Efficient and QoS Guaranteed Data Transport in Heterogeneous Wireless Mobile Networks

A Dissertation
Presented to
The Academic Faculty

by

Sung-Eun Kim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology

May 2006

Efficient and QoS Guaranteed Data Transport
in Heterogeneous Wireless Mobile Networks

Approved by:

Dr. John A. Copeland, Advisor
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Henry L. Owen
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Gee-Kung Chang
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Krishna V. Palem
School of Electrical and Computer Engineering
Georgia Institute of Technology

Dr. Jun Xu
College of Computing
Georgia Institute of Technology

Date Approved: April 5, 2006

To my family

for their tremendous love, support, and belief in me.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my advisor, Dr. John A. Copeland, for his valuable guidance and support throughout this research. Besides providing excellent academic guidance, he has been incredibly encouraging, supportive and patient. It has been a great pleasure to have him as an advisor.

I would like to express my appreciation to my dissertation committee members, Dr. Henry L. Owen, Dr. Gee-Kung Chang, Dr. Krishna V. Palem, and Dr. Jun Xu for reviewing this dissertation and providing constrictive comments.

I am grateful to all the members of the Communications Systems Center (CSC) for their friendship. Conversations with the colleagues made me cheerful and technical discussions helped me to progress in this work. Special thanks goes to Cherita Corbett and Kulsoom Abdullah for their help, especially during the last stage of this work. Many thanks go to Dr. Raheem Beyah at Georgia State University, an alumnus of CSC, for his concern about my career as well as my research. I also thank to Ms. Kathy Cheek who has greatly helped me in administrative affairs.

Finally, I would like to give the deepest gratitude to my family. I am grateful to my parents, parents-in-law, sisters and brothers for their unlimited support, encouragement, and belief in me. Special thanks are due to my husband, Chang-Ho Lee, and my son, David Woochang Lee, for their love, patience, sacrifice, and continuous motivation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

cwnd	Congestion Window
ssthresh	Slow start threshold
ACK	Acknowledgement
AIMD	Additive Increase, Multiplicative Decrease
BS	Base Station
CA	Congestion Avoidance
CAC	Call Admission Control
CDF	Cumulative Distribution Function
DUP_ACK	Duplicate Acknowledgement
FH	Fixed Host
FTP	File Transfer Protocol
GC	Guard Channel
IP	Internet Protocol
MAC	Medium Access Control
MH	Mobile Host
QoS	Quality of Service
RRC	Radio Resource Control
RSS	Received Signal Strength
RTO	Retransmission Time Out
RTT	Round Trip Time
RTT_VAR	RTT_variation

SACK	Selective Acknowledgement
SL	Service Level
SLA	Service Level Agreement
SRTT	Smoothed RTT
SS	Slow Start
TCP	Transport Control protocol
3G	Third Generation
3 GPP	Third Generation Partnership Project
WCDMA	Wideband Code Division Multiple Access
WLAN	Wireless Local Area Network
WMAN	Wireless Metropolitan Area Networks

SUMMARY

The objective of this research is to investigate and develop an efficient and seamless data transport protocol for a heterogeneous wireless mobile network.

In next-generation network, most of heterogeneous wireless mobile networks will be combined and complementarily constitute a hierarchical network. To integrate different networks, many challenging issues should be solved. In this thesis, an efficient and seamless data transport mechanisms are explored.

We investigate the problems that the current transport control protocol (TCP) will experience within the heterogeneous mobile network. In a heterogeneous network, a mobile host experiences drastic changes in network condition during a session. Traditional TCP struggles with abrupt network changes by intersystem handoff and cannot work efficiently in this environment. We propose a TCP scheme to be tailored to the heterogeneous mobile network to support seamless data transport. In the proposed scheme, a TCP is informed the impending handoff events and works differently based on a handoff type. Simulation results present the proposed algorithm improves throughput, stabilizes data transmission rapidly, and provides a seamless data transfer.

We also propose an adaptive resource management scheme within a 3G cellular network based on a user's priority level to reduce the call dropping and blocking rates. In a heterogeneous network, a network that provides smaller bandwidth may struggle with handed-off calls being served with a higher bandwidth. Therefore, a resource management algorithm should be defined so that an ongoing call is not dropped by a handoff and provides seamless data transfer. We propose an adaptive resource management scheme based on downgrading the quality of some existing services in a 3G cellular network. We analyze the system capacity, call blocking rate and call dropping rate of the proposed algorithm, and simulate the performance variation of the downgraded traffic. The results show that the proposed scheme increases system capacity, and decreases the call dropping rate at the cost of small delay of the downgraded data traffic.

CHAPTER 1

INTRODUCTION

1.1 Motivation

The number of mobile service users has been tremendously increased during the last few years, and the types of mobile services are diversified from pure voice to data and multimedia services, such as audio/video streaming, email, file transfer, and web browsing. These services have quite different requirements for service quality. Some services supporting real time applications have stricter requirements for delay and delay jitter, while some services for non-real time data applications are stricter for low error rate. Some services need low bandwidth while some need wide bandwidth. As a result of the service diversity, mobile users need diverse quality of service (QoS) and mobility requirements. To answer this demand, various wireless access networks such as third generation (3G) cellular networks, wireless local area networks (WLAN), wireless metropolitan area networks (WMAN), ad Hoc networks, and sensor networks have been actively researched and developed. These networks differ in terms of physical link characteristics, medium access control (MAC) technology, coverage, capacity, network architecture, cost, service data rate, and target application. Mobile users are not equally distributed in a service area. As a result, hot spot areas exist, such as airports, hotels, or large buildings. No one access system can cost effectively provide all the diverse services. These needs justify the development of a heterogeneous mobile wireless network

In next-generation networks, most of the heterogeneous wireless networks will be combined and constitute a hierarchical network, and operate complementarily. In a hierarchical network, a larger mobile network supporting lower bandwidth services overlays the smaller mobile networks supporting higher bandwidth services. Mobile hosts (MH) have multiple physical interfaces and select their active physical interface based on the received signal strength (RSS), the location and velocity of the MH, service requirements, and network load. A mobile user should be able to roam between different wireless networks while maintaining ongoing sessions and transporting data seamlessly.

Currently the WLANs and cellular networks have been immensely deployed, and the two networks are independently developed to support different services. WLANs provide relatively high data rates with limited mobility and the main applications are data services. On the other hand, cellular networks provide lower data rate mainly supporting voice and short message services with high mobility. 3G cellular networks will be widely deployed in the near future, and can provide well-defined quality of multimedia services with up to 2 Mbps data rate.

The WLANs and cellular networks can complement each other by acquiring the merits of each network and overcoming the drawbacks. Current trends show that 3G cellular networks and WLANs will co-exist and interwork each other to improve system efficiency and performance. Creating complimentary WLANs and 3G cellular networks and providing seamless data service to mobile users includes many technical challenges. Mobility management and intersystem handoff techniques with the proper network selection scheme should be defined for seamless data transfer between different networks. Resource assignment and call admission control schemes should be efficiently defined. MAC and scheduling algorithms are also important. Security, authentication, accounting and billing policies should be determined. There are also various topics of research to guarantee predefined QoS level.

Current transport protocols were not designed to work well within a mobile network because unstable link characteristics or user mobility were not considered. The transport control protocol (TCP) is the most dominant transport protocol in these days. The TCP has been performing very well in the current Internet, however should be redesigned to operate in a heterogeneous mobile network. Within a heterogeneous network, a MH experiences drastic changes in network environment by an intersystem handoff. Traditional TCP will struggle with the intersystem handoff and have significant performance degradation. Therefore, developing an efficient transport protocol is crucial to provide seamless data transfer for the network.

When a MH is handed off to a different type of network, the MH might not get enough resources to continue the service. In particular, when a handoff occurs from a WLAN to a cellular network, a MH may have high probability of call drop because a cellular network has less resource than a WLAN. A 3G cellular network may also struggle with handed-off calls from a WLAN because the handed-off calls have been working with higher bandwidth in a WLAN. Therefore, adaptive resource management scheme in a cell is a challenging issue to decrease the handoff failure and to efficiently use limited resources. Moreover, future mobile networks will provide different types of multimedia services and each service has different quality requirements. Therefore, definition of adaptive resource management is crucial for the heterogeneous mobile network.

1.2 Research objectives and contributions

The objective of this research is to develop an efficient and seamless data transport protocol for heterogeneous mobile network.

To achieve the objective, we investigate the heterogeneous mobile network, which will be deployed in a near future and define the research challenges for seamless data transmission within the network. We also study the basic TCP algorithms and some problems that the current TCP experiences within a wireless network and a mobile cellular network.

Then, we evaluate the effect on the TCP performance by a handoff. Especially, we analyze the variation of the retransmission time out (RTO) value caused by an intersystem handoff, which causes abrupt changes in network environment. We also analyze how the RTO variation affects to the throughput. Then we develop a transport protocol that enhances the TCP performance and provides seamless data transport for a MH, which is roaming within a heterogeneous mobile network. We simulate the proposed scheme and demonstrate the proposed scheme improves TCP performance by comparing the results with other schemes.

We also research QoS management in a wideband code division multiple access (WCDMA) network, which is a 3G cellular network specified by the third generation partnership project (3GPP), to reduce call dropping by intersystem handoff. In a heterogeneous network, each network has different capacity and provides different bandwidth to a user. Especially, a network that provides lower bandwidth should have an adaptive resource management algorithm to support handoff calls being served with a higher bandwidth. We study QoS requirements and rate control in WCDMA network.

Then, a resource management algorithm based on user's priority level is proposed to decrease a call drop of handoff calls and a call block of new calls. We analyze the call blocking rate and call dropping rate to present performance enhancements. We also simulate the proposed scheme to show the performance variation of QoS attributes.

1.3 Thesis outline

The remainder of the thesis is organized as follows.

Chapter 2 discusses some issues in heterogeneous wireless network. First, we describe network architecture of heterogeneous wireless mobile network and present the types of handoff within the network. Then, we define research issues to guarantee seamless data transfer among different mobile networks.

Chapter 3 describes TCP algorithms. We explain the basic TCP algorithm, which is originally designed for reliable wired network. We present problems that the TCP has within a wireless network and shows proposed schemes to address the problems. We also investigate some problems of TCP within a cellular network, which has a large error rate and long delay.

Chapter 4 proposes a TCP scheme tailored to the heterogeneous mobile network to support seamless data transfer. First, we introduce Freezing TCP algorithm proposed for intra-system handoff. For seamless data transfer during an inter-system handoff, we study variation of round trip time (RTT) and RTO by the handoff. We also research how the RTO value varies with different parameters and how the TCP throughput varies according to the RTO variation. Then we explain a multimode MH for heterogeneous mobile network and define message flows that should be implemented in the MH. And then, we develop a TCP for heterogeneous mobile network and show the evaluation results of the proposed TCP.

In Chapter 5, a QoS management scheme in WCDMA network is proposed. WCDMA network has well-defined QoS requirement. We first introduce QoS attributes and classes defined in WCDMA system. We explain how the WCDMA system can dynamically manage the data rate according to the network environment and briefly survey interference-based call admission control algorithm. Then we propose an algorithm that manages the resources within WCDMA network to decrease call dropping rate. We present the results of performance evaluation including system capacity, call dropping rate, and transmission delay for a flow.

Finally, Chapter 6 summarizes the research results and contributions, and suggests future research directions.

CHAPTER 2

HETEROGENEOUS MOBILE NETWORKS

2.1 Heterogeneous mobile networks

Wireless network technologies are developing rapidly and widely. Wireless networks are evolving to provide diverse types of services and traffic such as data messaging, web browsing, file transfer, video, multimedia streaming, as well as traditional voice. The networks should provide seamless service while a user is roaming with high mobility and be able to guarantee predefined levels of QoS allowing the diverse service types.

Currently, WLAN and cellular networks are the most dominant wireless networks. The followings are some possible integrated networks, which would be deployed in the near future :

- a network integrating different types of cellular networks such as GSM/2.5G General Packet Radio Service (GPRS) networks and 3G cellular networks such as WCDMA and CDMA-2000 systems ;
- a network integrating different wireless networks based on IEEE 802 standards such as WLAN implemented by 802.11b, 802.11a, 802.11g and WMAN by 802.16 standards ;
- a network integrating WLANs and 3G cellular networks.

A WLAN system is a cost-effective alternative to cellular access networks in hot spot areas. However, the network integrating different types of WLAN systems has a mobility limitation, and the network integrating 2.5G and 3G cellular networks has a bandwidth limitation and high investment cost. Therefore, there is a strong need and consensus to integrate WLANs with 3G cellular networks and develop heterogeneous mobile data networks capable of ubiquitous data services and very high data rates [1]. Figure 1 shows a network architecture that integrates a WLAN and a 3G cellular network.

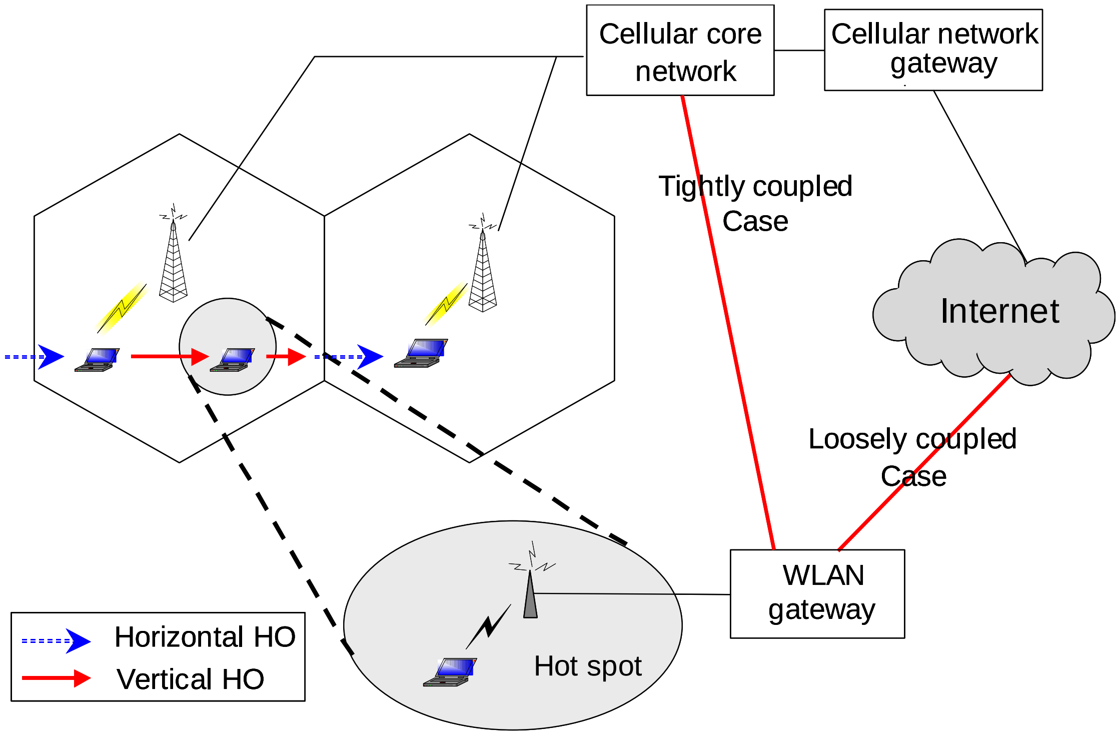


Figure1. Network architecture for interworking between the WLAN and 3G cellular networks.

As shown in Figure 1, the 3G cellular network covers a wider area and the WLAN is used for a hot spot area. A number of WLAN networks may be connected to each other to cover a wider area.

Depending on the degree of inter-dependence between a WLAN and a 3G cellular network, the network architecture can be defined as a tightly coupled integration or a loosely coupled integration [2]. In the tightly coupled integration, a WLAN is connected to a 3G cellular core network as a 3G access network. Therefore these two networks have to share authentication, mobility management, and radio resource management schemes. The 3G core network should dynamically manage the radio resources of the WLAN as well as those of the cellular network. As a consequence, the WLAN has to communicate with the 3G core network using pre-defined control messages. Currently, WLANs are already deployed and widely used. To implement the tightly coupled integration, the WLAN has to be modified so that it can be controlled by a cellular network, and this gives a higher processing load to the cellular core network. However, this approach provides shorter handoff latency and more flexible QoS management.

