v,l) - v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)\\
&\leq v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l) - v_\beta^{k+1}(i,b_1,d_1,b_2-1,d_2,v,l),
\end{aligned} \quad (35)$$

or

$$\begin{aligned}
&v_\beta^{k+1}(i,b_1,d_1,b_2+1,d_2,v,l) - v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)\\
&-\left(v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l) - v_\beta^{k+1}(i,b_1,d_1,b_2-1,d_2,v,l)\right) \leq 0. \quad (36)
\end{aligned}$$

We assume

$$v_\beta^{k+1}(i,b_1,d_1,b_2+1,d_2,v,l) = Q_\beta^{k+1}([i,b_1,d_1,b_2+1,d_2,v,l],a_2),$$

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l) = Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1),$$

and

$$v_\beta^{k+1}(i,b_1,d_1,b_2-1,d_2,v,l) = Q_\beta^{k+1}([i,b_1,d_1,b_2-1,d_2,v,l],a_0),$$

for some actions $a_2,a_1,a_0 \in \mathbf{A_s}$. Thus, we can re-write (35) as

$$\begin{aligned}
&Q_\beta^{k+1}([i,b_1,d_1,b_2+1,d_2,v,l],a_2) - Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)\\
&-\left(Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1) - Q_\beta^{k+1}([i,b_1,d_1,b_2-1,d_2,v,l],a_0)\right)\\
&\leq 0, \quad (37)
\end{aligned}$$

or

$$\begin{aligned}
&\underbrace{Q_\beta^{k+1}([i,b_1,d_1,b_2+1,d_2,v,l],a_2) - Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_2)}_{W_1}\\
&\underbrace{+Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_2) - Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)}_{X_1}\\
&\underbrace{-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1) + Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_0)}_{Y_1}\\
&-\Big(\underbrace{Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_0) - Q_\beta^{k+1}([i,b_1,d_1,b_2-1,d_2,v,l],a_0)}_{Z_1}\Big)\\
&\leq 0,
\end{aligned}$$

where $X_1 \leq 0$ and $Y_1 \leq 0$ by optimality. Note that in X_1 and Y_1 the optimal action is a_1.

In addition, it follows from the induction hypothesis that

$$\begin{aligned}
W_1 &= r([i,b_1,d_1,b_2+1,d_2,v,l],a_2;\beta) - r([i,b_1,d_1,b_2,d_2,v,l],a_2;\beta)\\
&\quad + \sum_{\mathbf{s}'\in\mathbf{S}} \lambda \Big(P[b_1',d_1' \mid b_1,d_1]P[b_2',d_2' \mid b_2+1,d_2]P[v' \mid v]P[l' \mid l]\\
&\quad \times v_\beta^k(j,b_1',d_1',(b_2+1)',d_2',v',l') - P[b_1',d_1' \mid b_1,d_1]\\
&\quad \times P[b_2',d_2' \mid b_2,d_2]P[v' \mid v]P[l' \mid l]v_\beta^k(j,b_1',d_1',b_2',d_2',v',l')\Big)\\
&\leq r([i,b_1,d_1,b_2,d_2,v,l],a_0;\beta) - r([i,b_1,d_1,b_2-1,d_2,v,l],a_0;\beta)\\
&\quad + \sum_{\mathbf{s}'\in\mathbf{S}} \lambda \Big(P[b_1',d_1' \mid b_1,d_1]P[b_2',d_2' \mid b_2,d_2]P[v' \mid v]P[l' \mid l]\\
&\quad \times v_\beta^k(j,b_1',d_1',b_2',d_2',v',l') - P[b_1',d_1' \mid b_1,d_1]P[b_2',d_2' \mid b_2-1,d_2]\\
&\quad \times P[v' \mid v]P[l' \mid l]v_\beta^k(j,b_1',d_1',(b_2-1)',d_2',v',l')\Big).
\end{aligned}$$

The right-hand side (RHS) of the inequality comes from the expansion of Z_1 which implies that $W_1 \leq Z_1$. Therefore, it is shown that $v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)$ satisfies (35), which implies that $Q_\beta(\mathbf{s},a)$ is submodular in (b_2,a). ■

