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

by

Chi Sun

B.Sc., Electrical Engineering, University of Alberta, Canada, 2006

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

( Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

July 2008

© Chi Sun, 2008
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 thesis 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 the 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.
Contents

Abstract ............................................................... ii

Contents ............................................................. iv

List of Tables ......................................................... vii

List of Figures ......................................................... viii

List of Acronyms ..................................................... x

List of Symbols ...................................................... xiii

Acknowledgements ................................................... xvii

1 Introduction ......................................................... 1

1.1 Related Work ................................................... 5

1.2 Motivations and Objectives .................................... 13

1.3 Contributions .................................................. 15

1.4 Structure of the Thesis ......................................... 16
# Contents

2 CMDP Formulation of the Vertical Handoff Decision Problem ........................................ 17
   2.1 CMDP Formulation and System Model ................................................................. 17
      2.1.1 States, Actions, and Transition Probabilities .............................................. 20
      2.1.2 Rewards ........................................................................................................ 26
      2.1.3 Costs ............................................................................................................ 29
   2.2 CMDP Optimality Equations .................................................................................... 30
   2.3 Monotone Optimal Vertical Handoff Policy ......................................................... 35
      2.3.1 Threshold Structure of Unconstrained MDP .................................................. 36
      2.3.2 Threshold Structure of Constrained MDP ...................................................... 40

3 Numerical Results and Discussions ........................................................................... 43
   3.1 Comparison between Different Vertical Handoff Decision Algorithms ............... 43
      3.1.1 Results for CBR Voice Traffic over UDP ...................................................... 46
      3.1.2 Results for FTP Data Traffic over TCP ....................................................... 52
   3.2 Structure of the Optimal Policy ............................................................................ 55

4 Conclusions and Future Work .................................................................................... 58
   4.1 Conclusions .......................................................................................................... 58
   4.2 Future Work .......................................................................................................... 59

Bibliography .................................................................................................................. 60

A Proofs of Lemma 1 and Theorems 1, 2, and 3 .............................................................. 67
   A.1 Proof of Lemma 1 .................................................................................................. 67
<table>
<thead>
<tr>
<th>Section</th>
<th>Proof of Theorem 1</th>
<th>Proof of Theorem 2</th>
<th>Proof of Theorem 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>69</td>
<td>71</td>
<td>73</td>
</tr>
</tbody>
</table>
List of Tables

3.1 Summary of simulation parameters. ......................... 44
List of Figures

2.1 Coverage area of different networks. ........................................... 20

3.1 Expected total reward under different discount factor $\lambda$ for CBR traffic. ... 47

3.2 Expected total reward under different mean of user’s velocity $\mu$ for CBR traffic. ................................................................. 48

3.3 Expected total reward under different user’s budget on expected total access cost $C_{\text{max}}$ for CBR traffic. ........................................... 49

3.4 Expected total reward under different switching cost $K_{i,a}$ for CBR traffic. ... 50

3.5 Expected total reward under different access cost of the cellular network $C_1$ for CBR traffic. ................................................................. 51

3.6 Expected total reward under different mean of user’s velocity $\mu$ for FTP traffic. ................................................................. 52

3.7 Expected total reward under different user’s budget on expected total access cost $C_{\text{max}}$ for FTP traffic. ........................................... 53

3.8 Expected total reward under different access cost of the cellular network $C_1$ for FTP traffic. ................................................................. 54
3.9 Structure of the unconstrained MDP optimal vertical handoff policy on the available bandwidth when \( s = [i = 3, b_1 = 1, d_1 = 3, d_2 = 1, d_3 = 3, v = 2, l = 3] \). .................................................. 55

3.10 Structure of the unconstrained MDP optimal vertical handoff policy on the packet delay when \( s = [i = 2, b_1 = 1, b_2 = 4, b_3 = 2, d_3 = 3, v = 1, l = 3] \). .................................................. 56

3.11 Structure of the unconstrained MDP optimal vertical handoff policy on the mean of user’s velocity when \( s = [i = 3, b_1 = 3, d_1 = 3, b_2 = 4, d_2 = 1, d_3 = 3, l = 3] \). .................................................. 57
# List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>3GPP</td>
<td>3rd Generation Partnership Project</td>
</tr>
<tr>
<td>4G</td>
<td>4th Generation</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>CBR</td>
<td>Constant Bit Rate</td>
</tr>
<tr>
<td>CDMA</td>
<td>Code Division Multiple Access</td>
</tr>
<tr>
<td>CMDP</td>
<td>Constrained Markov Decision Process</td>
</tr>
<tr>
<td>CTMDP</td>
<td>Continuous-time Markov Decision Process</td>
</tr>
<tr>
<td>FC</td>
<td>Feature Collector</td>
</tr>
<tr>
<td>FNQD</td>
<td>Fuzzy Logic-based Normalized Quantitative Decision</td>
</tr>
<tr>
<td>FTP</td>
<td>File Transfer Protocol</td>
</tr>
<tr>
<td>GPRS</td>
<td>General Packet Radio Service</td>
</tr>
<tr>
<td>GSM</td>
<td>Global System for Mobile Communications</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>------------------------------------------</td>
</tr>
<tr>
<td>HDM</td>
<td>Handoff Decision Manager</td>
</tr>
<tr>
<td>HHO</td>
<td>Horizontal Handoff</td>
</tr>
<tr>
<td>MADM</td>
<td>Multiple Attribute Decision Making</td>
</tr>
<tr>
<td>MDP</td>
<td>Markov Decision Process</td>
</tr>
<tr>
<td>MIH</td>
<td>Media Independent Handover</td>
</tr>
<tr>
<td>MT</td>
<td>Mobile Terminal</td>
</tr>
<tr>
<td>NHM</td>
<td>Network Handling Manager</td>
</tr>
<tr>
<td>PDA</td>
<td>Personal Digital Assistant</td>
</tr>
<tr>
<td>PEV</td>
<td>Performance Evaluation Value</td>
</tr>
<tr>
<td>PIA</td>
<td>Policy Iteration Algorithm</td>
</tr>
<tr>
<td>PR</td>
<td>Predicted RSS</td>
</tr>
<tr>
<td>QoS</td>
<td>Quality of Service</td>
</tr>
<tr>
<td>RSS</td>
<td>Received Signal Strength</td>
</tr>
<tr>
<td>SAW</td>
<td>Simple Additive Weighting</td>
</tr>
<tr>
<td>SINR</td>
<td>Signal to Interference and Noise Ratio</td>
</tr>
<tr>
<td>SMDP</td>
<td>Semi-Markov Decision Process</td>
</tr>
<tr>
<td>Acronym</td>
<td>Description</td>
</tr>
<tr>
<td>---------</td>
<td>-------------</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal to Noise Ratio</td>
</tr>
<tr>
<td>TCP</td>
<td>Transmission Control Protocol</td>
</tr>
<tr>
<td>TOPSIS</td>
<td>Technique for Order Preference by Similarity to Ideal Solution</td>
</tr>
<tr>
<td>UDP</td>
<td>User Datagram Protocol</td>
</tr>
<tr>
<td>UMTS</td>
<td>Universal Mobile Telecommunications System</td>
</tr>
<tr>
<td>VHDF</td>
<td>Vertical Handoff Decision Function</td>
</tr>
<tr>
<td>VHM</td>
<td>Vertical Handoff Manager</td>
</tr>
<tr>
<td>VHO</td>
<td>Vertical Handoff</td>
</tr>
<tr>
<td>WiMAX</td>
<td>Worldwide Interoperability for Microwave Access</td>
</tr>
<tr>
<td>WLAN</td>
<td>Wireless Local Area Network</td>
</tr>
</tbody>
</table>
## List of Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\cdot</td>
</tr>
<tr>
<td>$\times$</td>
<td>Cartesian product</td>
</tr>
<tr>
<td>$\mathcal{A}_s$</td>
<td>Action set when the current state is $s$</td>
</tr>
<tr>
<td>$\mathcal{B}$</td>
<td>Set of available bandwidth</td>
</tr>
<tr>
<td>$b^m_{\max}$</td>
<td>Maximum bandwidth available in network $m$</td>
</tr>
<tr>
<td>$C_{\max}$</td>
<td>User’s monetary budget</td>
</tr>
<tr>
<td>$C^n(s)$</td>
<td>Expected discounted total access cost given policy $\pi$ and initial state $s$</td>
</tr>
<tr>
<td>$c(s,a)$</td>
<td>Cost function given current state $s$ and chosen action $a$</td>
</tr>
<tr>
<td>$\mathcal{D}$</td>
<td>Set of packet delay values</td>
</tr>
<tr>
<td>$d^m_{\max}$</td>
<td>Maximum packet delay when using network $m$</td>
</tr>
<tr>
<td>$E$</td>
<td>Expectation</td>
</tr>
</tbody>
</table>
List of Symbols