In the loosely coupled integration, a WLAN does not have a direct link to a 3G cellular network, and a minimum functionality is needed for interworking, such as Mobile IP. The handoff latency increases with this approach because it requires a longer processing time for authentication, and some packet drops may occur during the handoff period. The loosely coupled integration allows independent deployment of WLANs and 3G cellular networks [16]. Therefore, this loosely coupled integration is more feasible considering the current situation, with WLANs already deployed over wide areas.

Providing seamless data service to mobile users requires the implementation of an intersystem handoff between WLAN and 3G cellular networks. Since a MH may experience frequent handoffs while it moves within this network, a smooth and seamless handoff is crucial. In the heterogeneously integrated mobile network, handoffs should be classified based on a MH's movement. In [3], the authors define a horizontal handoff as a handoff between base stations (BSs) that use the same type of wireless network interface, which is the traditional definition of a handoff for homogeneous cellular systems. The authors also define a vertical handoff as a handoff between BSs that use different wireless network technologies such as WLANs and 3G cellular networks. They divide vertical handoffs into upward vertical handoffs and downward vertical handoffs. An upward vertical handoff is a handoff from a smaller network with higher bandwidth to a larger network with lower bandwidth. A downward vertical handoff is that from a larger network to a smaller network. Figure 2 shows the definitions of the handoffs. The downward and upward handoffs have a different impact on data transmission. Therefore, these handoffs must be handled separately. The vertical handoff and horizontal handoff are also called as an intersystem handoff and an intrasystem handoff, respectively. In Figure 1, a downward vertical handoff occurs when a MH moves from a cellular network to a WLAN, and an upward vertical handoff occurs when it moves from a WLAN to a cellular network.

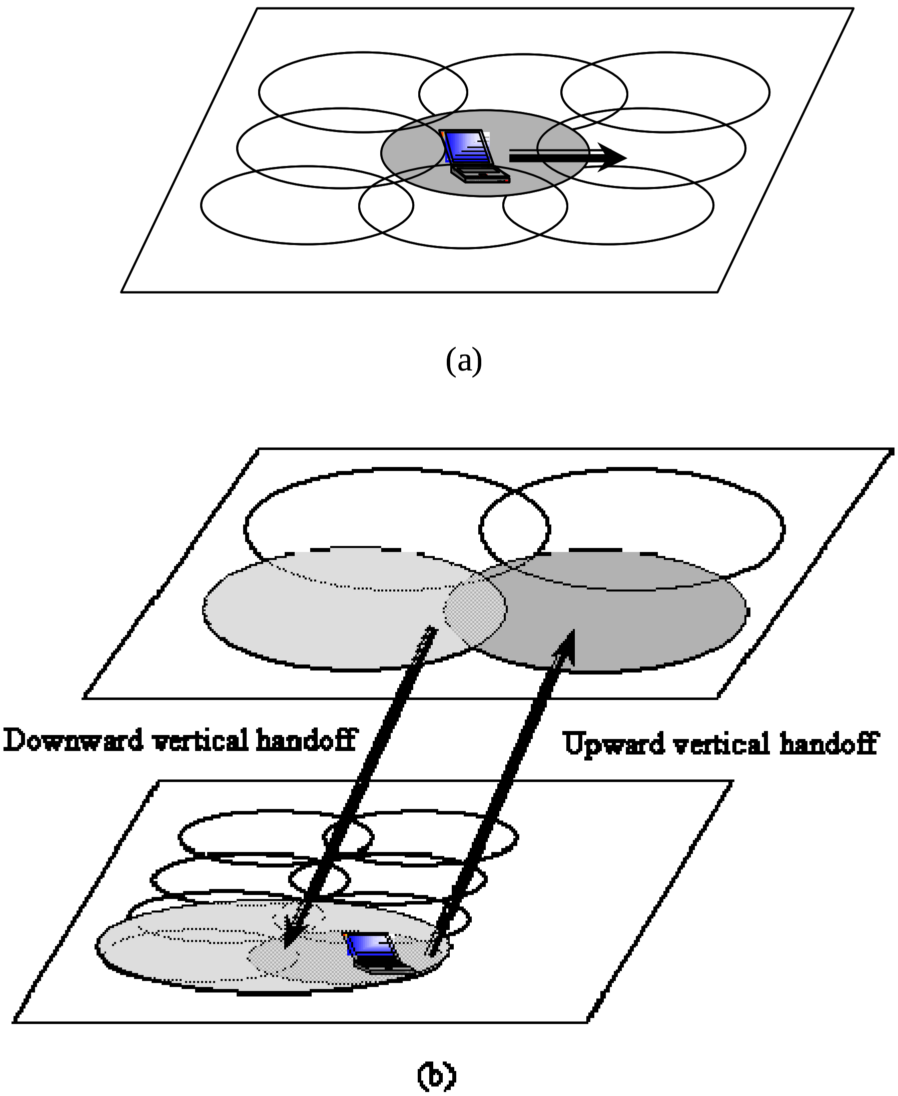


Figure 2. Types of handoffs within a heterogeneous mobile network (a) horizontal handoff, (b) vertical handoff : downward and upward.

2.2 Challenges of heterogeneous mobile networks

The interaction among the different mobile access networks poses several challenges [2][4][6][7][28][29][30]. One challenge is to implement a seamless mobility management scheme, which massive numbers of research proposals have addressed [29][35][36]. The current trend indicates that Mobile IP [32] will be used to provide global roaming within heterogeneous mobile networks.

Other challenges are to develop an authentication procedure with low latency and to solve security and billing issues [43].

Developing a seamless handoff procedure is an important challenge [6][7][27][37]. Handoff decision criteria, handoff triggering time, and network selection criteria should be defined. Handoff decisions depend on RSS at a MH from both networks and on the velocity and the location of a MH. If a vertical handoff is triggered only by RSS, it may cause unnecessary frequent handoffs, which degrade performance of a user moving with a high velocity since each WLAN covers only a small area. For a user with high velocity, 3G networks may support the user more efficiently since they cover wider areas. User's location should be also considered to prevent unexpected ping-pong effect, which is happened near cell borders and means a MH keeps being handed off between two cells. Therefore, network selection criteria in heterogeneous networks must be defined efficiently. Another concern is the management of handoff triggering time. The latency of the vertical handoff is usually longer than that of the horizontal handoff since an authentication procedure may be necessary when a MH enters a different network. The handoff latency is longer in the case of a loosely coupled integration, which is a more feasible solution. When an upward handoff from a smaller cell to a larger cell occurs, a MH loses more data packets as the handoff latency increases while moving outside coverage of the previous cell. When a downward handoff occurs, a MH can still receive data packets since it is still in the larger cell.

An additional challenge for a heterogeneous network is to ensure QoS [28][38]. Each kind of network supports different service quality to a user. Therefore, a scheme must be developed with which the heterogeneous network can manage these QoS changes with minimum effects while a MH roams across the different access networks.

Another challenge of these networks is to define an efficient data transport protocol, such as TCP. In contrast with internet protocol (IP), TCP has received little attention from researchers. Traditional TCP does not work efficiently in wireless networks since it cannot distinguish congestion losses from wireless link losses. As a result, TCP unnecessarily decreases its data rate and lowers its throughput, even when losses are actually caused by wireless link errors rather than by network congestion. Many suggestions for addressing this problem have been proposed. So far these suggestions focus primarily on a heterogeneous network with wired and wireless links, not on a heterogeneous mobile network with different wireless access systems. To provide seamless services to a MH that has multiple wireless access interface, transport protocols should be designed.

CHAPTER 3

TRANSPORT PROTOCOLS

3.1 Transport Control Protocol

TCP is the predominant transport protocol in current Internet and implemented on the top of the Internet protocol (IP) layer. The main objectives of TCP can be specified as the followings:

- reliable data transfer by error recovery algorithms
- flow control to prevent network congestion

The basic algorithms of TCP are defined by [18][46][55]. Added to the basic TCP, several TCP extensions have been proposed by [51][13][52].

To guarantee reliable data transfer, TCP provides error recovery mechanism by retransmitting lost packets. When a TCP sender transmits a data packet, the receiver sends back an acknowledgement (ACK) packet to the sender to inform that the receiver acquires the packet correctly and in-orderly. In the ACK packet, the next expected sequence number is included. Data are not always delivered to the TCP receiver in-orderly because some data are lost by buffer overflow or physical link error. When the receiver acquires a packet correctly but out-of-orderly, it sends back a duplicate ACK (Dup_ACK), which contains the same sequence number as the last sent ACK, to the sender so that the sender retransmits the missing packets. No feedback is provided for packets received in error. ACKs may be generated for every segment, or for every other segment if the delayed-ACK mechanism [46] is used.

The packet retransmission can be triggered by one of the following events : reception of three Dup_ACKs, or expiration of a RTO timer. TCP sender keeps information of transmitted packet such as sequence number and the time of sending the packet. The sender waits for an ACK before sending the next set of packets. It maintains some timers to define a RTO period. The timers are continuously updated based on a weighted average of previous RTT. RTO value is important to improve TCP throughput, especially in an unreliable wireless network. Shorter RTO leads some spurious retransmissions while longer RTO slows down the error recovery. When the RTO timer expires, the sender considers that the transmitted packet is dropped, and retransmits it. The sender also retransmits the packet as soon as it receives three Dup_ACKs, not waits until the RTO expires. This is called fast retransmit [55]. TCP can reduce the recovery time from packet loss with the fast retransmit scheme.

TCP provides flow and congestion control algorithms to prevent network congestion and packet drops and to fully utilize the available bandwidth by adjusting its data rate according to the received acknowledgements. Congestion occurs when routers are overloaded by traffics being delivered with higher data rate than that the network can support. Congestion causes packet drops, long delay, and low throughput. The flow and congestion control algorithm is implemented through a sliding window [54][55]. TCP defines two windows: a congestion window and an advertised window. A congestion window is defined at a TCP sender and the size of the window, *cwnd*, is determined by slow start (SS) or congestion avoidance (CA) algorithm so that packets in transit do not incur network congestion. An advertised window is defined at a TCP receiver and indicates remaining buffer size of the receiver. The size of the advertised window is informed to the sender within every ACK packet. The minimum of the congestion window and the advertised window determines the amount of outstanding data that a TCP sender can transmit

TCP has two states: the SS state in which a sender exponentially increases its data rate and the CA state in which a sender linearly increases its data rate.

When a TCP session is opened, the TCP sender initiates SS state by transmitting one packet and waiting for an ACK. When the sender receives ACK packet, it increases its *cwnd* from one to two. During the SS state, *cwnd* is doubled in each time when a TCP packet is positively acknowledged. This state is called SS because the starting congestion window is one, however, it produces an aggressive exponential increase of the congestion window. The SS state stops when the *cwnd* reaches the slow start threshold (*ssthresh*) or a loss is detected. *Ssthresh* defines a threshold value that a TCP changes its state from SS to CA. When the *cwnd* reaches the *ssthresh*, the *ssthresh* value is set to the half of the current *cwnd* and the TCP operates in a CA state. When a loss is detected, TCP updates its parameters such as *cwnd* and *ssthresh*, and retransmits the lost packet. When a loss is detected by the RTO expiration, the *ssthresh* is set to half of the current *cwnd* value, *cwnd* is set to 1, and the TCP operates in a SS state. When a loss is detected by receiving three Dup_ACK packets, the *ssthresh* is set to half of the current *cwnd* value, and TCP resumes in a CA state with a being halved *cwnd* value. As the sender is receiving Dup_ACKs, the sender recognizes the receiver is receiving packets sent. Therefore, the sender assumes the network is mildly congested and decreases its *cwnd* in half. This procedure is called as fast recovery [55].

In a CA state, the *cwnd* is increased by one packet per RTT and this increment produces a linear increase.

Figure 3 shows the basic TCP operation based on the *cwnd*. The graph shows an exponential increment during the SS state and a linear increment during the CA state. Since TCP traffic regulates its data rates based on the SS and CA mechanism, a lot of traffic can share the available bandwidth efficiently and fairly. These flow and rate control algorithm is essential to stability and success in the current Internet.

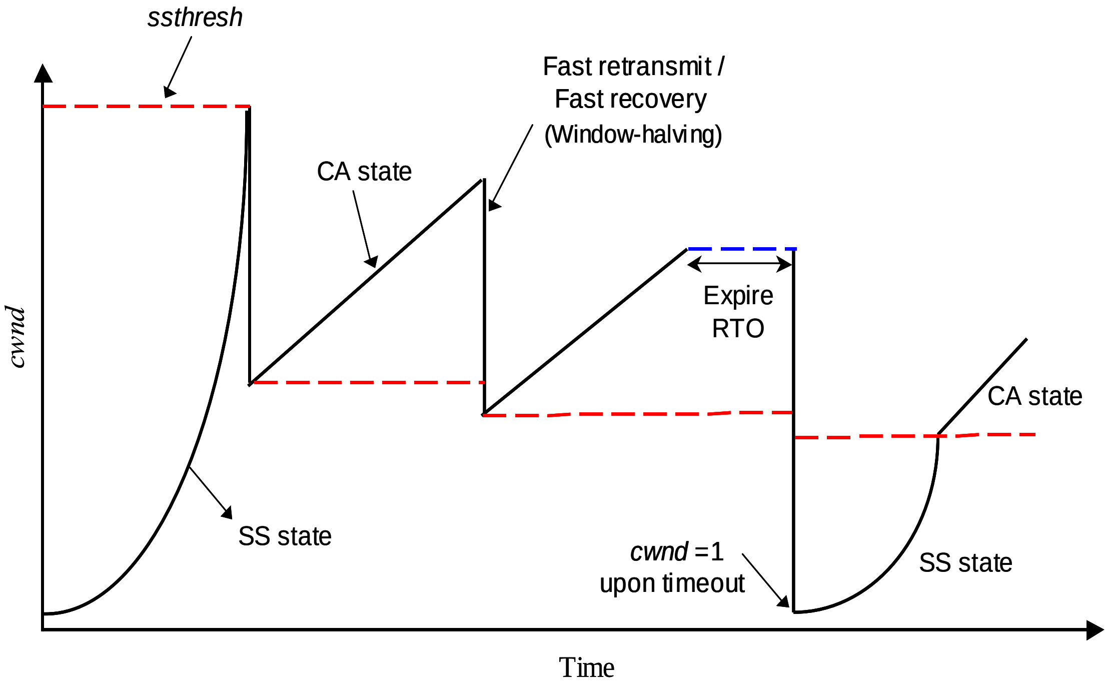


Figure 3. Basic TCP operation.

3.2 TCP in wireless networks

TCP was designed to work on wired fixed networks with negligible packet loss by the physical medium. In such networks, TCP assumes that the packet losses are caused by congestion and adjusts the TCP sender's data rate accordingly. In a wireless mobile network, however, losses are more likely the result of radio link characteristics such as a high bit-error rate, the fading effect, or a temporary disconnection caused by a handoff.

Traditional TCP does not work efficiently in wireless networks because it cannot distinguish and isolate congestion losses from wireless link losses. As a result, the congestion window of TCP decreases drastically, even when losses are actually caused by wireless link errors instead of network congestion. Many approaches have been proposed to address this problem. These proposed solutions can be categorized into three types: link layer solutions, split connection approaches, and end-to-end protocols as defined in [8].

The most popular example of the link layer solution is the Snoop protocol [8]. A Snoop agent resides on an intermediate host such as a base station (BS), caches packets from the sender, and retransmits the lost packet to the MH when packet losses occur, and suppresses Dup_ACKs sent from the MH to the sender. Packet loss is detected by the arrival of a small number of Dup_ACKs from the receiver or by a local timeout. The Snoop module should keep track of all the acknowledgments sent from the MH. By not propagating duplicate acknowledgments, the Snoop agent at a BS hides the packet loss over the wireless link from the correspondent node, thereby preventing unnecessary invocations of the TCP congestion control mechanism at the sender.