Appendix C. Proof of Theorem 2

To show that the optimal policy is monotonically non-decreasing in the delay, we need to prove that $Q_\beta(\mathbf{s},a)$ is supermodular in (d_2,a). We will prove via mathematical induction that, for a suitable initialization, (34) is monotonically non-decreasing in the delay d_2. The above holds if $v_\beta(j,b_1',d_1',b_2',d_2',v',l')$ has non-increasing difference in d_2. Select $v_\beta^0(j,b_1',d_1',b_2',d_2',v',l')$ with non-increasing difference in d_2. Assume that $v_\beta^k(j,b_1',d_1',b_2',d_2',v',l')$ has non-increasing difference in d_2, which implies that $Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a)$ is supermodular in (d_2,a). We will now prove that $v_\beta^{k+1}(j,b_1',d_1',b_2',d_2',v',l')$ also has non-increasing difference in d_2. That is,

$$\begin{aligned}
&v_\beta^{k+1}(i,b_1,d_1,b_2,d_2+1,v,l) - v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)\\
&\leq v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l) - v_\beta^{k+1}(i,b_1,d_1,b_2,d_2-1,v,l),
\end{aligned} \quad (38)$$

or

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2+1,v,l)-v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)$$
$$-\Big(v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)-v_\beta^{k+1}(i,b_1,d_1,b_2,d_2-1,v,l)\Big)\le 0. \quad (39)$$

We assume

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2+1,v,l)=Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2+1,v,l],a_2),$$

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)=Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1),$$

and

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2-1,v,l)=Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2-1,v,l],a_0),$$

for some actions $a_2,a_1,a_0\in\mathbf{A_s}$. Thus, we can re-write (38) as

$$Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2+1,v,l],a_2)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)$$
$$-\Big(Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)$$
$$-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2-1,v,l],a_0)\Big)\le 0, \quad (40)$$

or

$$\underbrace{Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2+1,v,l],a_2)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_2)}_{W_2}$$
$$\underbrace{+Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_2)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)}_{X_2}$$
$$\underbrace{-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)+Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_0)}_{Y_2}$$
$$-\Big(\underbrace{Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_0)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2-1,v,l],a_0)}_{Z_2}\Big)$$
$$\le 0,$$

where $X_2\le 0$ and $Y_2\le 0$ by optimality. Note that in X_2 and Y_2 the optimal action is a_1.

In addition, it follows from the induction hypothesis that

$$\begin{aligned}W_2 &= r([i,b_1,d_1,b_2,d_2+1,v,l],a_2;\beta)-r([i,b_1,d_1,b_2,d_2,v,l],a_2;\beta)\\ &\quad+\sum_{\mathbf{s}'\in\mathbf{S}}\lambda\Big(P[b_1',d_1'\mid b_1,d_1]P[b_2',d_2'\mid b_2,d_2+1]P[v'\mid v]P[l'\mid l]\\ &\quad\times v_\beta^k(j,b_1',d_1',b_2',(d_2+1)',v',l')-P[b_1',d_1'\mid b_1,d_1]\\ &\quad\times P[b_2',d_2'\mid b_2,d_2]P[v'\mid v]P[l'\mid l]v_\beta^k(j,b_1',d_1',b_2',d_2',v',l')\Big)\\ &\le r([i,b_1,d_1,b_2,d_2,v,l],a_0;\beta)-r([i,b_1,d_1,b_2,d_2-1,v,l],a_0;\beta)\\ &\quad+\sum_{\mathbf{s}'\in\mathbf{S}}\lambda\Big(P[b_1',d_1'\mid b_1,d_1]P[b_2',d_2'\mid b_2,d_2]P[v'\mid v]P[l'\mid l]\\ &\quad\times v_\beta^k(j,b_1',d_1',b_2',d_2',v',l')-P[b_1',d_1'\mid b_1,d_1]\\ &\quad\times P[b_2',d_2'\mid b_2,d_2-1]P[v'\mid v]P[l'\mid l]\\ &\quad\times v_\beta^k(j,b_1',d_1',b_2',(d_2-1)',v',l')\Big).\end{aligned}$$

The RHS of the inequality comes from the expansion of Z_2 which implies that $W_2\le Z_2$. Therefore, it is shown that $v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)$ satisfies (38), which implies that $Q_\beta(\mathbf{s},a)$ is supermodular in (d_2,a). ■