- $f(s, a)$: Total benefit function given current state $s$ and chosen action $a$
- $f_b(s, a)$: Bandwidth benefit function given current state $s$ and chosen action $a$
- $f_d(s, a)$: Delay benefit function given current state $s$ and chosen action $a$
- $g(s, a)$: Total penalty function given current state $s$ and chosen action $a$
- $g_{drop}(s, a)$: Call dropping penalty function given current state $s$ and chosen action $a$
- $g_{switch}(s, a)$: Switching cost penalty function given current state $s$ and chosen action $a$
- $K_{i,a}$: Normalized switching cost from network $i$ to $a$
- $\mathcal{L}$: Set of location types
- $l_{max}$: Maximum number of location types
- $\mathcal{M}$: Set of available network ID
- $N$: Connection termination time
- $P(\ )$: Transition probability function
**List of Symbols**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_l$</td>
<td>Total area of location type $l$</td>
</tr>
<tr>
<td>$\dot{Q}_l$</td>
<td>Effective area of location type $l$</td>
</tr>
<tr>
<td>$Q_\beta(s,a)$</td>
<td>State-action reward function given Lagrange multiplier $\beta$, current state $s$, and chosen action $a$</td>
</tr>
<tr>
<td>$r(s,a)$</td>
<td>Reward function given current state $s$ and chosen action $a$</td>
</tr>
<tr>
<td>$S$</td>
<td>State space</td>
</tr>
<tr>
<td>$T$</td>
<td>Set of decision epochs</td>
</tr>
<tr>
<td>$\mathcal{V}$</td>
<td>Set of possible velocity values</td>
</tr>
<tr>
<td>$V_{max}$</td>
<td>Maximum velocity threshold</td>
</tr>
<tr>
<td>$V_{min}$</td>
<td>Minimum velocity threshold</td>
</tr>
<tr>
<td>$v_{max}$</td>
<td>Maximum possible velocity</td>
</tr>
<tr>
<td>$v_\beta(s)$</td>
<td>Value function given Lagrange multiplier $\beta$ and initial state $s$</td>
</tr>
<tr>
<td>$v^\pi(s)$</td>
<td>Expected discounted total reward given policy $\pi$ and initial state $s$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Memory level</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Lagrange multiplier</td>
</tr>
<tr>
<td>Symbol</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------</td>
</tr>
<tr>
<td>( \delta )</td>
<td>Stationary policy</td>
</tr>
<tr>
<td>( \delta_t )</td>
<td>Decision rule at time ( t )</td>
</tr>
<tr>
<td>( \zeta )</td>
<td>Gaussian process</td>
</tr>
<tr>
<td>( \kappa )</td>
<td>User’s risky index</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>Discount factor</td>
</tr>
<tr>
<td>( \mu )</td>
<td>Mean value of velocity</td>
</tr>
<tr>
<td>( \Pi )</td>
<td>Set of all possible policies</td>
</tr>
<tr>
<td>( \pi )</td>
<td>Policy</td>
</tr>
<tr>
<td>( \pi^* )</td>
<td>Optimal policy</td>
</tr>
<tr>
<td>( \rho_l )</td>
<td>User density of location type ( l )</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>Standard deviation of velocity</td>
</tr>
<tr>
<td>( \phi )</td>
<td>Weight given to the switching cost factor</td>
</tr>
<tr>
<td>( \omega )</td>
<td>Weight given to the available bandwidth factor</td>
</tr>
</tbody>
</table>
Acknowledgements

First of all, I would like to express my deepest gratitude to my supervisor, Dr. Vincent Wong, for his patient guidance, constant encouragement, and excellent advice throughout my graduate study. Without his invaluable help, this work would not be possible.

A special thanks goes to my colleague, Enrique Stevens-Navarro, who provided me with precious assistance during the completion of this thesis. I am also thankful to all my colleagues in Dr. Wong’s group: Yuxia Lin, Amir Hamed Mohsenian-Rad, Vahid Shah-Mansouri, Mani Malekesmaeili, Keivan Ronasi, Man Hon Cheung, and Joo-Han Song, as well as other friends in the data communications group, for sharing their experiences and knowledge during the time of my study.

Finally, I take this opportunity to express my profound gratitude to my beloved parents and grandparents for their understanding, patience, support, and endless love during my study in Canada. To them I dedicate this thesis.
Chapter 1

Introduction

Various wireless technologies have evolved rapidly during the past decade. Mobile devices and gadgets (e.g., cellular phones, personal digital assistants (PDAs), laptops) supported by these technologies are becoming more popular and important in people’s everyday life. Wireless local area networks (WLANs), Worldwide Interoperability for Microwave Access (WiMAX), and cellular networks are three paradigms of such technologies in the present wireless realm.

WLANs, which are based on the IEEE 802.11 standards, are able to provide services with data rate up to 11 Mbps (802.11b) or 54 Mbps (802.11a/g) at a relatively low access and deployment cost. Moreover, 802.11n promises to offer data rate higher than 100 Mbps \[1\]. However, the coverage area of WLANs is typically less than 100 meters, making them only suitable for hotspot regions such as hotels, libraries, airports, and coffee shops.

On the other hand, WiMAX technology can provide broadband wireless services to users in a variety of ways. This technology is based on the IEEE 802.16 standards, in which the 802.16e-2005 (also called mobile WiMAX) introduces the support of mobility. The WiMAX mobile network deployment is expected to provide up to 15 Mbps of capacity within a typical cell radius of up to 3 km \[2\].
Chapter 1. Introduction

The typical supported service data rates of cellular networks such as GSM (Global System for Mobile Communications), GPRS (General Packet Radio Service), UMTS (Universal Mobile Telecommunications System), or CDMA2000 (Code Division Multiple Access 2000) are within the range of a few *kbps* to several *Mbps*. The cost of accessing and deploying cellular networks is much higher than that of the WLANs.

Driven by the complementary characteristics of these wireless technologies (e.g., high-rate, low-cost, small coverage area of WLANs versus low-rate, high-cost, large coverage area of cellular networks), there is a trend to interconnect these access networks into an integrated system. The inspiration behind it, which is also the goal of the 4th Generation (4G) wireless communication systems, is to utilize these 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) [3], 3GPP2 [4], and the IEEE 802.21 Media Independent Handover (MIH) working group [5] 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 UMTS to benefit from good quality of service (QoS), WiMAX to achieve a high data rate, or 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.

Consider the following scenario. Imagine that you are at a coffee shop, waiting for your friends Amy and Bob to give you a ride to school. You start to check email using your PDA while waiting for them, using the coffee shop’s WLAN. You find out that the bidding of an eBay item that you are interested in is going to end soon, so you go to the eBay website and start to bid on it. In the meantime, your friends arrive and you hop on their car. When the car moves away from the coffee shop, your PDA switches to the UMTS cellular network automatically, keeping you connected to the eBay website. After you have bought the item, Amy picks up your PDA and starts to watch a drama on YouTube, which she left unfinished before she went out. When the car arrives at the campus, the PDA detects the campus-wide WLAN and then switches to it automatically. Due to the higher bandwidth available in WLANs, the video that Amy is watching can now be played in a higher resolution.

In the above scenario, in order to maintain the continuity of the sessions (e.g., eBay webpage, video from YouTube), the switching between different networks (e.g., coffee shop’s WLAN to UMTS, UMTS to campus-wide WLAN) is important. This kind of switching is known as vertical handoff, as opposed to horizontal handoff, which is the switching between cells of the same network (e.g., different cells in the UMTS network). In other words, vertical handoff is responsible for service continuity when a connection needs to migrate across heterogeneous wireless access networks, and is also referred as intersystem handoff, while horizontal handoff is responsible for service continuity when
a connection needs to migrate across homogeneous wireless access networks, and is also referred as *intrasytem* handoff.

Vertical handoff generally involves three phases [6, 7]: *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. Those information may include the access costs and current QoS parameters for different services. Since the QoS provided by each network may change with time, the system discovery phase should be invoked periodically.

In the 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 it 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. Various vertical handoff execution schemes have been proposed in the literature (e.g., [8, 9], and [10]) with the aim of reducing the time and complexity spent during the vertical handoff procedure. However, the investigation on the vertical handoff execution is beyond the scope of this thesis.
Chapter 1. Introduction

1.1 Related Work

Vertical handoff decision algorithms provide guidance on how to choose a target network when multiple choices are present. In other words, these algorithms specify the aspects that need to be considered when making the vertical handoff decisions. We now summarize some of the recent work on vertical handoff decision algorithms in heterogeneous wireless networks.

In [11], a vertical handoff decision algorithm based on fuzzy logic-based normalized quantitative decision (FNQD) algorithm is proposed. Network parameters such as the current received signal strength (RSS), predicted RSS (PR), user’s velocity, and the available bandwidth are considered when making the decision. The MT samples the RSS periodically. With the sampled RSSs, the PR can be obtained by the forward differential prediction algorithm. This algorithm also has a pre-decision (PD) module which is applied before the handoff decision module, so that unnecessary data can be filtered out to provide an accurate handoff trigger. Specifically, the MT’s velocity and the PR are considered as PD metrics to filter data for the FNQD. There are three sub-procedures in the FNQD module: fuzzification, normalization, and quantitative decision. First, the input parameters (i.e., current RSS, PR, and user’s velocity) are calculated, and fed into the fuzzification procedure. After this, the FNQD normalizes these parameters to obtain their normalized evaluation values. Then, the FNQD determines a performance evaluation value (PEV) for each candidate network. Finally, the handoff decision algorithm is as follows: the greater the PEV of a certain network is, the higher the probability that
this network becomes the target network to switch to.

In [12], a middleware solution called Vertical Handoff Manager (VHM) is implemented to address the vertical handoff decision problem. The architecture of the VHM consists of three components: Network Handling Manager (NHM), Feature Collector (FC), and Artificial Neural Networks (ANN) selector.

The NHM is mainly responsible for handling interactivity between VHM and the available wireless networks, as well as between VHM and wireless devices. NHM handles requests made by wireless devices when initiating a handoff, and contacts the available wireless networks to begin the feature collection process. The FC module retrieves features that are used as attributes to feed into the neural network. These features can be divided into three categories: network-specific, user-specific, and device-specific. The network-specific features include the network’s access cost, security, transmission range, and data rate. The user-specific features are the user’s preferences, such as cost, security level, and power consumption. The device-specific features provide information about the mobile device, and are used for compatibility issues and other technical information necessary when making a handoff decision. After all the features are collected by the FC module, the ANN is utilized to find the best match between what the user wants and what is available across these wireless networks. Once a match is found, the VHM will initiate a message exchange between the mobile device and the chosen wireless network to execute the vertical handoff procedure.

In [13], a handoff decision manager (HDM) is proposed for network selection. The
proposed HDM algorithm considers different factors such as the candidate network’s cost, coverage, data rate, and traffic load. First, the HDM considers the availability of lower overlay networks in order to decide whether to perform downward vertical handoff (e.g., from UMTS to WLAN). If there are available lower overlay networks, HDM calculates the utilities of these candidates and then compares these utilities with that of the current serving network. If one of the candidates has a consistently better utility, then the HDM decides to make a downward vertical handoff to that network. Secondly, if handoff is not invoked, then the algorithm will consider the RSSs and use the conventional 2-level threshold scheme to decide whether to perform horizontal handoff. Finally, if no handoff has been triggered yet, then the HDM will consider the availability of upper overlay networks for upward vertical handoff (e.g. from WLAN to UMTS). Since the aim of an upward vertical handoff is to maintain the connectivity rather than improving QoS, this decision should be made quickly. As a result, the HDM uses RSS as an indication. If the RSS of the current network decays consistently, then the HDM will decide to switch to an upper overlay network.