Indirect-TCP (I-TCP) and TCP for mobile cellular networks (M-TCP) are "split connection" approaches. I-TCP [9] splits the TCP connection as a wired part between a fixed host (FH) and a BS, and a wireless part between a BS and a MH. The BS maintains two TCP connections, buffers the data, and sends acknowledgement packets (ACKs) to the sender as soon as it receives data packets. The BS has a role to reliably deliver the data to the MH. The link between BS and MH may not be a TCP protocol. I-TCP scheme violates TCP’s end-to-end semantics. M-TCP [10] also splits up the connection between FH and MH in two parts: FH to BS and BS to MH. Whenever the BS detects a disconnection or packet loss, it sends an ACK back to the sender with zero advertisement window size to force the sender into a persistent mode and to prevent TCP from dropping its congestion window. The BS relays ACKs back to the sender only when the MH has ACKed data. M-TCP maintains end-to-end semantics even though it splits the connection into two parts. I-TCP and M-TCP present a solution for TCP to function properly within a network that has a wireless link as the last part of the connection. However, both have drawbacks. In both of these schemes, the BS caches the packets and retransmits the ones lost. Consequently, the BSs must have a large buffer and high processing capacity. Moreover, when the MH is handed off to a neighboring cell, the entire status of the connection and the buffered data needs to be handed off to the new BS, which imposes tremendous data overhead and causes packet loss during the handoff. Although these schemes are beneficial for fixed wireless users, they are unsuitable for mobile wireless users.

TCP-Westwood proposed in [12][40][41][42] is one of the end-to-end approaches. This proposed solution is not just for wireless networks and works well in cases where the system has a long RTT. Traditional TCP follows the additive increase, multiplicative decrease (AIMD) scheme. TCP-Westwood complies with the basic TCP-Reno behavior, which follows the AIMD scheme. However, after a loss is detected, the *cwnd*, which defines the maximum number of packets that can be sent out without overloading the networks, and *ssthresh*, which defines the steady-state network capacity, are set based on the measured bandwidth estimation, rather than using the conventional multiplicative decrease scheme. The TCP sender continuously estimates the available data rate of the connection by monitoring the ACK reception rate, and this value is used to compute *cwnd* and *ssthresh* settings after a congestion episode. Therefore, speedy recovery is ensured. However, this scheme is unsuitable for the vertical handoff environment because each link of the heterogeneous networks has different bandwidth and delay. Therefore, the sender cannot use the estimated value that was calculated in the former network.

WTCP [11] is one of the end-to-end approaches and proposed for wireless wide area networks (WWAN), which provide low bandwidth and high latency. WTCP was designed to solve the bandwidth asymmetry problem within WWAN. It conducts rate-based control rather than the traditional window-based control. It estimates the data rate based on the inter-packet separation. The receiver computes the desired sending rate based on a rate control scheme. The sending rate is determined by the ratio of the sender’s inter-packet delay and the receiver’s inter-packet delay, and notified to the sender by an ACK packet. The sender monitors the information and adjusts the sending rate accordingly. If the sender does not receive an ACK for a certain time, it goes into blackout mode and periodically sends probe packets to receive ACKs from the receiver. The receiver also maintains a history of packet losses when the network is predicted to be uncongested, and computes the expected average and deviation of the non-congestion-related packet losses over a time window. Based on the statistic data, it identifies the cause of the packet loss and informs the sender to properly adjust the transmission rate. The WTCP receiver has a very complex algorithm to calculate the desired sending rate. It needs a large processing capacity and buffer size, and incurs a large power consumption. Therefore, the WTCP is not suitable for a MH that has a limited power, memory and processing resources.

The selective acknowledgments (SACKs) algorithm [13], which can be classified as an end-to-end approach, shows better performance for wireless links that have random burst errors. Because standard TCP uses a cumulative ACK scheme, it often does not provide the sender with sufficient information to recover quickly from multiple packet losses within a single transmission window. In SACK, each ACK contains information about up to three non-contiguous blocks of data that have been successfully received. Each block of data is described by its starting and ending sequence numbers. This scheme works well within wireless networks. However, SACK is effective when multiple TCP packets are lost in a single TCP window and was not designed for a heterogeneous mobile network that has handoffs and abrupt changes in network condition. Also, it slightly increases the packet size and makes a slightly bigger load over wireless link.

Random early detection (RED) [48] is an active queue management mechanism in routers, which detects congestion before the queue overflows and provides an indication of this congestion to the end nodes. A RED router drops packets to inform incipient congestion to the TCP sender before real congestion occurs. RED router maintains minimum threshold, min_{th} and maximum threshold, max_{th} of buffer size. RED router probabilistically drops packets if the average queue size is between the $\min_{th}$ and the $\max_{th}$. Explicit congestion notification (ECN) [49] [50] is an extension of RED. When the average queue size is between the $\min_{th}$ and the $\max_{th,}$, a router that has a ECN function marks packets, rather than drops the packets. If the TCP receiver acquires the marked packet, it informs incipient congestion to the TCP sender by marking the subsequent ACK packet. As a consequence, the TCP sender triggers congestion avoidance algorithm and reduces its data rate.

Originally, these approaches were proposed for current wired Internet, rather than wireless networks. However, these schemes can improve TCP performance in wireless network if these are implemented in a node within radio access networks, such as a base station or an access point. The buffer in these nodes may experience frequent overflow because wireless link is usually the bottleneck. Therefore, the RED or the ECN can prevent frequent buffer overflow within a radio access network and congestion control in the TCP sender, and improve overall performance.

The schemes that have been proposed so far focus on heterogeneous networks composed of wired and wireless links. A new TCP scheme should be developed for the heterogeneous mobile network that integrates different mobile networks.

3.3 TCP in 3G network

TCP performance is considerably affected by link layer characteristics in 3G cellular networks [22]. The followings are various link characteristics of 3G cellular networks.

- Data rates

Initial 3G cellular systems are expected to provide bit rates around 384 kbps in downlink and 64Kbps in uplink. The data rate depends on the velocity of a MH and will be beyond 2Mbps at a pedestrian speed.

- Latency

The latency of traffic that goes through cellular network is large because of the extensive processing on the physical layer for forward error correction (FEC) and interleaving, and link-level retransmissions. To mitigate the effects of packet losses over unstable wireless link, 3G cellular networks provide extensive local retransmission mechanisms. These mechanisms decrease packet losses over wireless link, however these generate a large delay and a delay variation. A typical RTT within a cellular network varies from a few hundred milliseconds to one second. [22]

- Delay spikes

A delay spike is a sudden increase in the latency and 3G links may experience frequent delay spikes by several reasons. Retransmissions on the link layer and FEC on the physical layer cause delay spikes. Delay spikes are also occurred by forwarded packets from an old BS to a new BS caused by a handoff. Delay spikes may cause spurious retransmissions and throughput degradation.

- Bandwidth oscillation

A MH can switch its data rate according to the physical link conditions or various scheduling mechanisms to increase system throughput. This causes bandwidth oscillation and spurious retransmissions. In [23][53], the authors show that the bandwidth oscillation can be the most important factor in reducing throughput. To reduce spurious retransmission, Eifel algorithm was proposed in [20][21] and it recommends using time stamp option in TCP.

- Intersystem handoff

3G cellular systems need to implement the backward compatibility with 2.5G systems, which provide lower data rate. 3G cellular network will also need to work with different types of mobile networks based on the 802 standard technologies or other wireless technologies. These needs create a MH that has an intersystem handoff as well as an intrasystem handoff. The intersystem handoff can adversely affect ongoing TCP connections since the network characteristics are radically changed.

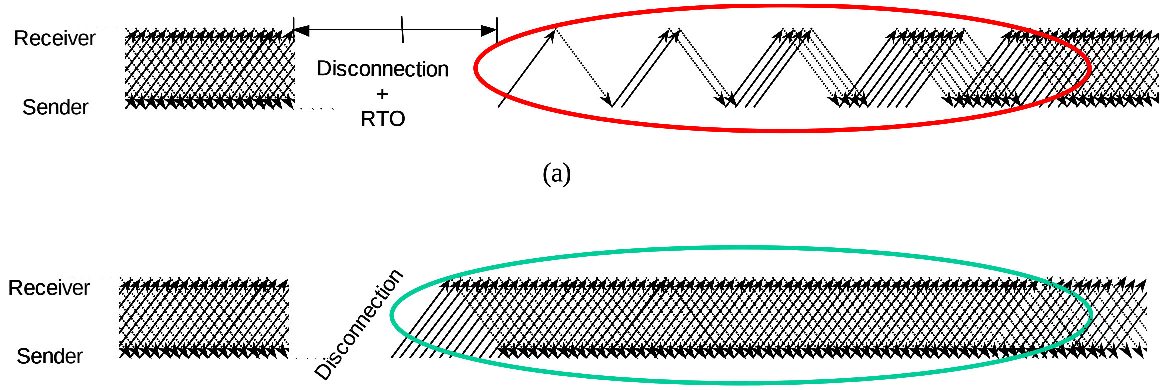
CHAPTER V

DESIGN OF A TCP FOR HETEROGENEOUS MOBILE NETWORKS

4.1 Freezing TCP during a handoff

The most important issue for a heterogeneous network consisting of wired and wireless links is the capability to distinguish packet losses over the wireless link from losses caused by network congestion. The capability to make this distinction will prevent the sender from diagnosing link errors as network congestion and decreasing its data rate. As a consequence, the network can better perform data transmission over wireless links.

In a heterogeneous mobile network integrating different types of wireless mobile networks, one task required to provide seamless service is to quickly adjust the sender's data rate as the sender moves into new network environments. Each mobile network within a heterogeneous wireless network has different characteristics in terms of capacity, bandwidth, delay, and coverage. Since each mobile network provides different bandwidth and delay to a user, a drastic change in data rate and delay may occur when a MH is handed off. When the TCP does not adjust its data rate quickly enough, low bandwidth utilization occurs immediately after a downward handoff, while continuous packet losses by buffer overflow are triggered immediately after an upward handoff. The latency of the vertical handoff is usually longer than that of the horizontal handoff, and this latency increases within a network integrated by the loosely coupled method. It is because a MH needs to do some authentication and authorization procedures for working in a new network and these procedures take longer time in the loosely coupled network since the integrated networks are independently deployed and belong to other service providers. Long handoff latency causes some packets to drop, requiring retransmissions and an exponentially increased RTO value. Longer handoff latency is critical to the upward handoff because a MH may leave the small coverage area and not receive a signal. Considering these facts, TCP has a better performance if the sender halts its data transmission during a handoff so that packet drops and timeouts are avoided. In [5], the authors proposed a Freeze-TCP scheme that avoids performance degradation caused by handoffs. In this approach, to prevent timeouts at the sender, a receiver sends a zero window advertisement (ZWA) to the sender when a handoff is impending and shrinks its *cwnd* to 1. The sender freezes all timeout timers and halts the data transmission when it receives the ZWA. Immediately after the handoff, the receiver sends three Dup_ACKs with information of the available receiving window size. Then, the sender resumes data transmission with its old *cwnd*. Figure 4 shows TCP packet flows when a handoff occurs. The regular TCP waits until the RTO expires, shrinks the *cwnd* to 1, and resumes data transmission. On the other hand, Freeze-TCP resumes data transmission immediately after the handoff, not waiting for the RTO, and avoids a drastic shrinkage of the *cwnd* caused by packet loss during handoff. RTO is exponentially increased by packet losses. Therefore, when the handoff latency is long, the RTO could be very large because of the consecutive packet losses. In this case, data cannot be transmitted by a TCP, even though the network can support it. Freeze-TCP improves throughput and system utilization during handoffs and works better as the disconnection time increases. Unfortunately for its potential as an overall solution, the Freeze-TCP scheme is designed for horizontal handoffs between homogeneous mobile networks, not vertical handoffs between heterogeneous mobile networks. When a MH undergoes a vertical handoff, network conditions change abruptly and drastically. Therefore, a MH should quickly adjust the parameters for data transmission to a new network. To support seamless data service, the TCP scheme must be modified to adjust its operation for different types of handoffs.



(a)

(b)

Figure 4. Performance improvement by Freeze-TCP (a) regular TCP, (b) Freeze TCP.

4.2 RTT and RTO variation during a handoff

RTT and RTO are important factors to improve TCP performance. RTT is the time that takes to send a packet from a sender to a receiver, get it processed at the receiver, and send a corresponding packet such as an ACK back to the sender. A flow’s RTT varies dynamically, depending on available bandwidth and queuing delays within a network. .

TCP uses a retransmission timer to trigger data transmission when the feedback messages from the receiver are dropped. The duration of the timer is referred to as RTO.

The calculation of the RTO value uses smoothed_RTT (SRTT), which averages the RTT, and RTT_variation (RTT_VAR), which is the mean deviation of the RTT [18][19].

These values are defined as the following:

$$\text{RTT\_VAR} = (1 - \beta) * \text{RTT\_VAR} + \beta * |\ \text{SRTT} - \text{RTT}\ | \tag{1}$$

$$\text{SRTT} = (1 - \alpha) * \text{SRTT} + \alpha * \text{RTT} \tag{2}$$

where α is defined as 1/8 and β is defined as 1/4 [19]. As seen by the equations, SRTT is a low-pass filter that memorizes a connection’s RTT history with a fixed weighing factor of 7/8, while RTT_VAR is a low-pass filter that keeps a memory of a connection’s RTT deviation history with a fixed weighing factor of 3/4.

After computing the SRTT and RTT_VAR, the sender updates its RTO value with the following equation:

$$\text{RTO} = \text{SRTT} + k * \text{RTT\_VAR} \tag{3}$$

where k is 4.

If the RTO is set with a small value, it causes spurious retransmissions [20]. If the RTO is set with a large value, it degrades TCP throughput over unreliable links. Using TCP over unreliable links incurs more frequent packet drops. The TCP sender has to wait until the large RTO expires to retransmit the lost packets. The RTO is exponentially increased if a TCP sender is unable to get feedback from the receiver during the RTO. This causes a long packet delay and low system utilization. Furthermore, a user might think the call is dropped because of a long idle time. This potential impact on the user increases the importance of setting proper RTO values.

In a heterogeneous network, which integrates a WLAN and a cellular network, the RTT in the cellular network is several times larger than that in the WLAN because cellular networks provide a retransmission function over the wireless links to compensate for unreliable links. As shown in the above equations, RTT and RTT_VAR values are smoothly updated by averaged history data. Therefore, these two values cannot be updated quickly after a vertical handoff occurs. The RTT variation caused by a vertical handoff differs from one caused by network congestion. A smoothing update is a reasonable way to compensate for a network fluctuation caused by congestion. However, this approach does not work well with vertical handoffs that typically involve sudden changes in delay and bandwidth.

Figure 5 shows the mathematical results of the RTO calculated by equations (1) to (3). Figure 5 shows the RTO values for a downward vertical handoff, in which the assumption is that a user moves from a larger network supporting lower bandwidth to a smaller network supporting higher bandwidth within an overlaid network. The RTT in the larger network is 300 msec and the RTT in the smaller network is 100 msec. The x axis indicates the number of RTT updates, and the y axis indicates the times. When a handoff occurs, the real RTT value decreases from 300 msec to 100 msec. However, SRTT decreases smoothly because it is updated with the stored history value as well as with the current value. The RTO increases rapidly and does not immediately converge into the proper value. This is mainly the result of the large RTT variation caused by a vertical handoff. In this case, the TCP sender has a larger RTO value until it converges, even though it has a much smaller RTT value after a downward vertical handoff. If a packet loss occurs in this period, a TCP sender has to wait until a long RTO period expires. This degrades performance and lowers system utilization.