Appendix D. Proof of Theorem 3

To show that the optimal policy is monotonically non-decreasing in the velocity, we need to prove that $Q_\beta(\mathbf{s},a)$ is supermodular in (v,a). We will prove via mathematical induction that for a suitable initialization, (34) is monotonically non-decreasing in the velocity v. It holds if $v_\beta(j,b_1',d_1',b_2',d_2',v',l')$ has non-increasing difference in v. Select $v_\beta^0(j,b_1',d_1',b_2',d_2',v',l')$ with non-increasing difference in v. Assume that $v_\beta^k(j,b_1',d_1',b_2',d_2',v',l')$ has non-increasing difference in v, which implies that $Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a)$ is supermodular in (v,a). We will now prove that $v_\beta^{k+1}(j,b_1',d_1',b_2',d_2',v',l')$ also has non-increasing difference in v. That is,

$$\begin{aligned}&v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v+1,l)-v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)\\ &\le v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)-v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v-1,l),\end{aligned} \quad (41)$$

or

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v+1,l)-v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)$$
$$-\Big(v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)-v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v-1,l)\Big)\le 0. \quad (42)$$

We assume

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v+1,l)=Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v+1,l],a_2),$$

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)=Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1),$$

and

$$v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v-1,l)=Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v-1,l],a_0),$$

for some actions $a_2,a_1,a_0\in\mathbf{A_s}$. Thus, we can re-write (41) as

$$Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v+1,l],a_2)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)$$
$$-\Big(Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)$$
$$-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v-1,l],a_0)\Big)\le 0, \quad (43)$$

or

$$\underbrace{Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v+1,l],a_2)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_2)}_{W_3}$$
$$\underbrace{+Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_2)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)}_{X_3}$$
$$\underbrace{-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_1)+Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_0)}_{Y_3}$$
$$-\Big(\underbrace{Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v,l],a_0)-Q_\beta^{k+1}([i,b_1,d_1,b_2,d_2,v-1,l],a_0)}_{Z_3}\Big)$$
$$\le 0,$$

where $X_3\le 0$ and $Y_3\le 0$ by optimality. Note that in X_3 and Y_3 the optimal action is a_1.

In addition, it follows from the induction hypothesis that

$$
\begin{aligned}
W_3 = {} & r([i,b_1,d_1,b_2,d_2,v+1,l],a_2;\beta) - r([i,b_1,d_1,b_2,d_2,v,l],a_2;\beta) \\
& + \sum_{\mathbf{s}'\in\mathbf{S}} \lambda \Big(P[b_1',d_1' \mid b_1,d_1]P[b_2',d_2' \mid b_2,d_2]P[v' \mid v+1]P[l' \mid l] \\
& \times v_\beta^k(j,b_1',d_1',b_2',d_2',(v+1)',l') - P[b_1',d_1' \mid b_1,d_1] \\
& \times P[b_2',d_2' \mid b_2,d_2]P[v' \mid v]P[l' \mid l]v_\beta^k(j,b_1',d_1',b_2',d_2',v',l') \Big) \\
\leq {} & r([i,b_1,d_1,b_2,d_2,v,l],a_0;\beta) - r([i,b_1,d_1,b_2,d_2,v-1,l],a_0;\beta) \\
& + \sum_{\mathbf{s}'\in\mathbf{S}} \lambda \Big(P[b_1',d_1' \mid b_1,d_1]P[b_2',d_2' \mid b_2,d_2]P[v' \mid v]P[l' \mid l] \\
& \times v_\beta^k(j,b_1',d_1',b_2',d_2',v',l') - P[b_1',d_1' \mid b_1,d_1]P[b_2',d_2' \mid b_2,d_2] \\
& \times P[v' \mid v-1]P[l' \mid l]v_\beta^k(j,b_1',d_1',b_2',d_2',(v-1)',l') \Big).
\end{aligned}
$$

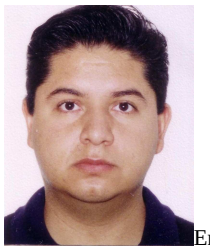
The RHS of the inequality comes from the expansion of Z_3 which implies that $W_3 \leq Z_3$. Therefore, it is shown that $v_\beta^{k+1}(i,b_1,d_1,b_2,d_2,v,l)$ satisfies (41), which implies that $Q_\beta(\mathbf{s},a)$ is supermodular in (v,a). ■