In [14], the vertical handoff decision is made 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 access network capacity is calculated based on parameters such as upstream and downstream bandwidth, power consumption, and charging rate. Moreover, the algorithm gives priority to horizontal handoff by including the signalling cost of vertical handoff into the cost function. Fi-
nally, the local balancing factor is added into the cost function in order to distribute the traffic over different access networks in a manner such that the total resources of the heterogeneous wireless access networks can be used more efficiently. Given the cost function of each access network, the handoff procedure is as follows. First, it creates a set of candidates by choosing the access networks whose RSSs are higher than the predefined threshold. Next, it classifies the candidate networks by their cost functions. Finally, the chosen network is the one with the least cost function value.

In [15], the vertical handoff decision algorithm is built upon three conditions: the QoS parameters, the call dropping probability, and the number of ping-pong events. First, the minimum SNR threshold is determined, for which QoS parameters such as the maximum bit error rate, the maximum data block delay, and the minimum data rate should all be satisfied. Then, the minimum downlink and uplink SNR threshold values are determined to jointly guarantee the QoS requirements for all active downlink and uplink multi-service connections of the MT. Secondly, the handoff entry threshold is derived by the average number of ping-pong events during the entry to the WLAN system. The handoff exit threshold is derived by the call dropping probability. Furthermore, the WLAN is selected as the preferred network for the MT. The goal of this 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 maximum call dropping probability and the maximum average number of ping-pong events constraints.

In [16], 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. This scheme is widely used due to its ease of understanding and simplicity in implementation. For example, in [17], a distributed vertical handoff decision scheme is proposed using the SAW algorithm. The scheme includes the dropping probability, bandwidth, and cost as the handoff parameters in order to make a decision.

In TOPSIS, the idea is based on the principle that the chosen candidate should have the shortest distance from the ideal solution and the farthest distance from the negative-ideal solution. As a result, 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 [18], a vertical handoff decision algorithm based on fuzzy logic is presented. This scheme uses the following parameters as the decision metric: RSS, SINR, cost, bit error rate, latency, and data transmission rate. In [19], a vertical handoff decision scheme based on ELECTRE is proposed. ELECTRE is an MADM algorithm, which performs pair-wise comparisons among the alternatives. However, ELECTRE assumes a monotonic utility and does not provide a complete ranking of all the alternatives. Since a complete ranking is necessary to find the top ranking candidate network, Bari et al. modified ELECTRE so that it is able to provide complete ranking of alternatives even in scenarios where the utility of some attributes are non-monotonic, and used the adapted ELECTRE to make
vertical handoff decision. This modification allows the application of the algorithm in scenarios where the decision maker can choose the alternative that has the attributes closest to a reference set of attribute values. The attributes considered in [19] include the bandwidth, delay, packet jitter, packet loss, utilization, and network cost.

In [20], a vertical handoff decision algorithm which uses the received SINR from various access networks as the handoff criterion is proposed. Since the maximum achievable data rate is a function of the SINR, this SINR-based vertical handoff method is applicable. The received SINR from the cellular network is converted to the equivalent SINR required to achieve the same data rate in WLAN, and then compared with the actual received SINR from WLAN. With both SINR being considered, vertical handoff is triggered when the user can obtain higher equivalent SINR from another access network. It means that given the receiver end SINR measurements of both WLAN and the cellular network, the handoff mechanism has the knowledge of the estimated maximum possible receiving data rate a user can obtain from either WLAN or the cellular network. This gives the vertical handoff mechanism the ability to make handoff decisions with multimedia QoS considerations, 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 [21], a handoff management system which includes several modules and procedures is proposed to make handoff decisions in heterogeneous wireless networks. Some of these modules collect the link-layer and network-layer information which is useful for handoff
management, and other modules use this information to decide the appropriate time to initiate the handoff and execute the handoff procedure. When the handoff management system starts, it activates the network environment detection module and the traffic measurement module. The network environment detection module gathers the information about the available access networks and their performances. For example, it determines how many WLANs are currently available and their signal strengths. Meanwhile, the traffic measurement module determines the type of the current application and its QoS requirement. For instance, it considers whether the current application is a real-time application, and determines its throughput and delay requirements. If there are multiple networks available and the current access network cannot satisfy the QoS requirement of the running application, the handoff decision module will be invoked. It determines the destination network based on the sojourn time of the MT in the candidate networks and these networks’ QoS estimation, 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 [22], a vertical handoff decision algorithm based on RSS-based handoff trigger scheme is implemented. By applying the autoregressive integrated moving average model, the future RSS values can be predicted. The handoff decision can then be made according to these RSS predictions. In [23], a vertical handoff decision function (VHDF) is proposed for handoff decision. The VHDF is used to measure the improvement gained by switching to a particular network. It considers the following factors: cost of services, security, power
consumption, network condition, and network performance. As the MT roams across different networks, VHDF is evaluated for all accessible networks. The network with the highest calculated VHDF value is the most desirable network for the users to switch to based on their specified preferences.

In [24], a utility-based network selection strategy is presented. It captures the trade-offs between the users’ preferences and their vertical handoff decisions. Several utility functions are examined which explore different users’ attitudes on their current applications. For example, users who are willing to pay more money are more likely to obtain better services.

In [25], a vertical handoff decision algorithm based on dynamic programming is proposed. Since the enhancement of 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. Two models are proposed to capture the user’s mobility behavior. The open area model assumes that the MT’s velocity is independent on its location, while the roads model assumes that the users are moving along roads, in which the velocity can be considered as a scalar rather than a vector.

The user’s velocity and moving pattern are also considered in the vertical handoff decision algorithms in [26] and [27], where the work in [26] is an adaptive multi-criteria vertical handoff decision algorithm based on fuzzy-logic, and the work in [27] is a movement-aware vertical handoff decision algorithm between WLAN and the WiMAX network. Moreover,
in [28], a framework is proposed to evaluate different vertical handoff decision algorithms, in which the MT’s mobility is modeled by a Markov chain.

In [29], 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 situation, device capability, and the effects of the environment. In [30], a vertical handoff decision algorithm is proposed. The problem is formulated as an MDP, and the model considers the available bandwidth and delay of the candidate networks.

The model in this thesis is an extension of the one proposed in [30], 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.

1.2 Motivations and Objectives

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., consider only current QoS of the networks and current MT’s conditions). Handoff decisions should also consider the probabilistic outcomes of the future system states as a result of the current decision. Some work (e.g., [25], [29], and [30]) 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 environment. This includes the available bandwidth, delay, switching cost, and access cost information of the overlaying networks.

2. The state of the user. 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 wireless networks based on the constrained MDP (CMDP) model. The first three aspects are embedded into the system model, the fourth aspect is realized by the MDP formulation, and the last aspect is viewed as the constraint of the MDP problem. In our proposed CMDP model, a benefit function is used to capture the advantages, which are assessed in terms of bandwidth and delay that a user can gain by choosing an action. The actions available to the user are related to its location. In the meantime, there is a penalty function that represents the potential drawbacks associated with this action. Specifically, the penalty includes the signaling cost incurred and the call dropping probability of taking this action. A reward function is then used to represent the reward that a user can obtain by choosing an action, considering the associated benefit and penalty. An access cost function is used to capture the access cost that a user needs to spend for using a specific network, and the constraint of the problem is that the user has a
budget on the total access cost for the duration of each connection. 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. In other words, the goal of this problem is to determine the method of choosing networks such that the user can enjoy the best possible QoS during the connection while he/she will not be charged more than his/her budget for this connection.

1.3 Contributions

The main contributions of this thesis are as follows [31]:

- 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. 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 \[16\], ELECTRE \[19\]) 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.

### 1.4 Structure of the Thesis

The rest of the thesis is organized as follows. The system model and the CMDP formulation are described in Chapter 2 in which the structure of the optimal vertical handoff policy is also investigated. The numerical results and discussions are presented in Chapter 3. Conclusions and future work are given in Chapter 4.
Chapter 2

CMDP Formulation of the Vertical Handoff Decision Problem

In this chapter, we explain how to formulate the vertical handoff decision problem as a CMDP model in Section 2.1. In Section 2.2 we describe how to obtain the optimal solution. In Section 2.3 we investigate the structure of the optimal policy, and show that the optimal policy has structural results on some state variables.

2.1 CMDP Formulation and System Model

First, let’s take a look at what an MDP model is. People make decisions every day. Simple examples include whether to upgrade your car with snow tires in the winter, or whether to eat breakfast when you know that you are going for a lunch buffet. These decisions have not only immediate impacts, but also long-term consequences. One cannot make decisions only based on the current conditions, because your current decision will also affect the decisions in the future. If the decision maker does not consider the relationship between current and future decisions and the consequences, he/she may not
achieve the best overall performance. As an example, in a marathon race, a player decides to sprint at the beginning will quickly lose energy hence results in a poor finish.

The MDP is a model for sequential decision making under uncertainty. It takes into account both the outcomes of the current decisions and the future decision making opportunities. The key elements of the MDP model are as follows [32]:

1. A set of decision epochs.
2. A set of system states.
3. A set of available actions.
4. A set of state and action dependent immediate rewards.
5. A set of state and action dependent probabilities.

At each decision epoch (i.e., time), the system state provides the decision maker with all necessary information for choosing an action from the set of available actions in that particular state. After the decision maker makes a choice on the action in that state, two things will happen. First, the decision maker receives an immediate reward. Second, the system evolves to a new (and possibly different) state at the next decision epoch, according to some state transition probabilities. Both the immediate reward and the state transition probabilities depend on the state and the chosen action. As the process evolves through time, the decision maker receives a sequence of rewards.

A decision rule prescribes a procedure to choose an action in each state at a specified decision epoch. A policy specifies the decision rules for all decision epochs, and thus can
be viewed as a sequence of decision rules. The goal of the sequential decision problem is
to implement a policy such that the sum of the total rewards is maximized.