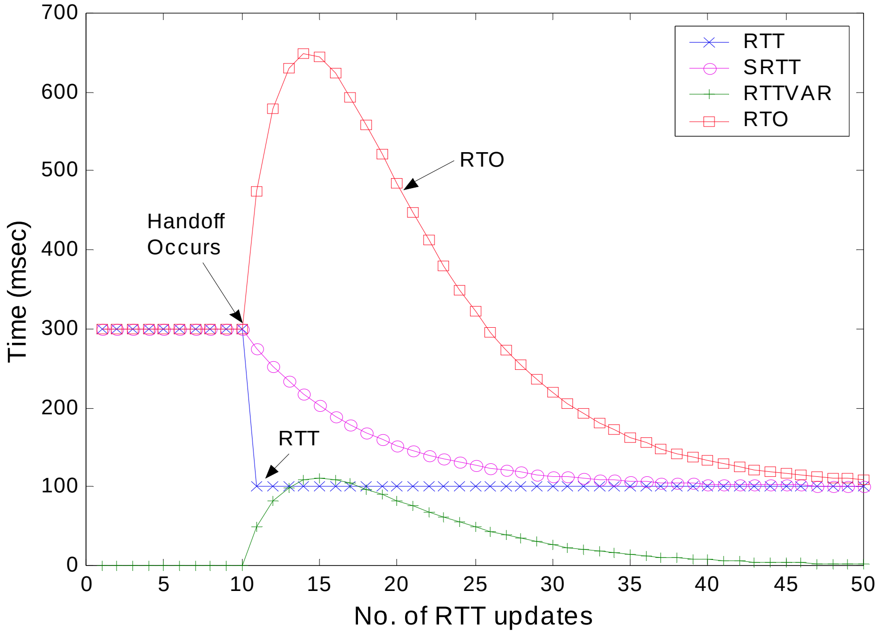


Figure 5. Calculated RTO variation for downward vertical handoff.

Figure 6 shows the RTO values in an upward vertical handoff. It indicates similar results as in the case of a downward vertical handoff. The RTO increases rapidly after a handoff, and this causes low system utilization and low throughput. Especially in the upward vertical handoff, a TCP sender may generate some spurious timeouts and spurious retransmissions. A spurious timeout occurs when the RTT suddenly increases to the extent that it exceeds the retransmission timer that had been determined a priori. Spurious timeouts can be caused by route changes, rapid increases in congestion at the bottleneck link, or by sudden decreases in bandwidth over the wireless link as the result of a vertical handoff.

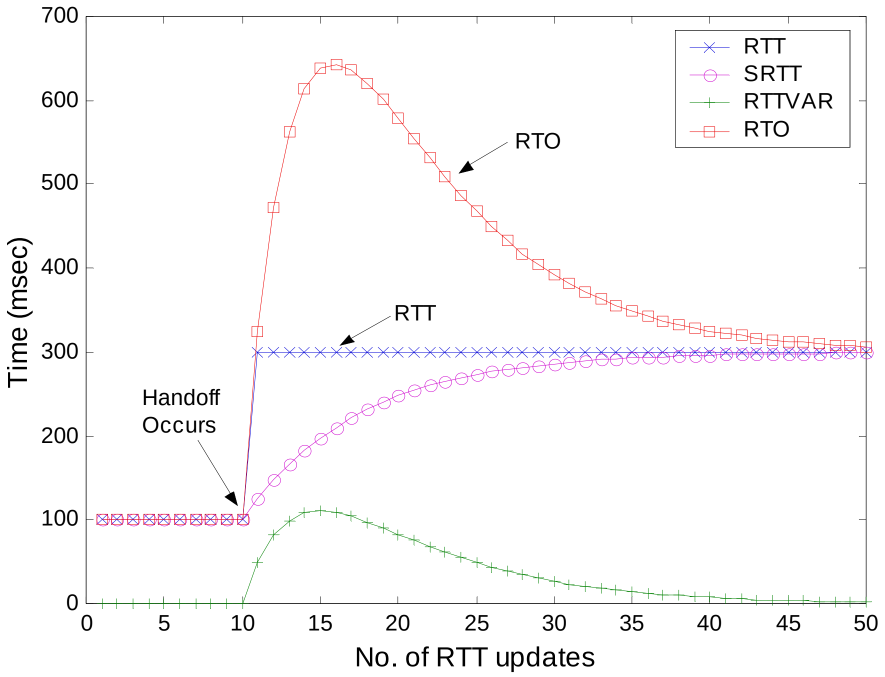


Figure 6. Calculated RTO variation for upward vertical handoff.

In this section, we present numerical results for a heterogeneous mobile network that integrates a WLAN and a 3G cellular network. We also considered some other networks that integrate different types of mobile networks : for example, WLAN and WMAN ; 3G cellular network and WMAN. Similar results are obtained for these networks. The numerical values are different because each network has different RTT and bandwidth. However, the trend of the RTO variation is similar and it generates similar problems regardless of the network types that are integrated.

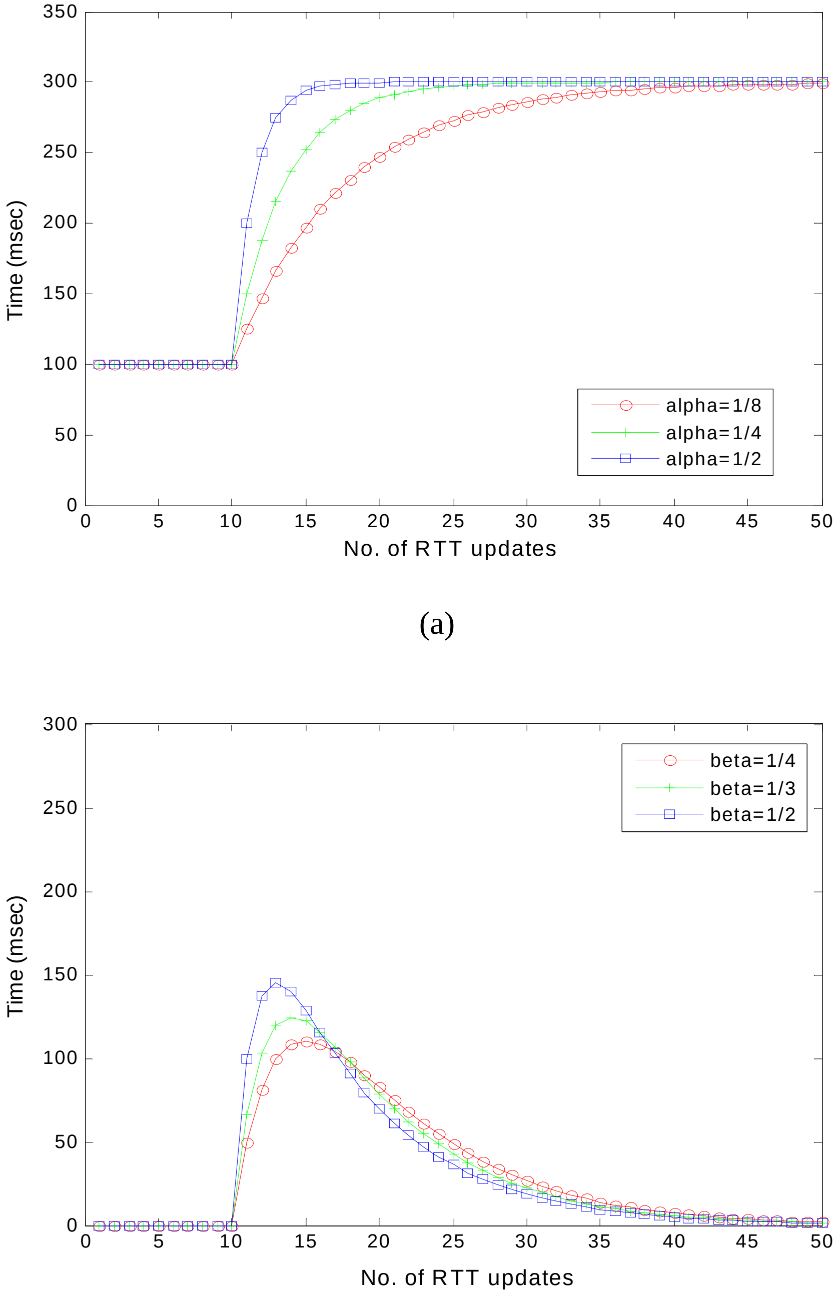
When a vertical handoff occurs, the RTO value should be promptly updated to prevent spurious retransmissions and quickly converged to a proper value to increase throughput. In a heterogeneous network, different types of handoffs occur in a MH. Therefore, TCP implemented within a MH needs to know the impending handoff situations and the type of a handoff so that the TCP operates differently after the completion of the handoff.

The impact of a vertical handoff can be diminished by using different values of α, β, and k. Figure 7 (a) shows the SRTT values calculated with different α value, and Figure 7 (b) shows the RTT_VAR calculated with different β value.

Figure 8 shows how the α, β, and k parameters affect to the RTO value. Figure 8 (a), (b), and (c) show the RTO values as α, β, and k vary, respectively. In our approach, the effect by the history data needs to be decreased. Therefore, we evaluate the results with larger α and β, and smaller k value.

α is used to calculate the SRTT. As increasing the α value, SRTT is calculated by a recent data. As shown in Figure 7 (a) and Figure 8(a), SRTT and RTO can converge quickly as increasing the α value and unnecessary RTO spike value decreases because the RTT variation decreases. β is used to calculate RTT variation. Since β is multiplied by the absolute value of the difference between the recently measured RTT and the average RTT, a large β augments the difference, increases the RTT variation, and decreases the convergence time a little bit. The results from Figure 7 and Figure 8 present the impact using different α is bigger than that using different β.

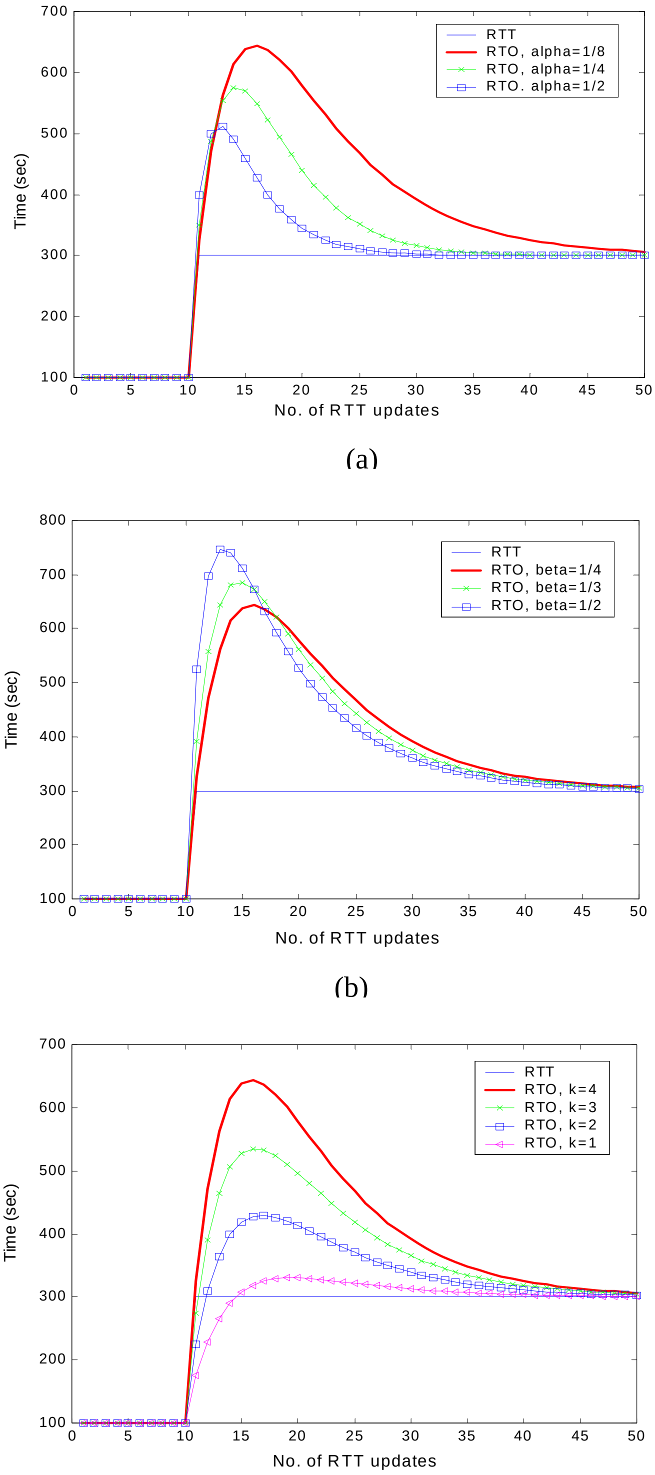
Because k amplifies the RTT variation when calculating the RTO, larger k can produce very large RTO, as shown in Figure 8 (c). This large RTO spike is not necessary in the vertical handoff case. However, larger k value helps to reduce the spurious retransmissions right after the vertical handoff in this case, as shown in Figure 8 (c).



(a)

(b)

Figure 7. RTT variation (a) with different α values, (b) with different β values.



(a)

(b)

(c)

Figure 8. RTO variation (a) with different α values (b) with different β values (c) with different k values.

Figure 9 shows the RTO variations by an upward vertical handoff when α, β, *and* k are defined with values other than α=1/8, β=1/4, and k=4 as recommended in [19]. When k equals 2, the drastic increases in RTO lessen because the impact of the RTT variation decreases. When α and β increase from 1/8 and 1/4 to 1/4 and 1/2, respectively, the RTO value converges more quickly into an optimum value as we can estimate by the previous results.

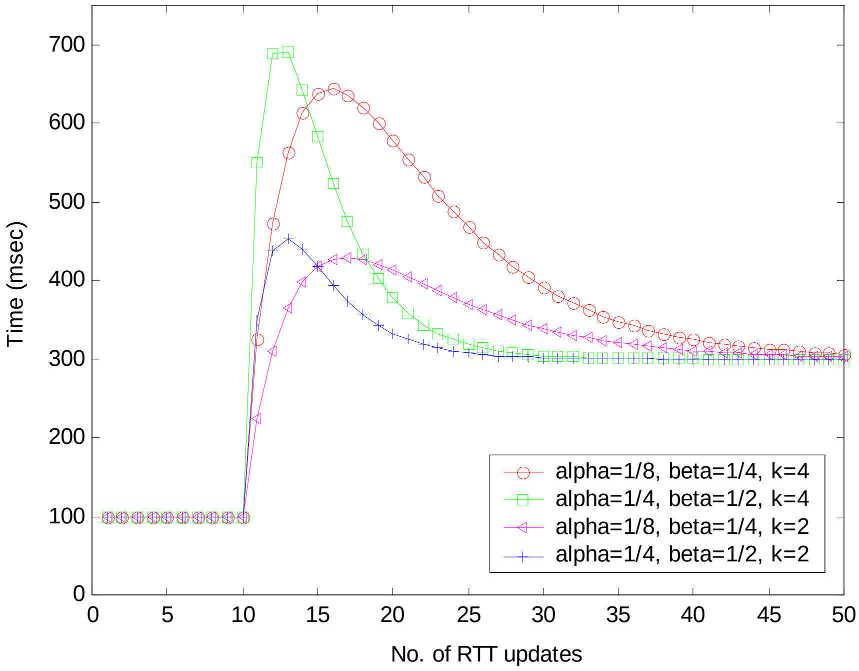


Figure 9. RTO variations with different α, β, k values : upward handoff case.

Figure 10 shows the RTO variations caused by a downward vertical handoff when α, β, *and* k are defined as the Figure 9. When k equals 2, the drastic increases in RTO lessen because the impact of the RTT variation decreases. When α and β increase from 1/8 and

1/4 to 1/4 and 1/2, respectively, the RTO value converges more quickly into an optimum value as we can estimate by the previous results.