References

1. 3rd Generation Partnership Project (3GPP), http://www.3gpp.org.
2. 3rd Generation Partnership Project 2 (3GPP2), http://www.3gpp2.org.
3. IEEE 802.21 Media Independent Handover Working Group, http://www.ieee802.org/21/.
4. J. McNair and F. Zhu, "Vertical Handoffs in Fourth-generation Multi-network Environments," *IEEE Wireless Communications*, vol. 11, no. 3, pp. 8–15, June 2004.
5. W. Chen, J. Liu, and H. Huang, "An Adaptive Scheme for Vertical Handoff in Wireless Overlay Networks," in *Proc. of ICPAD'04*, Newport Beach, CA, July 2004.
6. L. Xia, L. G. Jiang, and C. He, "A Novel Fuzzy Logic Vertical Handoff Algorithm with Aid of Differential Prediction and Pre-Decision Method," in *Proc. of IEEE ICC'07*, Glasgow, Scotland, June 2007.
7. N. Nasser, S. Guizani, and E. Al-Masri, "Middleware Vertical Handoff Manager: A Neural Network-based Solution," in *Proc. of IEEE ICC'07*, Glasgow, Scotland, June 2007.
8. M. Mani and N. Crespi, "Handover Criteria Considerations in Future Convergent Networks," in *Proc. of IEEE GLOBECOM'06*, San Francisco, CA, November 2006.
9. A. V. Garmonov, S. H. Cheon, D. H. Yim, K. T. Han, Y. S. Park, A. Y. Savinkov, S. A. Filin, S. N. Moiseev, and M. S. Kondakov, "QoS-Oriented Intersystem Handover between IEEE 802.11b and Overlay Networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1142–1154, March 2008.
10. W. Zhang, "Handover Decision Using Fuzzy MADM in Heterogeneous Networks," in *Proc. of IEEE WCNC'04*, Atlanta, GA, March 2004.
11. F. Bari and V. Leung, "Application of ELECTRE to Network Selection in a Heterogeneous Wireless Network Environment," in *Proc. of IEEE WCNC'07*, Hong Kong, China, March 2007.
12. K. Yang, I. Gondal, B. Qiu, and L. S. Dooley, "Combined SINR Based Vertical Handoff Algorithm for Next Generation Heterogeneous Wireless Networks," in *Proc. of IEEE GLOBECOM'07*, Washington, DC, November 2007.
13. M. Liu, Z. Li, X. Guo, and E. Dutkiewicz, "Performance Analysis and Optimization of Handoff Algorithms in Heterogeneous Wireless Networks," *IEEE Transactions on Mobile Computing*, vol. 7, no. 7, pp. 846–857, July 2008.
14. S. Chien, H. Liu, A. L. Y. Low, C. Maciocco, and Y. Ho, "Smart Predictive Trigger for Effective Handover in Wireless Networks," in *Proc. of IEEE ICC'08*, Beijing, China, May 2008.
15. O. Ormond, J. Murphy, and G. Muntean, "Utility-based Intelligent Network Selection in Beyond 3G Systems," in *Proc. of IEEE ICC'06*, Istanbul, Turkey, June 2006.
16. J. Zhang, H. C. Chan, and V. Leung, "A Location-Based Vertical Handoff Decision Algorithm for Heterogeneous Mobile Networks," in *Proc. of IEEE GLOBECOM'06*, San Francisco, CA, November 2006.
17. Q. Guo, J. Zhu, and X. Xu, "An Adaptive Multi-criteria Vertical Handoff Decision Algorithm for Radio Heterogeneous Networks," in *Proc. of IEEE ICC'05*, Seoul, Korea, May 2005.
18. W. Lee, E. Kim, J. Kim, I. Lee, and C. Lee, "Movement-Aware Vertical Handoff of WLAN and Mobile WiMAX for Seamless Ubiquitous Access," *IEEE Transactions on Consumers Electronics*, vol. 53, no. 4, pp. 1268–1275, November 2007.
19. A. Zahran and B. Liang, "Performance Evaluation Framework for Vertical Handoff Algorithms in Heterogeneous Networks," in *Proc. of IEEE ICC'05*, Seoul, Korea, May 2005.
20. J. Wang, R. V. Prasad, and I. Niemegeers, "Solving the Incertitude of Vertical Handovers in Heterogeneous Mobile Wireless Network Using MDP," in *Proc. of IEEE ICC'08*, Beijing, China, May 2008.
21. E. Stevens-Navarro, Y. Lin, and V. W. S. Wong, "An MDP-based Vertical Handoff Decision Algorithm for Heterogeneous Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 2, pp. 1243–1254, March 2008.
22. C. Sun, E. Stevens-Navarro, and V. Wong, "A Constrained MDP-based Vertical Handoff Decision Algorithm for 4G Wireless Networks," in *Proc. of IEEE ICC'08*, Beijing, China, May 2008.
23. C. Sun, "A Constrained MDP-based Vertical Handoff Decision Algorithm for Wireless Networks," Master's thesis, University of British Columbia, Vancouver, BC, Canada, July 2008.
24. E. Altman, *Constrained Markov Decision Processes*. Chapman and Hall, 1999.
25. B. Liang and Z. Haas, "Predictive Distance-Based Mobility Management for Multidimensional PCS Networks," *IEEE/ACM Transactions on Networking*, vol. 11, no. 5, pp. 718–732, October 2003.
26. S. Tang and W. Li, "Performance Analysis of the 3G Network with Complementary WLANs," in *Proc. of IEEE GLOBECOM'05*, St. Louis, MO, November 2005.
27. A. Doufexi, E. Tameh, A. Nix, S. Armour, and A. Molina, "Hotspot Wireless LANs to Enhance the Performance of 3G and Beyond Cellular Networks," *IEEE Communications Magazine*, vol. 41, no. 7, pp. 58–65, July 2003.
28. The Network Simulator - ns-2, http://www.isi.edu/nsnam/ns.
29. WiMAX Module for ns-2 Simulator, http://ndsl.csie.cgu.edu.tw/wimaxns2.php.
30. M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley and Sons, 1994.
31. D. P. Bertsekas, *Dynamic Programming and Optimal Control, Third Edition*. Athena Scientific, 2007.
32. V. Djonin and V. Krishnamurthy, "Q-Learning Algorithms for Constrained Markov Decision Process with Randomized Monotone Policies: Application to MIMO Transmission Control," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2170–2181, May 2007.
33. F. Beutler and K. Ross, "Optimal Policies for Controlled Markov Chains with a Constraint," *J. Math. Anal. Appl.*, vol. 112, pp. 236–252, 1985.
34. D. M. Topkis, *Supermodularity and Complementary*. Princeton University Press, 1998.