A CMDP model is an MDP model with constraints \[33, 34, 35\]. In this thesis, we
only consider the case where there is only one constraint. In the CMDP model, after
choosing an action at a decision epoch, the decision maker receives both an immediate
reward and an immediate cost. The goal is to maximize the sum of the expected total
rewards while the sum of the expected cost is less than a predefined value.

We now 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 \[33\]. At each
decision epoch, the MT has to choose an action (i.e., selects 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.
2.1.1 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 $S$, and we only consider 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 follows:

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

where $\times$ denotes the Cartesian product, $\mathcal{M}$ represents the set of available network ID that the MT can connect to, $\mathcal{B}^{m}$ and $\mathcal{D}^{m}$ denote the set of available bandwidth and delay

**Figure 2.1**: Coverage area of different networks.
of network $m \in \mathcal{M}$, respectively. $\mathcal{V}$ denotes the set of possible velocity values of the MT, and $\mathcal{L}$ denotes the set of location types that the MT can possibly reside in.

Since a finite countable state space is being considered in this thesis, the bandwidth and delay can be quantized into multiples of unit bandwidth and unit delay, respectively. Specifically, the set of available bandwidth of network $m$ is:

$$B^m = \{1, 2, \ldots, b_{\text{max}}^m\}, \quad m \in \mathcal{M},$$

where $b_{\text{max}}^m$ denotes the maximum bandwidth available to a connection from network $m$. For example, the unit bandwidth of WLAN and the UMTS network can be 500 $kbps$ and 16 $kbps$, respectively.

Similarly, the set of packet delay of network $m$ is:

$$D^m = \{1, 2, \ldots, d_{\text{max}}^m\}, \quad m \in \mathcal{M},$$

where $d_{\text{max}}^m$ denotes the maximum delay provided to a connection by network $m$. For example, the unit delay of WLAN and the UMTS 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:

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

where $v_{\text{max}}$ 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:

\[ \mathcal{L} = \{1, 2, \ldots, l_{\text{max}}\}, \]

where \( l_{\text{max}} \) 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. If the MT is within an area of location type \( l \), it implies that there are \( l \) kinds of wireless access networks that the MT can choose from. For instance, if an MT is in the location area covered by both WiMAX and UMTS networks, this place falls into the category of location type 2. However, if an MT is in the location area covered by only an UMTS network, this place is a location type 1 area.

Let vector \( s = [i, b_1, d_1, \ldots, b_{|\mathcal{M}|}, d_{|\mathcal{M}|}, v, l] \) denote the current state of the system, 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 \( s \), the MT chooses an action \( a \in \mathcal{A}_s \), where the action set \( \mathcal{A}_s \subset \mathcal{M} \) consists of the ID of the networks that the MT can potentially switch to. Given the current state is \( s \), if the chosen action is \( a \), the probability that the next state becomes \( s' = [j, b'_1, d'_1, \ldots, b'_{|\mathcal{M}|}, d'_{|\mathcal{M}|}, v', l'] \) is:

\[
P[s' \mid s, a] = \begin{cases} 
P[v' \mid v] P[l' \mid l] \prod_{m \in \mathcal{M}} P[b'_m, d'_m \mid b_m, d_m], & j = a, \\ 0, & j \neq a, \end{cases} \tag{2.1}
\]

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 explain how we obtain these transition probabilities as follows.

The transition probability of the MT’s velocity is obtained based on the Gauss-Markov mobility model from [36]. Since a mobile user usually travels with a destination in mind, and the change of a mobile user’s velocity within a short period of time is limited due to physical restriction, the MT’s future location and velocity are likely to be correlated with its past and current location and velocity. As a result, the memoryless nature of the random-walk model is not suitable to represent the MT’s behavior. Furthermore, the fluid-flow model that represents vehicular traffic is also not suitable to capture the pedestrian behavior of mobile users.

In the Gauss-Markov 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:

\[
v' = \alpha v + (1 - \alpha) \mu + \sigma \sqrt{1 - \alpha^2} \zeta, \tag{2.2}
\]

where \( v \) is the MT’s velocity at the current decision epoch, \( v' \) is the MT’s velocity at the next decision epoch, \( \alpha \) is the memory level (i.e., \( 0 \leq \alpha \leq 1 \)), \( \mu \) and \( \sigma \) are the mean and standard deviation of \( v \), respectively, and \( \zeta \) 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 (2.2), the MT’s velocity transition probability matrix (i.e., \( P[v' | v] \)) 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 network that has a larger coverage area (e.g., WiMAX). Although this assumption may not hold for the cases when $|\mathcal{M}|$ is large, it is still 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 $l \in \mathcal{L}$ as follows. Location type 1 is the area covered only by the UMTS network, location type 2 is the area covered by UMTS and WiMAX, but not WLAN, and location type 3 is the area covered by all three networks (i.e., UMTS, WiMAX, and WLAN).

Let $\Theta_l$ denote the total area of location type $l$ and $\rho_l$ denote the user density of location type $l$. The effective area of location type $l$ is defined as:

$$\hat{\Theta}_l = \Theta_l \rho_l, \quad l \in \mathcal{L}. \tag{2.3}$$

In practice, the user density in areas covered by different networks (e.g., WLAN and the UMTS network) are usually not the same [37, 38]. For example, the area covered by both WLAN and the UMTS network usually has more active connections than the area only covered by the UMTS 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\footnote{The time between two successive decision epochs is on the order of seconds.}. 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 is defined as follows:

$$P[l' \mid l] = \begin{cases} \frac{\hat{\Theta}_{l'}}{\sum_{\xi=1,l+1} \Theta_{\xi}}, & \text{for } l = 1, l' = 1, 2, \\ \frac{\hat{\Theta}_{l'}}{\sum_{\xi=1,l,l+1} \Theta_{\xi}}, & \text{for } l = 2, \ldots, l_{\text{max}} - 1, l' = l - 1, l, l + 1, \\ \frac{\hat{\Theta}_{l'}}{\sum_{\xi=1,l} \Theta_{\xi}}, & \text{for } l = l_{\text{max}}, l' = l_{\text{max}} - 1, l_{\text{max}}. \end{cases} \quad (2.4)$$

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:

$$P[b'_1, d'_1 \mid b_1, d_1] = \begin{cases} 1, & \text{if } b'_1 = b_1, d'_1 = d_1, \\ 0, & \text{otherwise}. \end{cases}$$

For WiMAX and WLAN, we estimate such probabilities in a simulation-based manner. In ns-2 simulator \cite{39}, typical IEEE 802.16 WiMAX \cite{40} and IEEE 802.11b WLAN are simulated in which the users arrive and depart from the networks with an average Poisson rate of 0.2 users per second. 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).
2.1.2 Rewards

When an MT chooses an action $a$ in state $s = [i, b_1, d_1, \ldots, b_{|M|}, d_{|M|}, v, l]$, it receives an immediate reward $r(s, a)$. 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 $s$ (recall $i$ denotes the ID of the current network):

$$
f_b(s, a) = \begin{cases} 
\frac{b_a - b_i}{\max_{k \in M} (b_k - b_i)}, & \text{if } b_a > b_i, \\
0, & \text{if } b_a = b_i, \\
-\frac{b_a - b_i}{\min_{k \in M} (b_k - b_i)}, & \text{if } b_a < b_i.
\end{cases}
$$

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 $s$, 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 $s$:

$$f_d(s, a) = \begin{cases} 
\frac{d_i - d_a}{\max_{k \in M} \{d_i - d_k\}}, & \text{if } d_a < d_i, \\
0, & \text{if } d_a = d_i, \\
-\frac{d_i - d_a}{\min_{k \in M} \{d_i - d_k\}}, & \text{if } d_a > d_i.
\end{cases}$$  \hspace{1cm} (2.6)

Given that the MT is currently connecting to network $i$, if the action $a$ leads to a network with a lower delay, then the benefit function value is represented by a fraction, in which the numerator is the MT’s actual decrease of delay by choosing action $a$ in state $s$, and the denominator is the MT’s maximum possible decrease of delay. 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 higher delay, then the benefit function value it can obtain 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.

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

$$f(s, a) = \omega f_b(s, a) + (1 - \omega)f_d(s, a), \quad s \in S, \ a \in A_s, \hspace{1cm} (2.7)$$

where $f_b(s, a)$ and $f_d(s, a)$ are both normalized (i.e., between 0 and 1), and $\omega$ is the weight given to the bandwidth aspect with $0 \leq \omega \leq 1$. 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).
Chapter 2. CMDP Formulation of the Vertical Handoff Decision Problem

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

\[
g_{\text{switch}}(s, a) = \begin{cases} 
K_{i,a}, & \text{if } i \neq a, \\
0, & \text{if } i = a, 
\end{cases}
\]  

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:

\[
g_{\text{drop}}(s, a) = \begin{cases} 
0, & \text{if } i = a, \\
0, & \text{if } i \neq a, 0 \leq v \leq V_{\text{min}}, \\
\frac{v-V_{\text{min}}}{V_{\text{max}}-V_{\text{min}}}, & \text{if } i \neq a, V_{\text{min}} < v < V_{\text{max}}, \\
1, & \text{if } i \neq a, v \geq V_{\text{max}}, 
\end{cases}
\]  

where \(V_{\text{max}}\) and \(V_{\text{min}}\) 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 UMTS network (hence lose the WLAN signal) before the WLAN-to-UMTS vertical handoff procedure is completed. As a result, the MT’s connection will be dropped.

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

\[
g(s, a) = \phi g_{\text{switch}}(s, a) + (1 - \phi) \kappa g_{\text{drop}}(s, a), \quad s \in S, \; a \in A_s, \quad 0 \leq \phi \leq 1
\]
where $\phi$ is the weight given to the switching cost factor with $0 \leq \phi \leq 1$, and $\kappa \in [0, 1]$ is the MT’s risky index. This risky index represents the user’s preference, since 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, whereas some others may refrain from switching whenever a risk is present.

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

$$
r(s, a) = f(s, a) - g(s, a), \quad s \in S, \ a \in A_s,
$$

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 risky preference.