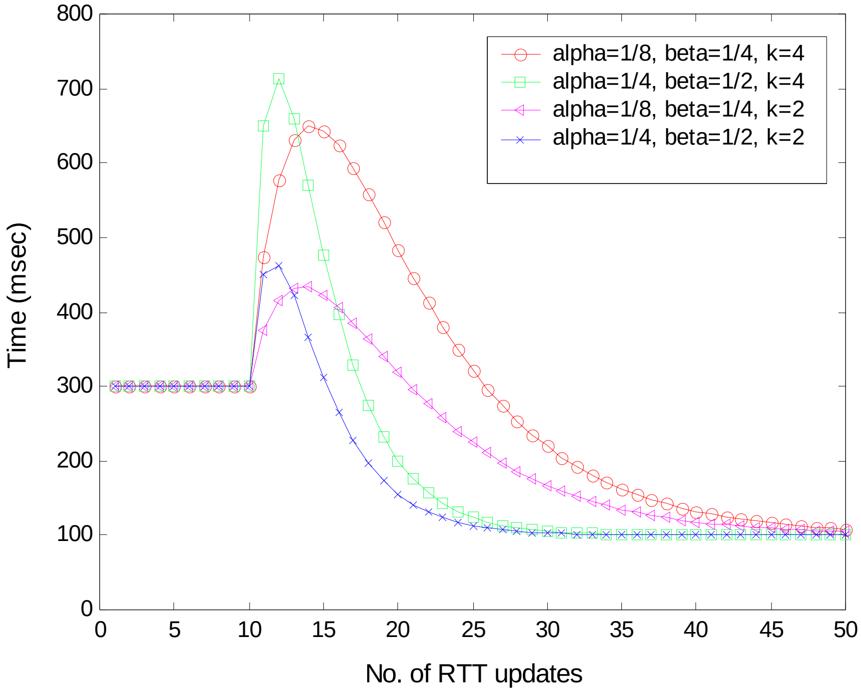


Figure 10. RTO variations with different α, β, k values : downward handoff case.

4.3 Throughput variation during a handoff

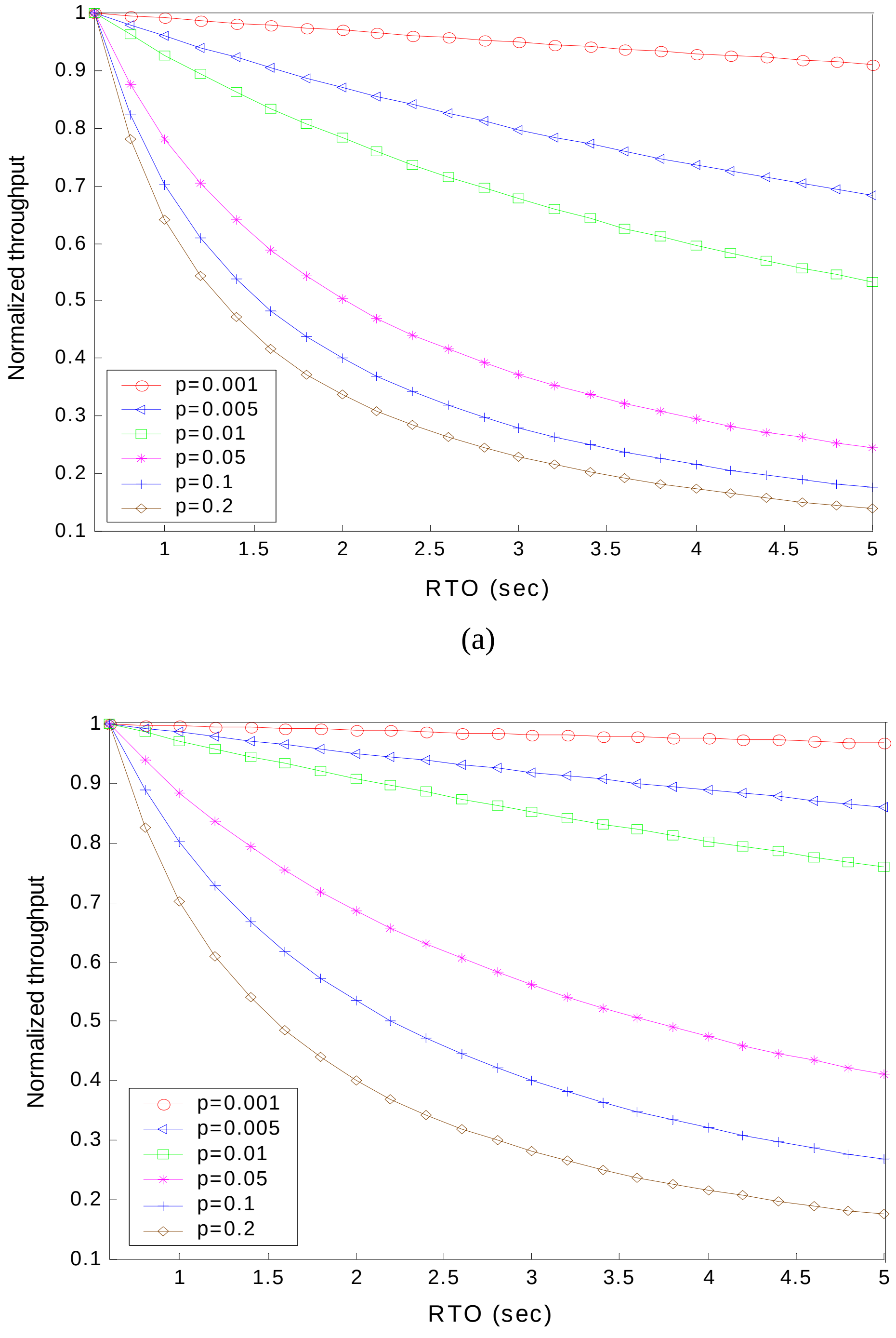
RTO is an important parameter to determine TCP throughput, and the consequences of the RTO become greater as the error rate increases. TCP throughput determined by RTT, RTO, and error rate, and it can be calculated by the following equation [44][45]:

$$B(RTT\ ,\ p, T_0)\ \approx\ \frac{S}{RTT\ \sqrt{\frac{2bp}{3}} + T_0\ \min\left(1, 3\sqrt{\frac{3bp}{8}}\right)\ p\ (1 + 32\ p^2)} \qquad (4)$$

where B is TCP throughput, p is error rate, T_0 is RTO, S is packet size, and b is the number of packets that are acknowledged by a received ACK. Many TCP receivers implement to send one cumulative ACK for two consecutive packets, therefore, b is usually 2.

As seen in equation (4), TCP throughput varies inversely with RTT, RTO, and error rate. TCP throughput decreases as the RTO increases, and the degree of decrease in throughput increases as the error rate grows because of the higher probability that the TCP sender will wait until the RTO expires

Figure 11 shows TCP throughput for various loss probabilities and RTOs. Figure 11 (a) shows the normalized throughput when RTT = 100 msec, while Figure 11 (b) shows that when RTT =300 msec. The top line in Figure 11 (a) and (b) shows the TCP throughput when the error rate equals 0.001. TCP throughput does not decrease much as the RTO increases because the error rate is low and TCP has little chance to wait until the RTO expires. However, as shown by the lowest line, a large RTO severely degrades TCP throughput when an error rate is high. The RTT in Figure 11 (a) is 100 msec, while the RTT in Figure 11 (b) is 300 msec. Therefore, the degree of decreased throughput in Figure 11 (a) is larger than that in Figure 11 (b) as RTO increases.



(a)

(b)

Figure 11. TCP throughput for various packet loss probability and RTO, (a) RTT=100ms, (b) RTT=300ms.

When a MH is moving within a heterogeneously integrated network, a vertical handoff increases the probability of packet drops. Therefore, it experiences severe throughput degradation for a certain time and cannot guarantee seamless data transmission to the user.

Figure 12 and Figure 13 show normalized throughput variation by the RTO spike in an upward and a downward vertical handoff case, respectively. During a vertical handoff, TCP has unnecessarily large RTO spike value and the figures show how the RTO spike degrades the TCP throughput. As shown in Figure 12 and Figure 13, the TCP throughput degradation by a downward handoff is larger than that by an upward handoff. The RTO spike values generated by an upward and a downward handoff are similar since the RTT_VARs are same in the case of these two handoffs. The RTT decreases after a downward handoff, therefore, the large RTO causes more severe degradation in a downward handoff case.

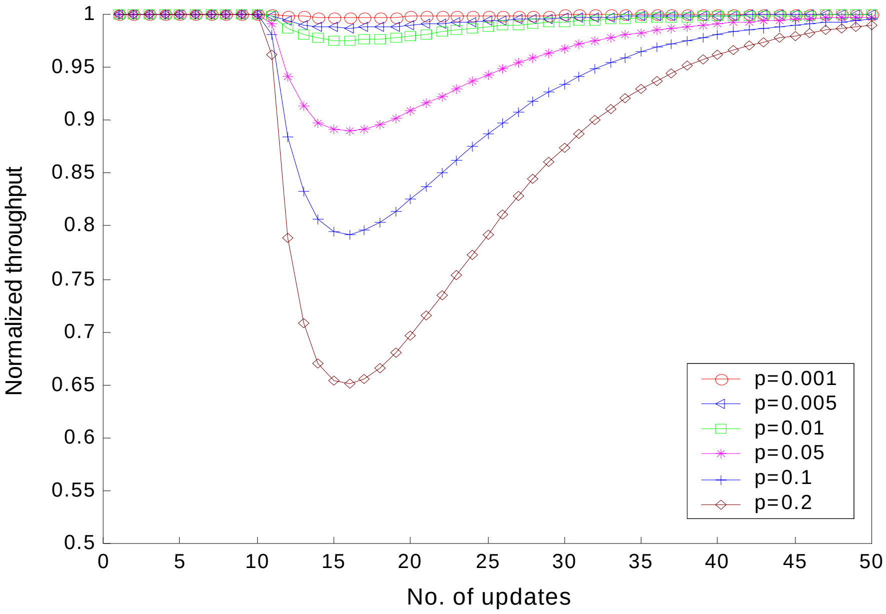


Figure 12. TCP throughput variation by RTO spike : upward handoff case.

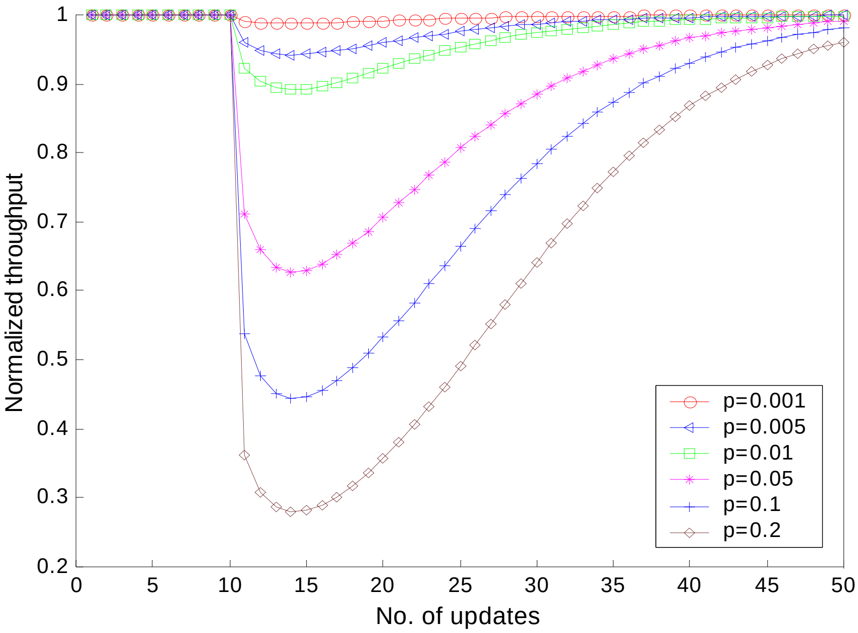


Figure 13. TCP throughput variation by RTO spike : downward handoff case.

4.4 Multimode MH for heterogeneous network

To support interworking within different wireless networks, a MH has a multi-mode wireless physical interface and switches its physical interface in response to a network environment. A MH has to implement a cross-layer interaction function through a radio resource control (RRC) module and report the measurement results of the RSS and the velocity to the RRC. The RRC module selects a physical interface based on a predefined scheme using the measurement data from the physical layer. It also must communicate with a network entity, such as a radio network controller (RNC), within the serving network to execute the handoff. To support the proposed TCP scheme, the RRC must notify the TCP layer of the impending handoff so that the TCP will respond properly.

Figure 14 shows the message flows for a handoff in a MH. When a handoff occurs, the control messages are generated as follows:

1. Both of the physical interface modules gather measurement data such as the RSS and the velocity of the MH and report it to the RRC module;
2. The RRC module selects a physical interface based on a predefined scheme using measurement data from the physical layer;
3. The RRC module notifies the TCP layer of an impending handoff so that the TCP receiver module will set its handoff optional field properly;
4. The RRC module triggers a horizontal or vertical handoff.

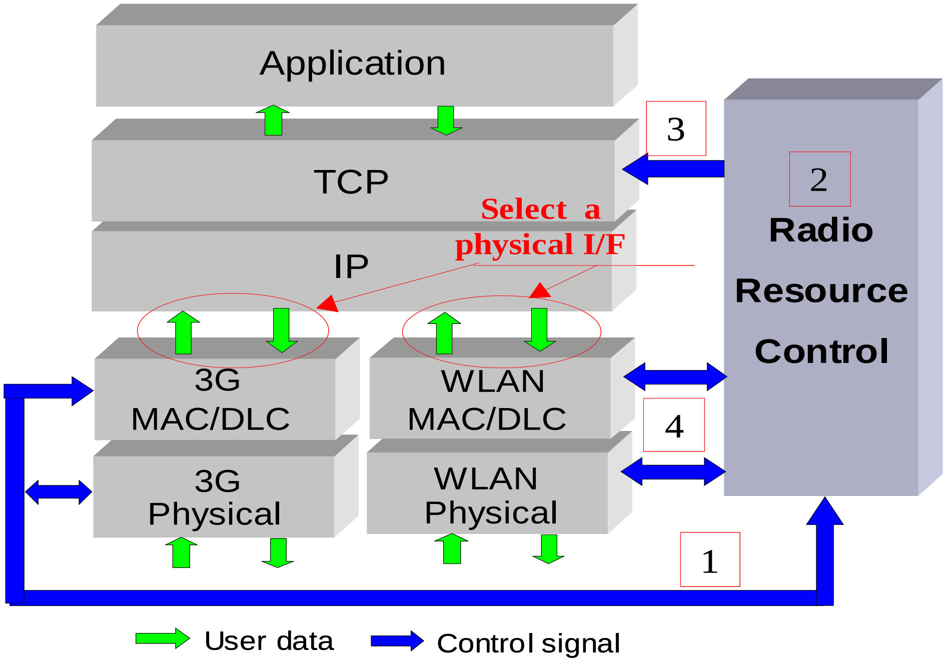


Figure 14. Control message flows for a handoff in a MH.

4.5 Design of a TCP for a heterogeneous mobile network

4.5.1 TCP receiver

After moving into a new network, TCP should adjust its operation to the type of handoff involved. For a horizontal handoff, the TCP sender needs as quickly as possible to resume its data transfer at the same rate as before. For a vertical handoff, TCP needs to adjust all parameters, such as RTT, RTO, *cwnd*, and *ssthresh,* to provide seamless data transmission to the user without overloading the new network.

As mentioned in Chapter 3, TCP has two states: SS state in which a sender exponentially increases its data rate, and a CA state in which a sender linearly increases its data rate. After a handoff, TCP needs to begin data transmission in a different state based on the type of handoff.

In the proposed scheme, TCP halts its data transmission during a handoff and resumes data transmission in either the SS or the CA state. If a MH undergoes a horizontal handoff, the TCP resumes in the CA state. If a MH experiences a vertical handoff, the TCP restarts in the SS state. Since the network environment is changed abruptly in a new wireless network after a vertical handoff, starting in the SS state and estimating the available bandwidth, rather than reverting to the same bandwidth as before the vertical handoff, yields improved performance and seamless transmission.

As shown in Figure 14, the physical layer in a MH reports RSS and its velocity to the RRC module. The RRC module determines the handoff triggering time and notifies the TCP layer of the impending handoff. When a MH is working as a TCP sender, the TCP sender halts its data transmission until the RRC module notifies a completion of the handoff. After the handoff, it resumes data transmission based on the type of the handoff. When a MH is working as a TCP receiver, the TCP receiver should inform the impending handoff and the completion of the handoff to the TCP sender.