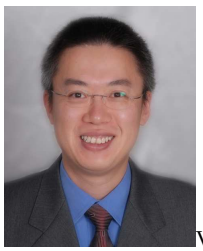
Biographies

Chi Sun received his B.Sc. degree from the University of Alberta in 2006, and M.A.Sc. degree from the University of British Columbia in 2008. He is currently working in China. His research interests are in the area of mobility management for wireless networks.



Enrique Stevens-Navarro received the B.Sc. degree from Universidad Autonoma de San Luis Potosi (UASLP), San Luis Potosi, Mexico in 2000, the M.Sc. degree from Instituto Tecnologico y de Estudios Superiores de Monterrey (ITESM), Monterrey, Mexico in 2002, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, Canada in 2008, all in electrical engineering. Currently, he is an Assistant Professor at the Faculty of Science of Universidad Autonoma de San Luis Potosi (UASLP), in San Luis Potosi, Mexico. His research interests are in mobility and resource management for heterogeneous wireless networks.

Vahid Shah-Mansouri received the B.Sc. and M.Sc. degrees in electrical engineering from University of Tehran, Tehran, Iran in 2003 and Sharif University of Technology, Tehran, Iran in 2005, respectively. From 2005 to 2006, he was with Farineh-Fanavar Co., Tehran, Iran. He is currently working towards the Ph.D. degree in the Department of Electrical and Computer Engineering at the University of British Columbia (UBC), Vancouver, BC, Canada. As a graduate student, he received the UBC Four Year Fellowship and UBC Faculty of Applied Science Award. His research interests are in design and mathematical modeling of radio frequency identification (RFID) systems and wireless networks.



Vincent W.S. Wong received the B.Sc. 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. From 2000 to 2001, he worked as a systems engineer at PMC-Sierra Inc. He joined the Department of Electrical and Computer Engineering at UBC in 2002 and is currently an Associate Professor. His research areas include protocol design, optimization, and resource management of communication networks, with applications to the Internet, wireless networks, smart grid, RFID systems, and intelligent transportation systems. Dr. Wong is an Associate Editor of the IEEE Transactions on Vehicular Technology and an Editor of KICS/IEEE Journal of Communications and Networks. He is the Symposium Co-chair of IEEE Globecom'11, Wireless Communications Symposium. He serves as TPC member in various conferences, including IEEE Infocom and ICC. He is a senior member of the IEEE.