### 2.1.3 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):

$$
c(s, a) = \frac{b_a C_a}{\max_{m \in M} \{b_m C_m\}}, \quad s \in S, \ a \in A_s,
$$

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.
2.2 CMDP Optimality Equations

In this section, we 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 $\delta_t : S \rightarrow A$, where $\delta_t$ 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 $v^\pi(s)$ denote the expected total reward between the first decision epoch and the connection termination, given that policy $\pi$ is used with initial state $s$. We can represent $v^\pi(s)$ as:

$$v^\pi(s) = E^\pi_s \left[ E_N \left\{ \sum_{t=1}^{N} r(s_t, a_t) \right\} \right], \quad s \in S,$$

(2.13)

where $E^\pi_s$ denotes the expectation with respect to the variables given policy $\pi$ and initial state $s$, 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 (2.13) can be re-written as follows:

$$v^\pi(s) = E^\pi_s \left\{ \sum_{t=1}^{\infty} \lambda^{t-1} r(s_t, a_t) \right\}, \quad s \in S,$$

(2.14)

where $\lambda$ can also be interpreted as the discount factor of the model (i.e., $0 \leq \lambda < 1$).

We need to find the policy that leads to this optimal solution while satisfying the expected total access cost constraint. We define a policy $\pi^* = (\delta^*_1, \delta^*_2, \ldots)$ to be optimal in $\Pi$ if $v^{\pi^*}(s) \geq v^\pi(s)$ for all $\pi \in \Pi$, where $\Pi$ is the set of all possible policies. We assume stationary policy in this model, and 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 $\pi$ simply by $\delta$. 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 [32, pp. 22].

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

$$\max v^\pi(s) = E^\pi_s \left\{ \sum_{t=1}^{\infty} \lambda^{t-1} r(s_t, a_t) \right\},$$

subject to

$$C^\pi(s) = E^\pi_s \left\{ \sum_{t=1}^{\infty} \lambda^{t-1} c(s_t, a_t) \right\} \leq C_{\text{max}},$$

where $C^\pi(s)$ denotes the expected discounted total access cost with respect to policy $\pi$ and initial state $s \in S$.

To solve (2.15) without the constraint on the expected discounted total access cost, we can use the Policy Iteration Algorithm (PIA) [32] to solve the following optimality equations:

$$v(s) = \max_{a \in A} \left\{ r(s, a) + \sum_{s'} \lambda P[s' | s, a] v(s') \right\}, \quad s \in S,$$

and the corresponding optimal policy is given as:

$$\delta^*(s) = \arg \max_{a \in A_s} \left\{ r(s, a) + \sum_{s'} \lambda P[s' | s, a] v(s') \right\}, \quad s \in S.$$  (2.17)

However, since (2.15) has a constraint in it, we cannot use the PIA directly. We need to first use the Lagrangian approach [33, 41] to reduce the CMDP problem into an equivalent unconstrained MDP problem. By including the Lagrange multiplier $\beta$ with $\beta > 0$, we have:

$$r(s, a; \beta) = r(s, a) - \beta c(s, a), \quad s \in S, \ a \in A_s,$$  (2.18)
Algorithm 1 - The Policy Iteration Algorithm (PIA):

1: Set \( k = 0 \), and select an arbitrary decision rule \( \delta^0_\beta(s) \in A_s \) for all \( s \in S \).

2: (Policy evaluation) Obtain \( v^k_\beta(s) \) for all \( s \in S \) by solving:

\[
(I - \lambda P) v^k_\beta = r_\beta,
\]

where \( I \) is an identity matrix of size \(|S|\).

3: (Policy improvement) Choose \( \delta^{k+1}_\beta(s) \) to satisfy:

\[
\delta^{k+1}_\beta(s) = \arg \max_{a \in A_s} \left\{ r(s, a; \beta) + \sum_{s' \in S} \lambda P[s' \mid s, a] v^k_\beta(s') \right\},
\]

for all \( s \in S \).

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

where \( r(s, a; \beta) \) is the Lagrangian reward function.

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

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

which can be solved by using the PIA with a fixed value of \( \beta \). The procedures of the PIA algorithm is described in Algorithm 1. We denote the vectors \( v^k_\beta = (v^k_\beta(s), \ s \in S) \), \( r_\beta = (r(s, \delta^k_\beta(s); \beta), s \in S) \), and the matrix \( P = (P[s' \mid s, \delta^k_\beta(s)], s \in S, s' \in S) \).

The solutions of (2.19) correspond to the maximum expected discounted total reward \( v_\beta(s) \) and the pure policy \( \delta^*_\beta(s) \) which satisfies:

\[
\delta^*_\beta(s) = \arg \max_{a \in A_s} \left\{ r(s, a; \beta) + \sum_{s' \in S} \lambda P[s' \mid s, a] v_\beta(s') \right\}, \quad s \in S.
\]
Algorithm 2 - The Q-learning Algorithm (Part I):

1: Set $\beta_1$ to an arbitrary number greater than zero, and set $n = 1$.

2: Solve for $\delta^*_\beta_n(s)$ in (2.20) via PIA for all $s \in S$.

3: Determine $C^{\delta^*_\beta_n}(s)$ in the following equation:

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

(2.21)

for all $s \in S$.

4: Update the Lagrange multiplier by:

$$
\beta_{n+1} = \beta_n + \frac{1}{n} \left( \bar{C}^{\delta^*_\beta_n} - C_{\text{max}} \right),
$$

(2.22)

where

$$
\bar{C}^{\delta^*_\beta_n} = \frac{1}{|S|} \sum_{s \in S} C^{\delta^*_\beta_n}(s).
$$

(2.23)

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

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

When the CMDP problem is converted to an unconstrained MDP problem by a Lagrange multiplier $\beta$ in (2.18), there is a relationship between the constraint (i.e., the user’s budget $C_{\text{max}}$) and the Lagrange multiplier (i.e., $\beta$). In this thesis, we use the Q-learning algorithm (Algorithm 2) proposed in [41] to determine the proper $\beta$ (i.e., $\beta^*$) for a feasible $C_{\text{max}}$. 
Algorithm 3 - The Q-learning Algorithm (Part II):

1: Perturb $\beta^*$ by some $\Delta \beta$ to obtain:

$$\beta^- = \beta^* - \Delta \beta,$$

$$\beta^+ = \beta^* + \Delta \beta.$$  

2: Calculate the pure policies $\delta^-$ and $\delta^+$ (using $\beta^-$ and $\beta^+$) in (2.20) using the PIA.

3: Calculate the average expected discounted total access costs $\bar{C}^-$ and $\bar{C}^+$ (using $\delta^-$ and $\delta^+$) in (2.23).

4: Solve for $q$ in the following equation:

$$q \bar{C}^- + (1-q) \bar{C}^+ = C_{max}. \quad (2.24)$$

5: The optimal policy $\delta^*$ can be obtained as:

$$\delta^* = q \delta^- + (1-q) \delta^+. \quad (2.25)$$

Once $\beta^*$ has been obtained, we follow the procedure (Algorithm 3 in [34]) to find the optimal policy for the CMDP problem. As discussed in [34], the optimal policy for a CMDP with single constraint is a mixed policy of two pure policies. First, we perturb $\beta^*$ by some $\Delta \beta$ to obtain $\beta^- = \beta^* - \Delta \beta$ and $\beta^+ = \beta^* + \Delta \beta$. Then, we calculate the pure policies $\delta^-$ and $\delta^+$ in (2.20) (using $\beta^-$ and $\beta^+$, respectively) via PIA and their corresponding average expected discounted total access costs $\bar{C}^-$ and $\bar{C}^+$ in (2.23) (using $\delta^-$ and $\delta^+$, respectively). Next, we define a parameter $q$ such that $q \bar{C}^- + (1-q) \bar{C}^+ = C_{max}$. Thus the optimal policy $\delta^*$ of the CMDP problem is a randomized mixture of two
policies (i.e., $\delta^-$ and $\delta^+$), such that at each decision epoch, the first policy is chosen with probability $q$ and the second one is chosen with probability $1-q$. In other words, the optimal policy can be obtained as follows:

$$\delta^* = q\delta^- + (1-q)\delta^+.$$ 

This also can be viewed as: at each decision epoch, the selection to be made between the two policies is by flipping a biased coin, whose head and tail have probabilities $q$ and $1-q$, respectively. If head appears, the optimal policy will be $\delta^-$, while if tail appears, the optimal policy will be $\delta^+$. 

### 2.3 Monotone Optimal Vertical Handoff Policy

In this section, we investigate the structure of the optimal policy. We provide conditions and show that the optimal policy is monotonic (non-decreasing or non-increasing) in some system states (e.g., available bandwidth, delay, velocity). The importance of these results regarding the structure of the optimal policy are as follows:\[32\]:

- They are attractive to decision makers.
- They are easy to implement.
- They enable efficient computation.

The last point is especially valuable because when the optimal policy has a simple form, specialized algorithms can be developed to search only among policies that have
the same form as the optimal policy. A simple example of such kind of policy can be expressed of the form:

\[
\delta^*(s) = \begin{cases} 
  a_1, & \text{if } s < s^*, \\ 
  a_2, & \text{if } s \geq s^*, 
\end{cases}
\]

where \( a_1 \) and \( a_2 \) are two different actions and \( s^* \) is a threshold. This policy can be interpreted as follows. When the state of the system is less than \( s^* \), it is optimal to use action \( a_1 \); and when the state of the system is \( s^* \) or greater, it is optimal to use action \( a_2 \). If an MDP problem has this kind of structure in the optimal policy, the problem of finding an optimal policy reduces to determining the threshold \( s^* \).  

Given the proper selection of the Lagrange multiplier \( \beta \) (i.e., \( \beta^* \)), the unconstrained MDP in (2.19) has a stationary optimal policy. It can be shown that for a scenario with two wireless 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 \( \mathcal{A}_s = \{1, 2\} \) define a threshold policy. The threshold policies are optimal policies with a special structure that facilitates computation and implementation.  