In the proposed scheme, a MH that is working as a TCP receiver uses an optional field in the TCP header to identify the handoff situation. The optional field is defined as follows:

Handoff (HO) optional field = 00 : no HO

HO optional field = 10 : horizontal HO

HO optional field = 11 : vertical HO

When a handoff is impending, the TCP receiver sends an ACK that sets the HO optional field to the proper value for the handoff type; then, as soon as the handoff is completed, the TCP receiver sends an ACK, which has 00 value in the HO optional field, so that the TCP sender can resume data transmission without waiting for a timeout. This ACK message, which informs the sender of an impending handoff, can be delivered multiple times, thus improving reliability. When a MH is working as a TCP sender, no information needs to be transferred. Figure 15 shows this procedure at a TCP receiver.

The proposed scheme uses an option filed in a TCP header. Therefore, it can resolve the issues of compatibility with existing TCPs.

```
/* at Receiver  */
/* when a HO occurs */
if  Horizontal HO
    ACK with HO option field = 10;
else      /* Vertical HO */
    ACK with HO option field = 11;
end
/* when HO is completed */
/* notify the sender HO is completed and ask to restart  data transmission
*/
Ack with HO option field = 00;
```

Figure 15. Procedures at a TCP receiver.

4.5.2 TCP sender

The TCP sender monitors the HO option field and adjusts its *cwnd*. If it detects that a horizontal handoff is occurring, the TCP sender stops the timeout timer and suspends data transmission until the handoff is completed. When the handoff is completed, the TCP sender resumes data transmission in the CA state with the same *cwnd* as before the handoff because the MH moves into the same wireless environment. If a vertical handoff is occurring, the TCP sender stops the timeout timer and holds data transmission until the handoff is completed. When the handoff is completed, the TCP sender resumes data transmission in the SS state with the *cwnd* = 1, which estimates the available bandwidth. Figure 16 shows this procedure for the TCP sender. After receiving three Dup_ACKs, or timeout, the procedure follows the traditional TCP algorithm.

```
/* at Sender */
 /* when HO is occurred */
 if  HO option field = 10                    /* Horizontal HO */
        Stop timeout timer;
        hold data transmission;
        if  HO option field = 00             /* HO is completed */
               restart timeout timer;
               cwnd = old_cwnd;              /* Keep cwnd value */
               ssthresh = old_ssthresh;
               resume data transmission in CA state;
        end;
 else if  HO option field = 11                 /* Vertical HO */
        Stop timeout timer;
```

```
    hold data transmission;
    if  HO option field = 00          /* HO is completed */
        restart timeout timer;
        cwnd = 1;
        ssthresh = 65535bytes;
        resume data transmission in SS state;
    end;
else if  HO option field = 00                   /* No HO */
     if  TCP is at SS state
            cwnd = cwnd + 1;          /* Same as TCP Reno*/
        continue data transmission staying in SS state;
    else if  TCP is in CA state
        cwnd = cwnd + 1/cwnd;    /* Same as TCP Reno*/
        continue data transmission staying in CA state;
end
```

Figure 16. Procedures at a TCP sender.

4.6 Simulation results

We evaluated the performance of the proposed TCP scheme using *ns-2* [55]. Simulation scenarios are defined for a heterogeneous network, which includes a WLAN and a 3G cellular network. We assumed the network is integrated by the loosely coupled method that does not have a direct connection for control messages or for forwarding data between the WLAN and the cellular network. As seen in Table 1, the data rate is 144 Kbps and end-to-end RTT is 300 msec for a 3G cellular network, while the data rate is 2 Mbps and end-to-end RTT is 100 msec for the WLAN network. A vertical handoff occurs at 70 sec and ends at 73 sec.

Table 1. Simulation parameters.

	3G network	WLAN
Data rate	144 kbps	2 Mbps
RTT	300 msec	100 msec

Figures 17 and 18 show the sequence numbers for transmitting data versus elapsed time for a downward vertical handoff and for an upward vertical handoff, respectively. After a downward vertical handoff from a 3G cellular network to a WLAN, the available bandwidth increases drastically. However, the sender does not use the full bandwidth available even though the WLAN can provide much higher bandwidth. The reason is that there is no method by which a TCP sender can detect the increase in bandwidth. On the other hand, the performance of the proposed scheme indicated by the red line in Figure 17 shows the TCP sender can re-estimate the available bandwidth immediately after completion of the handoff, rapidly increase its data rate, and stabilize its transmission rate. In contrast, the normal TCP indicated by the blue line experiences packet losses during the handoff and doubles its RTO. Therefore, it can resume data transmission after about 77 sec. Figure 18 shows the TCP performance of an upward vertical handoff in which the available bandwidth is decreased. Normal TCP has to wait until the RTO expires to resume its data transmission. Freeze TCP tries to transfer its data immediately after a handoff, however, it drops some packets by using large *cwnd* and needs some time to stabilize its data transmission.

Figures 19 and 20 show the congestion window value versus elapsed time for a downward vertical handoff and for an upward vertical handoff, respectively.

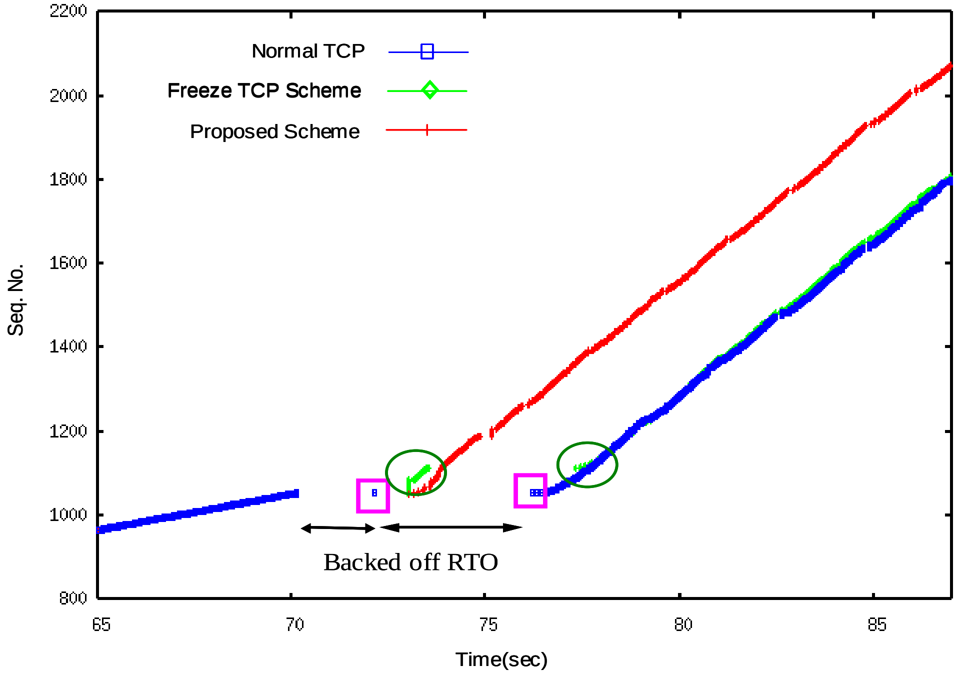


Figure 17. Sequence number vs. time for a downward vertical handoff.

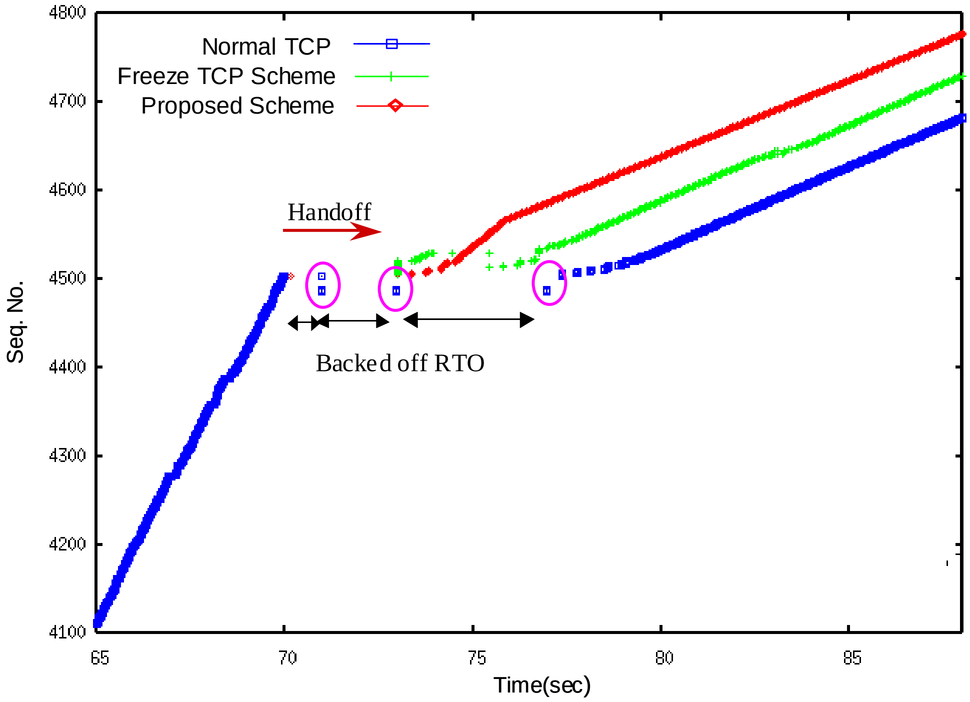


Figure 18. Sequence number vs. time for an upward vertical handoff.

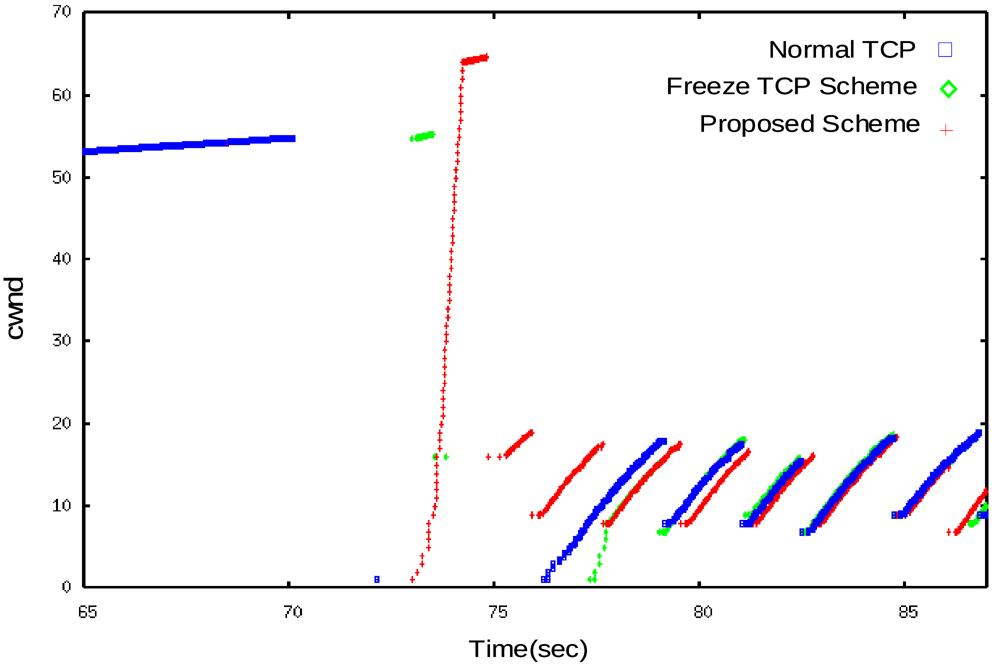


Figure 19. *cwnd*. vs. time for a downward vertical handoff.

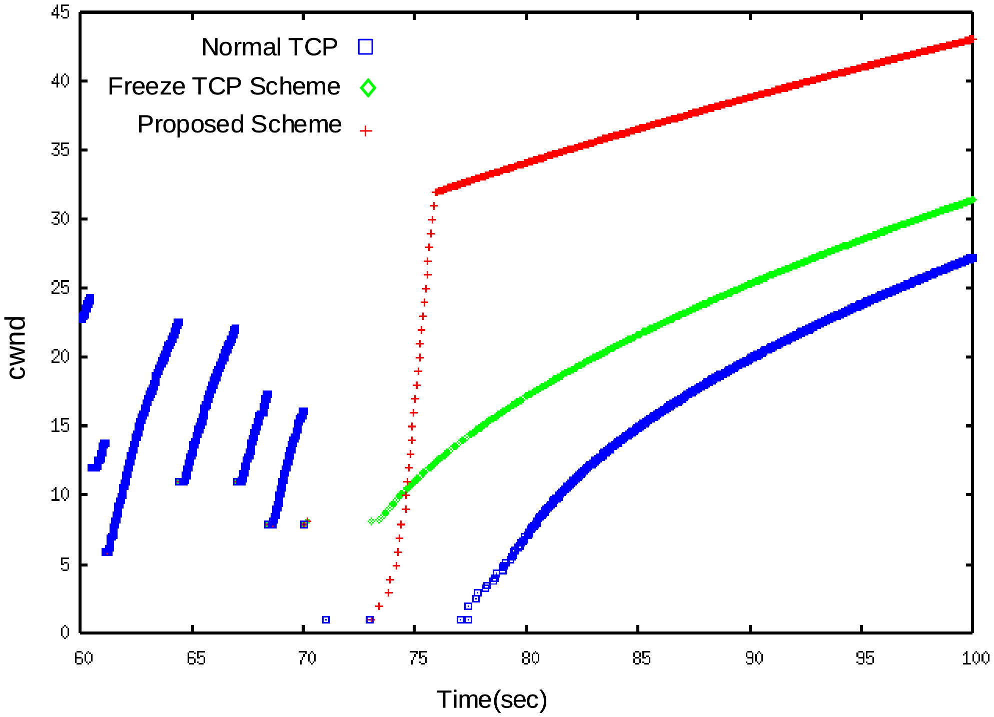


Figure 20. *cwnd*. vs. time for an upward vertical handoff.

The TCP scheme has AIMD characteristics in which it slowly increases its data rate and quickly decreases its data rate. Therefore, the performance of the proposed scheme improves most when the available bandwidth is abruptly increased by a downward handoff. In our simulation model, a WLAN has a short RTT and high bandwidth. Therefore, the TCP can increase its data rate relatively fast. However, a hierarchical network combined with two networks that have similar RTT performs much better with the proposed scheme.

Figures 21 and 22 show the RTO versus elapsed time. As shown in these figures, the RTO increases rapidly when a vertical handoff occurs, mainly caused by the large RTT variation. In this case, the RTT variation is not caused by network congestion, but by the move to a new network environment that has a different RTT value. Therefore, TCP has to update its RTO value quickly. The performance of TCP can improve if the SRTT value is updated frequently, not just once per RTT. As shown in Figure 21 and Figure 22, the sender has a large RTO for a certain period. Therefore, the TCP performance is low if burst packet errors occur during this period because the TCP sender must wait until the RTO expires. Users may misinterpret this waiting time as a call drop. For an upward handoff, the RTT increases immediately after the handoff. However the TCP does not update its RTO value quickly enough. As a result, TCP reacts as if some packets transferring through a 3G cellular network have been dropped. This causes some spurious retransmissions. After a downward handoff, TCP might perform poorly when packet loss occurs because the TCP updates its RTO value slowly and must wait until the expiration of the RTO, which is a large value calculated within a 3G network.

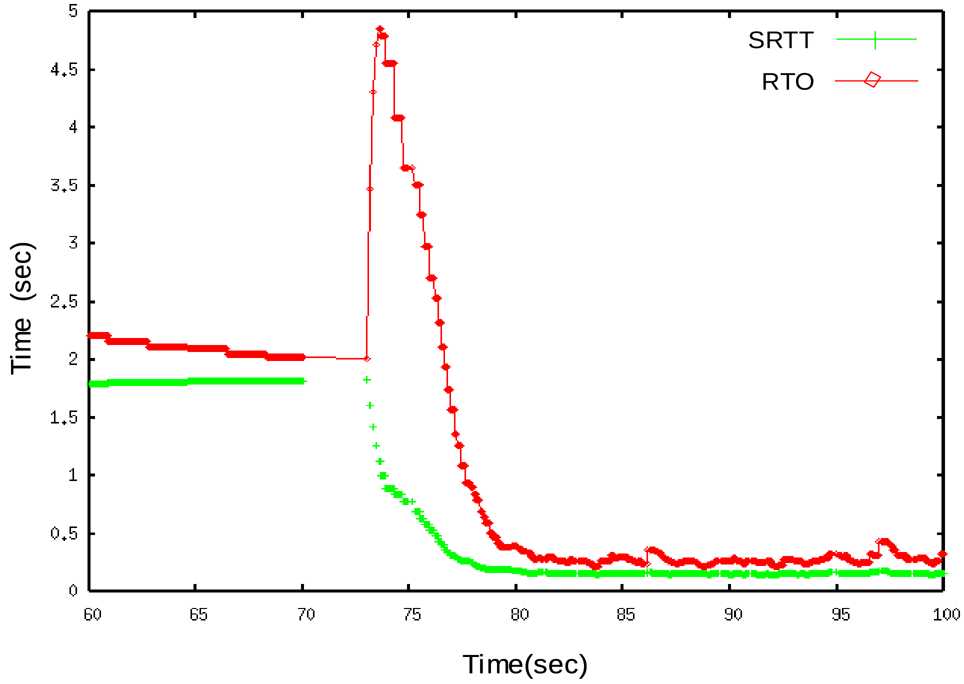


Figure 21. RTO & SRTT vs. time for a downward vertical handoff.

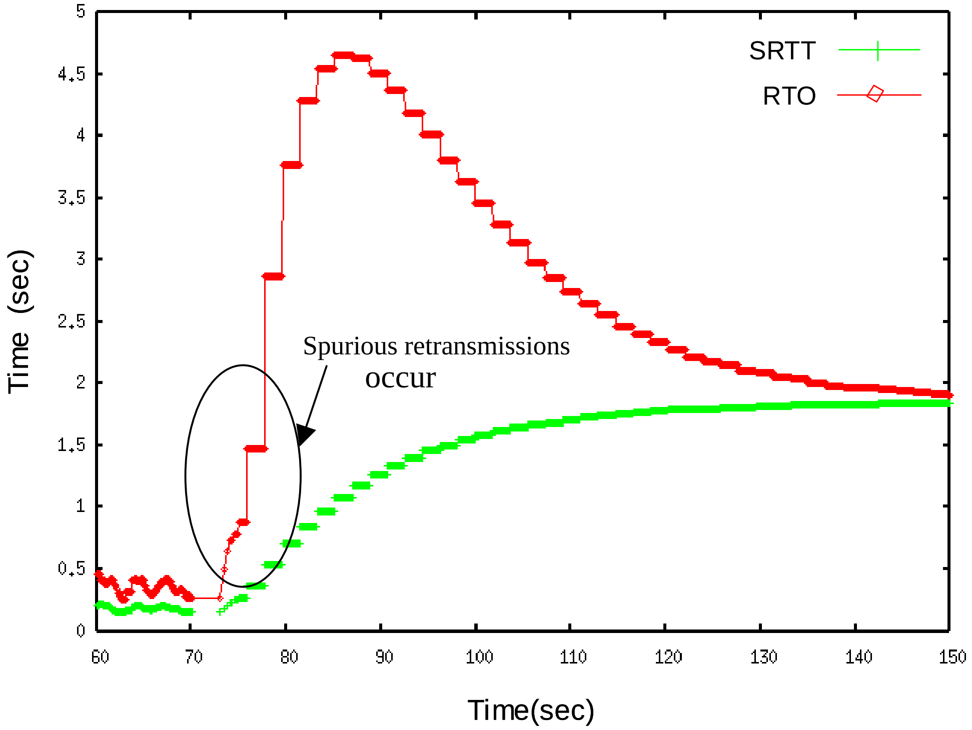


Figure 22. RTO & SRTT vs. time for an upward vertical handoff.

To solve this problem, we used a method that immediately after the handoff resets all the related parameters such as RTT, SRTT, RTT_VAR, and RTO. Figure 23 and Figure 24 show the results of the method. Figure 23 shows that the RTO spike problem can be solved for downward handoffs. However, in the case of upward handoffs, shown in Figure 24, the RTO still has some large values for a while, even though all the parameters have been reset. In this case, the new network has a large RTT and small bandwidth. Therefore, the real RTT varies for a while until the TCP sender stabilizes its *cwnd* value, which causes the large RTO values.

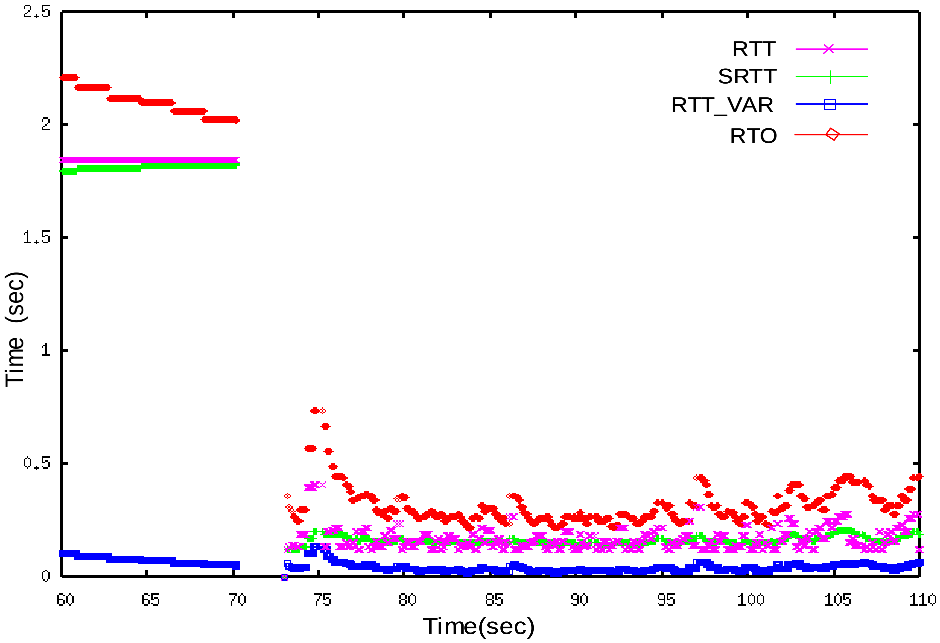


Figure 23. RTO vs. time after reset for a downward vertical handoff.

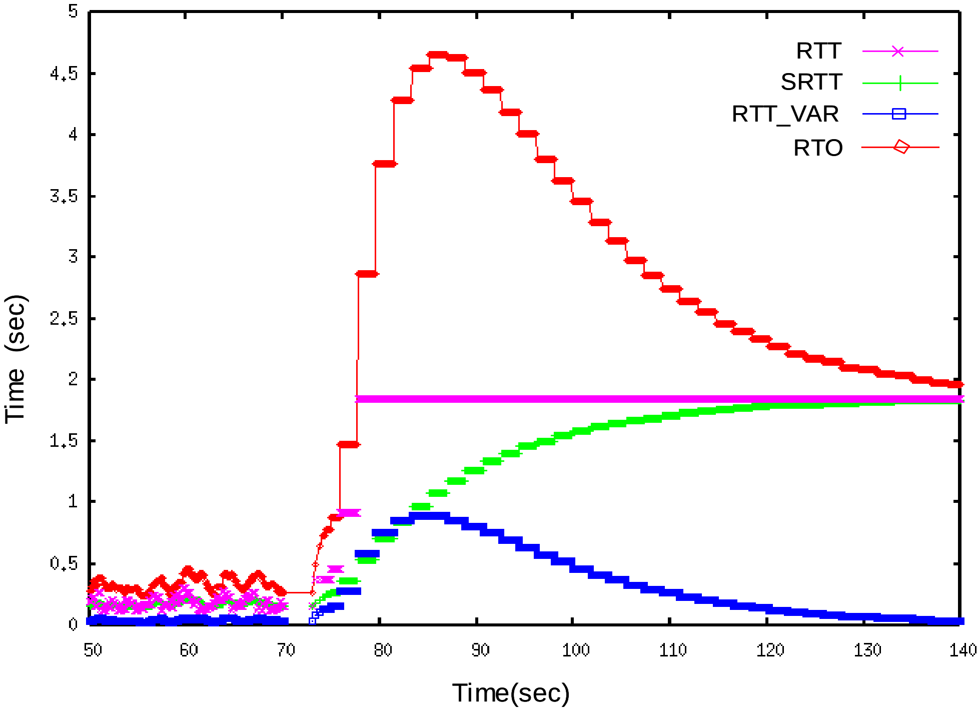


Figure 24. RTO vs. time after reset for an upward vertical handoff.

Figure 25 and Figure 26 show the throughput performance in a downward handoff and an upward handoff, respectively. As shown in Figure 25, the proposed TCP scheme has the best throughput and it has the shortest time to restart its data transfer in a stable state. It is improtant to provide seamless data transfer to a user.

As shown in Figure 26, the traditional TCP has the worst throughput. Since it cannot transmit data during five seconds even after the handoff completion, the user might feel the connection is dropped. The proposed shceme can redeem the data transmission immediately after a handoff, therefore it can provide seamless data transfer and decrease the service time.

Currently, TCP is mainly used for delay tolerant services, however, the service time is important to both of users and network providers. Also, in a heterogeneous mobile network, a MH has a high chance of handoffs caused by a high mobility. When we consider these conditions, the proposed scheme can enhance the TCP performance remarkably.

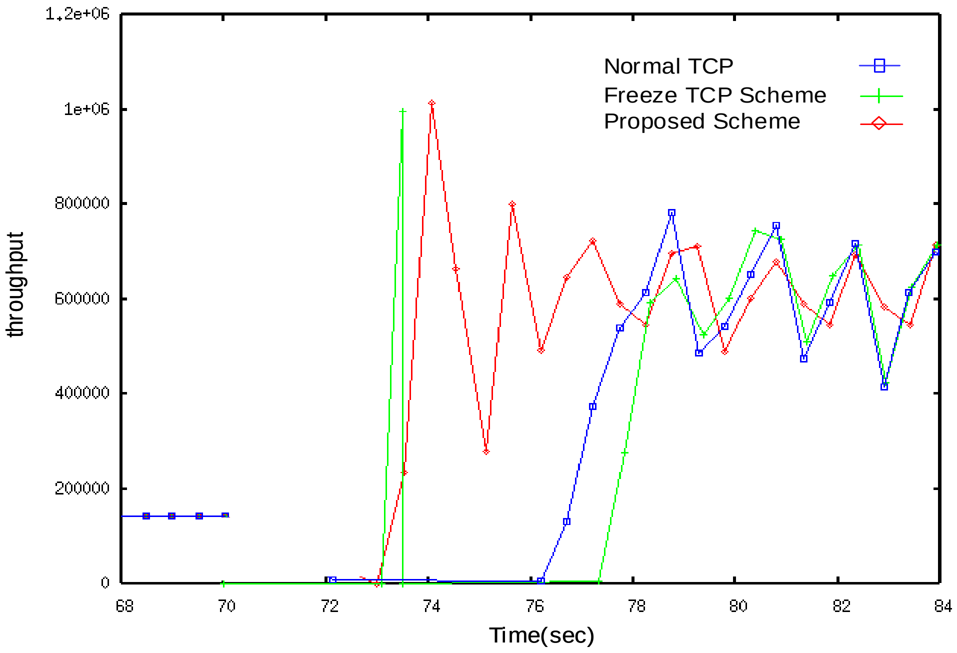


Figure 25. Throughput for a downward handoff.

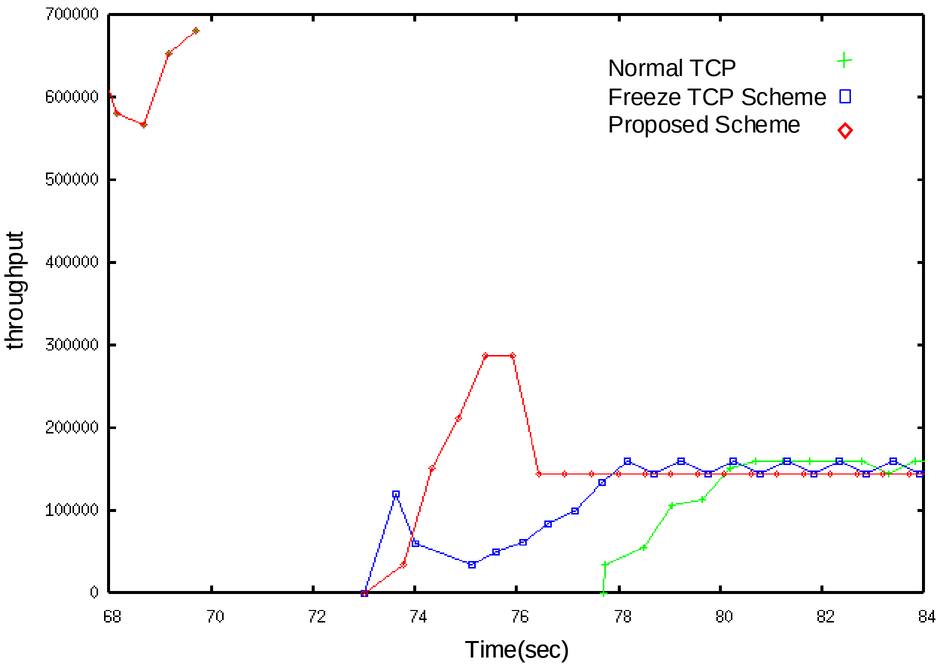


Figure 26. Throughput for an upward handoff.

To evaluate the performance improvement of the proposed scheme, file transferring time for each TCP scheme is simulated. Table 2 shows the results when a TCP sender transfers 500 KB file and 1MB file. As shown in Table 2, the file transfer time of the proposed TCP is shorter than the other TCP schemes. Table 3 shows the percentage of the performance improvements in file transfer. When a TCP sender transfers 500 KB file, the proposed scheme improves the performance in file transfer time as 28.2% and 21.2 % comparing with the normal TCP and Freeze TCP, respectively. The percentage of the improvements increases as the file size decreases and the number of handoffs increases during the file transfer.

Table 2. File transfer time.

	File size	
	500 KB	1 MB
Normal TCP	20.2 sec	31.8 sec
Freeze TCP	18.4 sec	29.7 sec
Proposed TCP	14.5 sec	26.3 sec

Table 3. Performance improvements in file transfer.

	500 KB	1 MB
Normal TCP → Proposed TCP	28.2 %	17.3 %
Freeze TCP → Proposed TCP	21.2 %	11.4 %

4.7 Summary

In this chapter, a TCP algorithm for enhancing the performance during a handoff within a heterogeneous mobile netowrk is proposed. In a heterogeneous network, different types of networks are integrated and a MH experiences drastic changes in network condition during a session. Traditional TCP cannot work efficiently in this environment since TCP does not know the cause of packet losses, regards it as a network congestion, and does flow control. Moreover, the TCP struggles with some troules by abrupt changes in network condition during an intersystem handoff.

In this chapter, an integrated network between WLANs and 3G cellular network is evaluated. Simulation results show the TCP has too large RTO caused by a RTT variation. The handoff latency could be quite large because of an authentication procedure between two networks and the packets that are on the flight to the old network might not be forwarded to a new network. Therefore, the probability of the packet loss increases severely. Moreover, the large RTO seriously decreases the network performance.

To prevent this situation, TCP sender needs to know the impending handoff situation for a seamless data transfer during and after a vertical handoff. We propose that the TCP uses two bits out of the TCP header's option field to recognize an impending handoff, type of the handoff, and a completion of the handoff. During the handoff, we propose the TCP halts its data transmission, since it prevents packet drops and a backed-off RTO value during a handoff.

After the handoff, the proposed TCP adjusts its data rate according to type of the handoff as soon as the handoff completes. If it was a vertical handoff, the TCP re-estimate the network capacity since the new network has different characteristics, in contrast with a horizontal handoff in which keeping the same data rate produces better performance. Also TCP has to update its RTO quickly enough after experiencing an abrupt change in bandwidth and delay.