2.3.1 Threshold Structure of Unconstrained MDP

Since two wireless networks are considered (i.e., \( \mathcal{M} = \{1, 2\} \)), the current system state is given by \( s = [i, b_1, d_1, b_2, d_2, v, l] \), 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.
Chapter 2. CMDP Formulation of the Vertical Handoff Decision Problem

Recall that the unconstrained MDP can be solved by using PIA. From Algorithm 1, in each iteration $k$, we have:

$$v^{k+1}_\beta(s) = \max_{a \in A_s} \left\{ r(s, a; \beta) + \sum_{s' \in S} \lambda P[s' | s, a] v^k_\beta(s') \right\}, \quad k = 0, 1, 2, \ldots \quad (2.26)$$

and

$$Q^{k+1}_\beta(s, a) = r(s, a; \beta) + \sum_{s' \in S} \lambda P[s' | s, a] v^k_\beta(s'), \quad k = 0, 1, 2, \ldots \quad (2.27)$$

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

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

If the state-action reward function $Q_\beta(s, a)$ is supermodular (or submodular) in action $a$ and another variable in the state $s$ (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 [32], [42]. Based on Definition 1, if $F(x, y)$ is supermodular (submodular) in $(x, y)$, then $y(x) = \arg \max_y F(x, y)$ is monotonically non-decreasing (non-increasing) in variable $x$ [42].
By supermodularity, we can see that the state-action reward function $Q_\beta(s, a)$ being submodular in $(b_2, a)$, and supermodular in $(d_2, a)$ and $(v, a)$ implies that the optimal vertical handoff policy $\delta_\beta(s)$ 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 two 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 straightforwardly.

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

**Proof.** Please refer to Appendix A.1

In the following theorems, without loss of generality, we assume the current network in use is network 2 (i.e., $i = 2$).

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

$$Q_\beta(i, b_1, b_2, d_2, v, l, 2) - Q_\beta(i, b_1, b_2, d_2, v, l, 1) \geq Q_\beta(i, b_1, b_2 + 1, d_2, v, l, 2) - Q_\beta(i, b_1, b_2 + 1, d_2, v, l, 1),$$

then the optimal policy is deterministic and monotonically non-increasing in the available bandwidth component of the state $s$. Consequently, $\delta_\beta^*(s)$ is of the form:

$$\delta_\beta^*(i, b_1, b_2, d_2, v, l) = \begin{cases} a \neq i, & \text{if } 0 \leq b_2 \leq \tau_b(i, b_1, d_1, d_2, v, l), \\ a = i, & \text{if } b_2 > \tau_b(i, b_1, d_1, d_2, v, l), \end{cases}$$

(2.29)

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

Proof. Please refer to Appendix [A.2]

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

$$Q_\beta(i, b_1, d_1, b_2, d_2, v, l, 2) - Q_\beta(i, b_1, d_1, b_2, d_2, v, l, 1) \leq Q_\beta(i, b_1, d_1, b_2, d_2 + 1, v, l, 2) - Q_\beta(i, b_1, d_1, b_2 + 1, v, l, 1),$$

(2.30)

then the optimal policy is deterministic and monotonically non-decreasing in the delay component of the state $s$. Consequently, $\delta_\beta^*(s)$ is of the form:

$$\delta_\beta^*(i, b_1, d_1, b_2, d_2, v, l) = \begin{cases} a = i, & \text{if } 0 \leq d_2 \leq \tau_d(i, b_1, d_1, b_2, v, l), \\ a \neq i, & \text{if } d_2 > \tau_d(i, b_1, d_1, b_2, v, l), \end{cases}$$

(2.31)
where \( \tau_d(i, b_1, d_1, b_2, v, l) \) defines the threshold for the rest of the elements of the state \( s \) for a given Lagrange multiplier \( \beta \).

Proof. Please refer to Appendix [A.3]

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

\[
Q_\beta(i, b_1, d_1, b_2, d_2, v, l, 2) - Q_\beta(i, b_1, d_1, b_2, d_2, v, l, 1) \leq Q_\beta(i, b_1, d_1, b_2, d_2, v + 1, l, 2) - Q_\beta(i, b_1, d_1, b_2, d_2, v + 1, l, 1),
\]

then the optimal policy is deterministic and monotonically non-decreasing in the velocity component of the state \( s \). Consequently, \( \delta^*_s(s) \) is of the form:

\[
\delta^*_s(i, b_1, d_1, b_2, d_2, v, l) = \begin{cases} 
    a = i, & \text{if } 0 \leq v \leq \tau_v(i, b_1, d_1, b_2, d_2, l), \\
    a \neq i, & \text{if } v > \tau_v(i, b_1, d_1, b_2, d_2, l),
\end{cases}
\]

(2.33)

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 \( s \) for a given Lagrange multiplier \( \beta \).

Proof. Please refer to Appendix [A.4]

### 2.3.2 Threshold Structure of Constrained MDP

Having shown the threshold structure of the unconstrained MDP optimal policy and based on the results described in Section 2.3.1 and 34, 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 $\delta^*$ that is the optimal solution of the CMDP equivalent to the one given in (2.15) such that $\delta^*$ is a randomized mixture of two threshold vertical handoff policies as follows:

$$\delta^*(s) = q\delta^*_- (s) + (1-q)\delta^*_+ (s), \quad s \in S,$$

(2.34)

where $\delta^*_- \text{ and } \delta^*_+$ are two unconstrained optimal policies with Lagrange multipliers $\beta^-$ and $\beta^+$ that are of the form (2.29) for the available bandwidth, (2.31) for the delay, or (2.33) for the velocity.

**Proof.** These results follow directly from Theorems 1, 2, or 3 and Theorem 4.3 of [34].

The existence of 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. [32]). These algorithms seek the optimal policy $\delta^*$ 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 [32], pp. 259] can be used. When an optimal policy is strictly increasing, the action set $A_s$ decreases in size with increasing $s$ (e.g., $b_2$) and hence reduces the number of actions which need to be evaluated by the algorithm. If at some state, say $\bar{b}_2 = \tau_b(i, b_1, d_1, d_2, v, l)$, the action set $A_s$ contains only one element (i.e., $a^*$), then no
further maximization is required because the action will be optimal for all $b_2 \geq \bar{b}_2$. Thus, the threshold value is $\bar{b}_2$ and the optimal action is $a^*$ for all $b_2 \geq \bar{b}_2$. 
Chapter 3

Numerical Results and Discussions

3.1 Comparison between Different Vertical Handoff Decision Algorithms

We compare the performance of our proposed CMDP-based vertical handoff decision algorithm with three other schemes. The first one is also based on CMDP but does not consider the impact of velocity in making the decisions (this scheme is denoted by CMDP-w/o-velocity). The second is the SAW algorithm introduced in [16], and the last algorithm we compare with is ELECTRE, which is also an MADM algorithm introduced in [19] 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).

We consider the scenario depicted in Fig. [2.1]. 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 successice decision epochs is 15 sec. The bandwidth importance weight $\omega$ 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 3.1.

Table 3.1: Summary of simulation parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$b_{1\max}$</td>
<td>Maximum available bandwidth for network 1</td>
<td>3 units</td>
</tr>
<tr>
<td>$b_{2\max}$</td>
<td>Maximum available bandwidth for network 2</td>
<td>4 units</td>
</tr>
<tr>
<td>$b_{3\max}$</td>
<td>Maximum available bandwidth for network 3</td>
<td>5 units</td>
</tr>
<tr>
<td>$C_1$</td>
<td>Access cost of the cellular network (in monetary units per bit)</td>
<td>2</td>
</tr>
<tr>
<td>$C_2$</td>
<td>Access cost of the WiMAX network (in monetary units per bit)</td>
<td>1.5</td>
</tr>
<tr>
<td>$C_3$</td>
<td>Access cost of the WLAN (in monetary units per bit)</td>
<td>1.2</td>
</tr>
<tr>
<td>$d_{1\max}$</td>
<td>Maximum packet delay for network 1</td>
<td>3 units</td>
</tr>
<tr>
<td>$d_{2\max}$</td>
<td>Maximum packet delay for network 2</td>
<td>3 units</td>
</tr>
<tr>
<td>$d_{3\max}$</td>
<td>Maximum packet delay for network 3</td>
<td>3 units</td>
</tr>
<tr>
<td>$K_{i,a}$</td>
<td>Switching cost ($i, a \in {1, 2, 3}, i \neq a$)</td>
<td>0.5</td>
</tr>
<tr>
<td>$</td>
<td>M</td>
<td>$</td>
</tr>
</tbody>
</table>
Table 3.1 – continued from previous page

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{\text{max}}$</td>
<td>Upper threshold of the MT’s velocity</td>
<td>3 units</td>
</tr>
<tr>
<td>$V_{\text{min}}$</td>
<td>Lower threshold of the MT’s velocity</td>
<td>1 unit</td>
</tr>
<tr>
<td>$v_{\text{max}}$</td>
<td>Maximum possible velocity of the MT</td>
<td>3 units</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>Memory level for the Gauss-Markov model</td>
<td>0.5</td>
</tr>
<tr>
<td>$\Theta_1$</td>
<td>Area of location type 1 (in terms of percentage of total area)</td>
<td>50%</td>
</tr>
<tr>
<td>$\Theta_2$</td>
<td>Area of location type 2 (in terms of percentage of total area)</td>
<td>25%</td>
</tr>
<tr>
<td>$\Theta_3$</td>
<td>Area of location type 3 (in terms of percentage of total area)</td>
<td>25%</td>
</tr>
<tr>
<td>$\kappa$</td>
<td>Risky index of the user</td>
<td>0.5</td>
</tr>
<tr>
<td>$\mu$</td>
<td>Mean value of the MT’s velocity</td>
<td>1 unit</td>
</tr>
<tr>
<td>$\rho_1 : \rho_2 : \rho_3$</td>
<td>Ratio between user densities</td>
<td>1:1:8</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Standard deviation of the MT’s velocity</td>
<td>0.1 unit</td>
</tr>
<tr>
<td>$\phi$</td>
<td>Importance weight for switching cost</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Note the bandwidth importance weight $\omega$ 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.

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 (2.25) is calculated for different discount factors (i.e., different average connection durations). Specifically, for \( \lambda \) equals to \([0.9, 0.95, 0.966, 0.975, 0.98]\), the corresponding 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 \( \lambda \) equals to \([0.9, 0.95, 0.966, 0.975, 0.98]\), the corresponding predefined constraints \( C_{\text{max}} \) are \([2, 4, 6, 8, 10]\).

### 3.1.1 Results for CBR Voice Traffic over UDP

The expected total reward of a user under different discount factors for CBR traffic is shown in Fig. 3.1. For all the four schemes considered here, the expected total reward increases as \( \lambda \) becomes larger. This is because the larger \( \lambda \) 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 \( \lambda \) equals to 0.975 (i.e., the average duration of connection is 10 mins), for which the predefined constraint is 8 monetary units, the CMDP algorithm achieves 25% higher expected total reward than the CMDP-w/o-velocity algorithm, 93% higher expected total reward than the SAW scheme, and 199% higher expected total reward than the ELECTRE algorithm. The reason is that when an MT does not con-
Figure 3.1: Expected total reward under different discount factor $\lambda$ for CBR traffic.

Consider its velocity, the connection might be dropped during the handoff process and needs to be re-established. The associated QoS degradation and extra signaling and processing costs decrease the actual reward it will gain by performing the handoff. SAW and ELECTRE algorithms achieve lower reward than the CMDP-w/o-velocity scheme, because they do not consider the user’s budget for the connection, neither 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. 3.2 shows the expected total reward of a user versus the mean of its velocity un-
Figure 3.2: Expected total reward under different mean of user’s velocity $\mu$ 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. For the CMDP-w/o-velocity algorithm, the expected total reward decreases as the user’s velocity increases. The reason is that as the user’s velocity becomes larger, the decrease in the actual reward (i.e., QoS degradation and extra signaling and processing costs) associated with the issue that the model does not consider the effect of user’s velocity becomes more significant. The SAW and ELECTRE algorithms still
The expected total reward a user can obtain versus its budget on the expected total access cost for CBR traffic is shown in Fig. 3.3. 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 algorithm always achieves higher reward than the CMDP-w/o-velocity, 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 CMDP-w/o-velocity, SAW, and ELECTRE schemes are reduced because...
Chapter 3. Numerical Results and Discussions

Figure 3.4: Expected total reward under different switching cost $K_{i,a}$ for CBR traffic.

of the dropped connections.

Fig. 3.4 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 four schemes decrease. The expected total reward of the CMDP algorithm decreases slower than the CMDP-w/o-velocity algorithm does. This is because as the switching costs increase, the decrease on the actual reward achieved by the CMDP-w/o-velocity scheme is also larger. Since for the same number of dropped calls, the extra signaling and processing costs increase as $K_{i,a}$ becomes larger.

Fig. 3.5 shows the expected total reward of a user versus the access cost of the cellular
Figure 3.5: Expected total reward under different access cost of the cellular network $C_1$ for CBR traffic.

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 become smaller for all four 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. From another point of view, this can also be seen as the user’s budget becomes smaller. Thus, the expected total reward of the user decreases. For the same constraint on the expected total access cost, the CMDP scheme achieves better expected total reward than the CMDP-w/o-velocity, SAW, and ELECTRE schemes.
Figure 3.6: Expected total reward under different mean of user’s velocity $\mu$ for FTP traffic.

3.1.2 Results for FTP Data Traffic over TCP

Fig. 3.6 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 ($\mu = 24 \text{ km/h}$), 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.
Figure 3.7: Expected total reward under different user’s budget on expected total access cost $C_{\text{max}}$ for FTP traffic.

The expected total reward a user can obtain versus its budget on the expected total access cost for FTP traffic is shown in Fig. 3.7. Similar to the CBR traffic case, we can see for the same budget, the CMDP algorithm always achieves higher expected total reward than the CMDP-w/o-velocity, 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. 3.3, however for CBR traffic this decrease is less noticeable.

Fig. 3.8 shows the expected total reward of a user versus the access cost of the cellular network under a budget of 8 monetary units for FTP traffic. For the same constraint on the expected total access cost, the CMDP scheme achieves higher expected total reward.
Figure 3.8: Expected total reward under different access cost of the cellular network $C_1$ for FTP traffic.

than the CMDP-w/o-velocity, SAW, and ELECTRE schemes. Note for FTP traffic, the reward does not decrease as much as in the CBR traffic case shown in Fig. 3.5. The reason is that since FTP traffic is more sensitive to bandwidth, the MT will sojourn more time in the WiMAX network and WLAN (i.e., less time in the cellular network) than in the CBR traffic case, such that it can take advantages of their higher bandwidth. As a result, the increase of the cellular network access cost does not influence the expected total reward in the FTP traffic case as much as in the CBR traffic case.
3.2 Structure of the Optimal Policy

Fig. 3.9 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 $s$ 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 $b_3 < 3$. 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 $b_2 < 3$. 

**Figure 3.9:** Structure of the unconstrained MDP optimal vertical handoff policy on the available bandwidth when $s = [i = 3, b_1 = 1, d_1 = 3, d_2 = 1, d_3 = 3, v = 2, l = 3]$. 

Chapter 3. Numerical Results and Discussions 55
Figure 3.10: Structure of the unconstrained MDP optimal vertical handoff policy on the packet delay when $s = [i = 2, b_1 = 1, b_2 = 4, b_3 = 2, d_3 = 3, v = 1, l = 3]$.

Fig. 3.10 shows the structure of the unconstrained MDP optimal vertical handoff policy by varying the delay of network 1 and network 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 $d_1 < 3$. Similarly, for a fixed $d_1$ (e.g., $d_1 = 2$), the optimal policy chooses network 1 when $d_2 \geq 3$, and selects network 2 when $d_2 < 3$.

Fig. 3.11 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
Figure 3.11: Structure of the unconstrained MDP optimal vertical handoff policy on the mean of user’s velocity when $s = [i = 3, b_1 = 3, d_1 = 3, b_2 = 4, d_2 = 1, d_3 = 3, l = 3]$. MT’s velocity. The condition of the current state $s$ 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 handoff when $v \geq 2$, and chooses to handoff when $v < 2$. If $b_3 = 2$, the optimal policy does not perform vertical handoff when $v \geq 3$, and chooses to handoff when $v < 3$. 
Chapter 4

Conclusions and Future Work

4.1 Conclusions

In this thesis, 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.

\section*{4.2 Future Work}

For future possible extension of this work, the Semi-Markov decision process (SMDP) can be used to formulate the vertical handoff decision problem. In the current work, the time between each decision epoch (i.e., state transition time) is fixed. By using the SMDP, the time between state transitions may follow an arbitrary probability distribution. A special case of the SMDP is the continuous-time Markov decision process (CTMDP), in which the inter-transition times are exponentially distributed. By using this algorithm, a more realistic system model can be formulated.

Another extension of this work can be in the direction of considering more constraints. For example, in addition to the constraint that users have monetary budgets for their connections, the power consumed for making a call should be less than some threshold too. In this case, the problem becomes a CMDP with multiple constraints \cite{43}.
Bibliography


Appendix A

Proofs of Lemma 1 and Theorems 1, 2, and 3

A.1 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 $r(s, a)$ is monotone in available bandwidth, delay, and velocity;

2. To prove the sum of transition probabilities $\sum_{s' \in S} P[s' \mid s, a]$ is monotone in available bandwidth, delay, and velocity.

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

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

For the velocity, the only part that relates to it in the reward function is \(-q(s, a)\). As we can see from the definition of \( q(s, a) \) \((2.9)\), when the velocity \( v \) becomes larger, the value of \( q(s, a) \) becomes larger too (or stays the same), which means \(-q(s, a)\) becomes smaller (or stays the same). Consequently, the reward function \( r(s, a) \) is monotonically non-increasing in velocity.

Second, we assume the transition probability function \( P[s' \mid s, a] \) 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. Specifically, take bandwidth for example, this implies the probability that the bandwidth in the current decision epoch exceeds some value will be greater when the bandwidth in the previous decision epoch is greater. In other words, the bandwidth in the next decision epoch is stochastically increasing with respect to the bandwidth in the current decision epoch. This is the condition under which the sum of transition probabilities (i.e., \( \sum_{s' \in S} P[s' \mid s, a] \)) is monotonically non-decreasing in the available bandwidth. Similarly, that 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., \( \sum_{s' \in S} P[s' \mid s, a] \))
is monotonically non-increasing in the delay (velocity).

\[ A.2 \text{ Proof of Theorem 1} \]

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

\[
Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], 2) - Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], 1)
= r([i, b_1, d_1, b_2, d_2, v, l], 2; \beta) - r([i, b_1, d_1, b_2, d_2, v, l], 1; \beta)
+ \sum_{s' \in S} \lambda P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2] P[v' | v] P[l' | l]
\times \left( v^k_\beta(j, b'_1, d'_1, b'_2, d'_2, v', l') - v^k_\beta(j, b'_1, b'_2, d'_2, v', l') \right),
\]

(A.1)

is monotonically non-increasing in 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 \( v^0_\beta(j, b'_1, d'_1, b'_2, d'_2, v', l') \) with non-increasing difference in \( b_2 \). Assume that \( v^k_\beta(j, b'_1, d'_1, b'_2, d'_2, v', l') \) has non-increasing difference in \( b_2 \), which implies that \( Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a) \) is submodular in \((b_2, a)\). We will now prove that \( v^{k+1}_\beta(j, b'_1, d'_1, b'_2, d'_2, v', l') \) also has non-increasing difference in \( b_2 \). That is,

\[
v^{k+1}_\beta(i, b_1, d_1, b_2 + 1, d_2, v, l) - v^{k+1}_\beta(i, b_1, d_1, b_2, d_2, v, l) \leq v^{k+1}_\beta(i, b_1, d_1, b_2, d_2, v, l) - v^{k+1}_\beta(i, b_1, d_1, b_2 - 1, d_2, v, l),
\]

(A.2)

or

\[
v^{k+1}_\beta(i, b_1, d_1, b_2 + 1, d_2, v, l) - v^{k+1}_\beta(i, b_1, d_1, b_2, d_2, v, l)
- (v^{k+1}_\beta(i, b_1, d_1, b_2, d_2, v, l) - v^{k+1}_\beta(i, b_1, d_1, b_2 - 1, d_2, v, l)) \leq 0.
\]

(A.3)
Appendix A. Proofs of Lemma 1 and Theorems 1, 2, and 3

Let's assume:

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

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

and

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

for some actions \( a_2, a_1, a_0 \in \mathcal{A}_s \). Thus, we can re-write (A.2) as:

\[
Q^{k+1}_\beta([i, b_1, d_1, b_2 + 1, d_2, v, l], a_2) - Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_1)
- \left( Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_1) - Q^{k+1}_\beta([i, b_1, d_1, b_2 - 1, d_2, v, l], a_0) \right) \leq 0,
\]

or

\[
\begin{align*}
W_1 = & Q^{k+1}_\beta([i, b_1, d_1, b_2 + 1, d_2, v, l], a_2) - Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_2) \\
& + Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_2) - Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_1) \\
X_1 = & -Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_1) + Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_0) \\
& - \left( Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a_0) - Q^{k+1}_\beta([i, b_1, d_1, b_2 - 1, d_2, v, l], a_0) \right) \leq 0,
\end{align*}
\]

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:

\[
W_1 = r([i, b_1, d_1, b_2 + 1, d_2, v, l], a_2; \beta) - r([i, b_1, d_1, b_2, d_2, v, l], a_2; \beta) \\
+ \sum_{s' \in S} \lambda \left( P[b'_1, d'_1 | b_1, d_1] \sum_{b_2} P[b'_2, d'_2 | b_2 + 1, d_2] P[\v' | v] P[\l' | l] v^k_\beta(j, b'_1, d'_1, (b_2 + 1)', d'_2, \v', \l') \right) \\
- P[b'_1, d'_1 | b_1, d_1] \sum_{b_2} P[b'_2, d'_2 | b_2, d_2] P[\v' | v] P[\l' | l] v^k_\beta(j, b'_1, d'_1, b_2, d'_2, \v', \l') \\
\leq r([i, b_1, d_1, b_2, d_2, v, l], a_0; \beta) - r([i, b_1, d_1, b_2 - 1, d_2, v, l], a_0; \beta) \\
+ \sum_{s' \in S} \lambda \left( P[b'_1, d'_1 | b_1, d_1] \sum_{b_2} P[b'_2, d'_2 | b_2 - 1, d_2] P[\v' | v] P[\l' | l] v^k_\beta(j, b'_1, d'_1, (b_2 - 1)', d'_2, \v', \l') \right).
\]

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 \( v^{k+1}_\beta(i, b_1, b_2, d_2, v, l) \) satisfies \( (A.2) \), which implies that \( Q_\beta(s, a) \) is submodular in \((b_2, a)\). \hfill \square

## A.3 Proof of Theorem 2

To show that the optimal policy is monotonically non-decreasing in the delay, we need to prove inductively that \( Q_\beta(s, a) \) is supermodular in \((d_2, a)\). We will prove via mathematical induction that, for a suitable initialization, \((A.1)\) 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 \( v^0_\beta(j, b'_1, d'_1, b'_2, d'_2, \v', \l') \) with non-increasing difference in \( d_2 \). Assume that \( v^k_\beta(j, b'_1, d'_1, b'_2, d'_2, \v', \l') \) has non-increasing difference in \( d_2 \), which implies that \( Q^{k+1}_\beta([i, b_1, d_1, b_2, d_2, v, l], a) \) is supermodular in \((d_2, a)\).

We will now prove that \( v^{k+1}_\beta(j, b'_1, d'_1, b'_2, d'_2, \v', \l') \) also has non-increasing difference in
Appendix A. Proofs of Lemma 1 and Theorems 1, 2, and 3

\[ v_{\beta}^{k+1}(i, b_1, d_1, b_2, d_2 + 1, v, l) - v_{\beta}^{k+1}(i, b_1, d_1, b_2, v, l) \leq v_{\beta}^{k+1}(i, b_1, d_1, b_2, d_2 - 1, v, l) \]  \hspace{1cm} (A.5)

or

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

\[ - (v_{\beta}^{k+1}(i, b_1, d_1, b_2, d_2, v, l) - v_{\beta}^{k+1}(i, b_1, b_2, d_2 - 1, v, l)) \leq 0. \]  \hspace{1cm} (A.6)

Let’s assume:

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

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

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 \(a_2, a_1, a_0 \in \mathcal{A}.\) Thus, we can re-write (A.5) as:

\[ Q_{\beta}^{k+1}([i, b_1, d_1, b_2, d_2 + 1, v, l], a_2) - Q_{\beta}^{k+1}([i, b_1, d_1, b_2, d_2, v, l], a_1) \]

\[ - (Q_{\beta}^{k+1}([i, b_1, d_1, b_2, d_2, v, l], a_1) - Q_{\beta}^{k+1}([i, b_1, d_1, b_2, d_2 - 1, v, l], a_0)) \leq 0. \]  \hspace{1cm} (A.7)

or

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

where \( X_2 \leq 0 \) and \( Y_2 \leq 0 \) 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:

\[
W_2 = r([i, b_1, d_1, b_2, d_2 + 1, v, l, a_2; \beta]) - r([i, b_1, d_1, b_2, d_2, v, l, a_2; \beta]) \\
+ \sum_{s' \in S} \lambda \left( P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2 + 1] P[v' | v] P[l' | l] v_{\beta}^k(j, b'_1, d'_1, b'_2, (d_2 + 1)', v', l') \\
- P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2] P[v' | v] P[l' | l] v_{\beta}^k(j, b'_1, d'_1, b'_2, v', l') \right) \\
\leq r([i, b_1, d_1, b_2, d_2 - 1, v, l, a_0; \beta]) \\
+ \sum_{s' \in S} \lambda \left( P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2 - 1] P[v' | v] P[l' | l] v_{\beta}^k(j, b'_1, d'_1, b'_2, (d_2 - 1)', v', l') \right).
\]

The RHS of the inequality comes from the expansion of \( Z_2 \) which implies that \( W_2 \leq Z_2 \).

Therefore, it is shown that \( v_{\beta}^{k+1}(i, b_1, d_1, b_2, d_2, v, l) \) satisfies (A.5), which implies that \( Q_{\beta}(s, a) \) is supermodular in \((d_2, a)\).

### A.4 Proof of Theorem 3

To show that the optimal policy is monotonically non-decreasing in the velocity, we need to prove inductively that \( Q_{\beta}(s, a) \) is supermodular in \((v, a)\). We will prove via mathematical induction that, for a suitable initialization, (A.1) 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 \( v \).  As-
sume that \( v^k_{\beta}(j, b'_1, d'_1, b'_2, d'_2, v', l') \) has non-increasing difference in \( v \), which implies that 

\[ Q^{k+1}_{\beta}([i, b_1, d_1, b_2, d_2, v, l], a) \]

is supermodular in \((v, a)\).

We will now prove that \( v^{k+1}_{\beta}(j, b'_1, d'_1, b'_2, d'_2, v', l') \) also has non-increasing difference in \( v \). That is,

\[
v^{k+1}_{\beta}(i, b_1, d_1, b_2, d_2, v + 1, l) - v^{k+1}_{\beta}(i, b_1, d_1, b_2, d_2, v, l) \leq v^{k+1}_{\beta}(i, b_1, d_1, b_2, d_2, v - 1, l),
\]

or

\[
v^{k+1}_{\beta}(i, b_1, d_1, b_2, d_2, v + 1, l) - v^{k+1}_{\beta}(i, b_1, d_1, b_2, d_2, v, l)
- (v^{k+1}_{\beta}(i, b_1, d_1, b_2, d_2, v, l) - v^{k+1}_{\beta}(i, b_1, d_1, b_2, d_2, v - 1, l)) \leq 0.
\]

Let’s assume:

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

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

and

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

for some actions \( a_2, a_1, a_0 \in \mathcal{A}_s \). Thus, we can re-write (A.8) as:

\[
Q^{k+1}_{\beta}([i, b_1, d_1, b_2, d_2, v + 1, l], a_2) - Q^{k+1}_{\beta}([i, b_1, d_1, b_2, d_2, v, l], a_1)
- \left( Q^{k+1}_{\beta}([i, b_1, d_1, b_2, d_2, v, l], a_1) - Q^{k+1}_{\beta}([i, b_1, d_1, b_2, d_2, v - 1, l], a_0) \right) \leq 0,
\]

or

\[
\underbrace{Q^{k+1}_{\beta}([i, b_1, d_1, b_2, d_2, v + 1, l], a_2) - Q^{k+1}_{\beta}([i, b_1, d_1, b_2, d_2, v, l], a_2)}_{w_3}
\]
Therefore, it is shown that the RHS of the inequality comes from the expansion of $W$ where $a_1 = 3$.

In addition, it follows from the induction hypothesis that:

\[
\begin{align*}
W_3 &= r([i, b_1, d_1, b_2, d_2, v + 1, l], a_2; \beta) - r([i, b_1, d_1, b_2, d_2, v, l], a_2; \beta) \\
&+ \sum_{s' \in S} \lambda \left( P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2] P[v' | v + 1] P[l' | l] v_{\beta}^{k}(j, b'_1, d'_1, b'_2, d'_2, (v + 1)', l') \\
&- P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2] P[v' | v] P[l' | l] v_{\beta}^{k}(j, b'_1, d'_1, b'_2, d'_2, v', l') \right) \\
&\leq r([i, b_1, d_1, b_2, d_2, v, l], a_0; \beta) - r([i, b_1, d_1, b_2, d_2, v - 1, l], a_0; \beta) \\
&+ \sum_{s' \in S} \lambda \left( P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2] P[v' | v] P[l' | l] v_{\beta}^{k}(j, b'_1, d'_1, b'_2, d'_2, v', l') \\
&- P[b'_1, d'_1 | b_1, d_1] P[b'_2, d'_2 | b_2, d_2] P[v' | v - 1] P[l' | l] v_{\beta}^{k}(j, b'_1, d'_1, b'_2, d'_2, (v - 1)', l') \right).
\end{align*}
\]

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 \(A.\), which implies that $Q_{\beta}(s, a)$ is supermodular in $(v, a)$. \(\blacksquare\)