From simulation results, we demonstrate the proposed scheme avoids packet loss during handoff, has better performance and reaches a stable condition rapidly immediately after the handoff because the TCP sender knows handoff events and readjusts its window size based on the capacity of a new network.

CHAPTER V

QOS MANAGEMENTS IN WCDMA NETWORK

5.1 QoS attributes

Future wireless networks will provide integrated multimedia services with diverse Quality of Service (QoS). For a wireless mobile network that has scarce resources, an adaptive and dynamic resource allocation according to the network condition is essential to guarantee QoS requirements of the diverse services. Especially within a heterogeneous mobile network, a MH should be roaming across different networks holding a session seamlessly. Consequently, adaptive resource management is crucial for the heterogeneous mobile network.

In a heterogeneous mobile network between WLAN and 3G cellular networks, the 3G cellular network may be struggling from handed-off calls from a WLAN since WLANs usually provide higher bandwidth to a user than cellular networks. The adaptive resource management should be efficiently defined in a cellular network, which has fewer resources, especially in this case.

As a 3G cellular network, we consider a WCDMA system using direct sequence-code division multiple access (DS-CDMA) technology. A WCDMA system specified by the 3GPP [66] provides well-defined QoS functions. QoS in the WCDMA system can be controlled by variable spreading gain, data rate, and transmit power.

The major QoS attributes are bit rate, throughput, delay, delay jitter, error ratio, call blocking rate, and call dropping rate. Real-time services have stricter requirements in delay and delay jitter, while non-real-time data services have stricter requirements in error ratio with looser delay requirements.

In particular, call blocking rate and call dropping rate have been considered as critical parameters in the mobile networks. Dropping a call in progress, mainly caused by handoff failure, is more annoying to users than blocking a new call. Therefore, a handoff connection usually has a higher priority than a new connection for assigning limited resources.

To reduce the call dropping rate caused by handoff failure, various schemes have been proposed [56][57][58][59][60][61]. The proposed schemes can be classified into three types. The first type is the guard channel (GC) scheme. The main idea of this scheme is to reserve a number of channels in each cell for further handoff requests. Various methods have been proposed to determine the number of GCs to be reserved. The number is determined by user’s mobility estimation [56][57] or handoff dropping probability [58]. This GC scheme reduces the handoff latency because each cell always has a certain number of GCs available for handoff calls. However, the GC scheme does not efficiently use limited radio resources, causes higher call blocking rate, and increases system complexity for estimating users’ mobility. Also, it was proposed for a network supporting a single type of service like voice, and cannot be implemented simply for diverse multimedia services since each service requires a different quality of radio resources. The second type is the handoff queuing (HQ) scheme in which the calls are queued until a channel is available [60]. This scheme is not appropriate for real-time services because of queuing delay. The third type is the resource adaptation scheme [59][60][61]. In [59], a channel sub-rating scheme had been proposed for a personal communications service (PCS) system, which serves only voice calls. When a handoff call arrives and there is no empty channel for the call, the on-going channel can be temporarily divided into two half-rate channels so that one half-rated channel serves the same on-going call and the other serves the handoff call. In [60], a similar approach had been proposed. They classified traffic into two classes: narrowband service, which cannot adjust its data rate; and wideband service, which can adjust its data rate. An adaptive rate control scheme, which is proposed in this research, is based on the third approach.

In this chapter, QoS priorities among the users in a cell are introduced, and a method that presents how the cell dynamically manages its resources to decrease the call dropping rate is proposed. Also, tradeoffs between call dropping rate and data performance, which is sacrificed for the overall QoS within the WCDMA system, are evaluated.

5.2 QoS classes in WCDMA system

QoS classes for the WCDMA system, which are also referred to as traffic classes, are defined in a technical specification [67] by 3GPP. There are four different QoS classes:

- Conversational class
- Streaming class
- Interactive class
- Background class.

The main distinguishing factor among these QoS classes is the delay sensitivity of the traffic. For example, the conversational class includes very delay-sensitive traffic, while the background class is the most delay-insensitive traffic class.

Conversational and streaming classes are intended to support real-time traffic flow. The main difference between the two classes is the degree of the delay sensitivity of the traffic. Conversational real-time traffic is the most delay-sensitive traffic. The most well-known types of traffic in this class are telephony voice, voice over IP, and video conferencing traffic. The maximum transfer delay is given by the human perception of video and audio conversation. Consequently, the limit for acceptable transfer delay is very low and strict. Streaming class traffic is used when a user is looking at real-time video or listening to real-time audio. This traffic is an one-way or asymmetric transport and characterized such that the time variation between information entities such as samples or packets should be preserved.

Interactive and background classes include traditional Internet applications like the world wide web (WWW), e-mail, telnet, file transfer protocol (FTP), and news. These types of traffic have looser delay requirements compared to conversational and streaming classes, but better error rate by means of channel coding and retransmission. The main difference between interactive and background classes is that the interactive class is mainly used by interactive applications such as web browsing, data base retrieval, or server access, while the background class is used by background traffic such as e-mail or file downloading. Round trip delay time is one of the key attributes of the interactive class. Therefore, traffic in the interactive class has higher priority in scheduling than background class traffic, and traffic in background classes uses transmission resources only when interactive applications do not need them. [67]

Table 4 [67] summarizes the QoS classes of the WCDMA network.

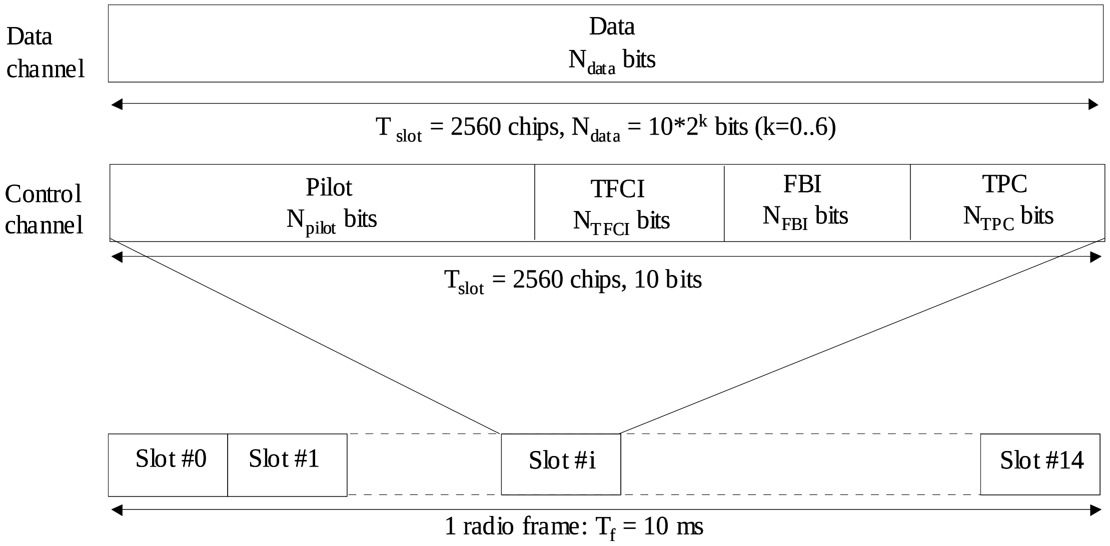
Table 4. QoS classes of WCDMA network.

Traffic class	Fundamental characteristics	Example of the application
Conversational class	- Preserve time relation (variation) between information entities of the stream - Conversational pattern (stringent and low delay)	Voice Video conferencing
Streaming class	- Preserve time relation (variation) between information entities of the stream	Streaming video
Interactive class	- Request response pattern - Preserve payload content	Web browsing
Background class	- Destination is not expecting the data within a certain time - Preserve payload content	Background download of emails / FTP

5.3 Rate control in WCDMA systems

A WCDMA system defines its channels based on the types of traffic transferred. Common physical channels are used for infrequent low bit rate traffic, while dedicated physical channels are used for high bit rate traffic. In this research, the dedicated channels are considered.

Figure 27 shows a frame structure of the uplink dedicated physical channel. The uplink dedicated physical channel is composed of two types of channels: One is the data channel that carries user data and can be integrated by one or up to 6 channels within one physical channel according to the user data rate and the channel condition; the other one is the control channel that carries control information such as pilot bits, transmit power control (TCP) commands, feedback information (FBI), and transport format combination indicator (TFI). The data channel and the control channel are in-phase/quadrature (I/Q) code multiplexed so as to transfer onto a dedicated physical channel. As shown in Figure 27, the control channel is continuously transmitted at a constant rate with a fixed spreading factor (SF) of 256. On the other hand, the data channel can change its data rate with a variable SF, which is predefined according to the channel condition and service traffic.



TFCI : Transport Format Combination Indicator
FBI : Feedback
TPC : Transmission Power

Figure 27. An uplink data frame format of a WCDMA system.

In a WCDMA system, user information bits are spread over a wide bandwidth by multiplexing the user data with quasi-random bits derived from CDMA spreading codes [63]. Chip rate is fixed as 3.84 Mcps; therefore, the user data rate and processing gain can be changed based on the SF. The range of the SF values is from 256 down to 4. The achievable data rates with different SFs are presented in Table 5 [63]. When the value of the SF is larger, the user's data rate is smaller and the user information can have a higher processing gain, which makes the data robust to the unstable radio link.

Table 5. Data rates of Uplink dedicated physical data channel.

Physical channel spreading factor	Physical channel bit rate (kbps)	Maximum user data rate with ½ rate coding (kbps)
256	15	7.5
128	30	15
64	60	30
32	120	60
16	240	120
8	480	240
4	960	480
4, with 6 parallel codes	5740	2.8M

A data channel can vary its data rate on a frame-by-frame basis, if necessary. The data rate of the current data channel is informed on the control channel with the transport format combination indicator (TFCI) field. If the TFCI is not decoded correctly, the whole data frame is lost. Data channel may consist of 1 to 6 sub-channels for higher bit rate traffic. This allows the user data rate to reach approximately up to 2 Mbps.

Since this SF can be changed dynamically according to the network conditions, the WCDMA system can perform adaptive and dynamic QoS management. Originally, the dynamic changing scheme of the SF is designed so that user traffic adapts its data transmission to the physical link conditions because the link conditions keep changing while a user moves.

However, this capability can be used to adaptively manage the total resource within a cell so as to distribute a network load and guarantee the requirements for call blocking or call dropping rate.

5.4 Interference based call admission control

A call admission control (CAC) algorithm is executed when a new call is set up or modified, and checks that the new call will not sacrifice the planned coverage area or the quality of the existing connections. The admission control algorithm estimates the load increase that the new call would cause in the network. The estimation should be separately conducted for the uplink and downlink. When the load increases by the call can be acceptable in both directions, the call is admitted; otherwise, it is rejected.

In a WCDMA system, there is no absolute maximum number of channels within a cell. The capacity is determined by the interference that is generated by all signals in its cell and neighbor cells. The interference level increased by a new user is a major critique for accepting a new call in the CAC. Consequently, any reduction on the interference converts directly into an increase of capacity. The following equations explain the interference based CAC scheme [63]. A new call request is accepted if the following requirement is satisfied:

$$I_{total\_old} + \Delta I \leq I_{threshold} \tag{5}$$

where I_{total_old} is the total interference level before the new call, ΔI is an increased interference by a new call, and $I_{threshold}$ is the maximum tolerant interference level generated by all of the users within a cell and defined by network operator's planning.

The uplink load factor, η_{UL} is normally used as the uplink load indicator. For example, η_{UL} = 0.7 means that 70% of the cell capacity is used. The η_{UL} is calculated by summing up all of the load factors generated by each call within a cell as shown in the following equation:

$$\eta_{UL} = (1 + i) \sum_{j=1}^{N} L_j \tag{6}$$

where i is the average other-to-own cell interference ratio, N is the number of calls within a cell, and Lj is the load factor of the jth call, which is defined as the following:

$$Lj = \frac{1}{1 + \frac{W}{R_j \times (E_b / N_o)_j \times v_j}} \tag{7}$$

where R_j is the bit rate of the jth call, $(E_b/N_o)_j$ is the required energy per bit to noise ratio of the jth call, and v_j is the activity factor of the jth call.

Now, the η_{UL} can be expressed as the following:

$$\eta_{UL} = (1+i)\sum_{j=1}^{N} \frac{1}{1+\frac{W}{R_j \times (E_b/N_o)_j \times v_j}} \qquad (8)$$

Finally, the *ΔI* can be calculated with the following equations:

$$\Delta I \approx \frac{I_{total}}{1-\eta} \times \Delta L = \frac{I_{total}}{1-\eta} \times \frac{1}{1+\frac{W}{R \times (E_b/N_o) \times v}} \qquad (9)$$

The interference is exponentially proportional to the number of users. Figure 28 shows the interference increases by a new call. As it approaches the maximum interference value that a cell can accept, the interference rapidly increases by a new call.

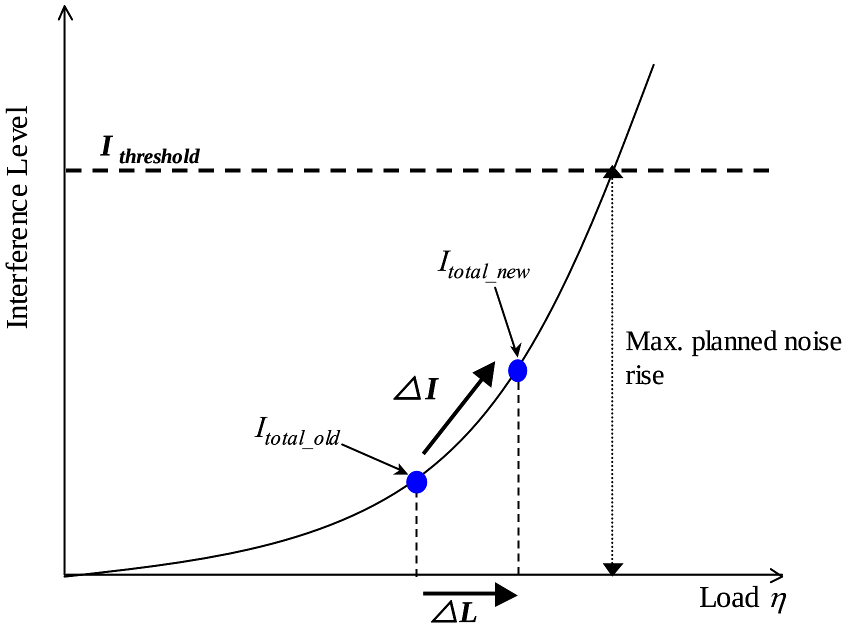


Figure 28. Interference increases according to the load.

As shown in the equation (9), the larger bandwidth traffic generates more interference. By reducing the data rate of some traffic, a cell can reduce the total interference level and get an extra radio resource that will be assigned to a new handoff call when there is no other resource available.

Reducing data rate has a different effect on each type of traffic. It causes a delay to non-real-time traffic and a lower rate of service to real-time traffic. Because real-time multimedia traffic is usually able to adopt a layered coding approach, a subset of the hierarchically coded data can be selectively chosen depending on the radio link capacity. If the cell is overloaded, the multimedia stream is filtered at the base station based on the radio resource management algorithm, and the MH only receives a subset of the multimedia stream, which causes a lower quality of service. A multimedia call dynamically changes its bandwidth depending on the network load situation during its lifetime. It guarantees a delay and delay jitter requirement by adaptively managing radio resources.

The other parameter, which can be adjusted, is the error rate. Figure 29 shows the block error rate vs. E_b/N_o when a block length, B is 100 bits and convolutional coding rates are 1/2 and 1/3 [69]. The curves can be approximated to first-order regression lines as defined in the following equation.

$$\log P_{100} \approx b_0 + b_1 * (E_b/N_0)_{dB} \qquad (10)$$

where P_{100} is a block error rate when the block length equals 100 bits, E_b/N_o is the required energy per bit to noise ratio, and b_0 , b_1 are regression coefficients that are defined as Table 6 for different coding rates.

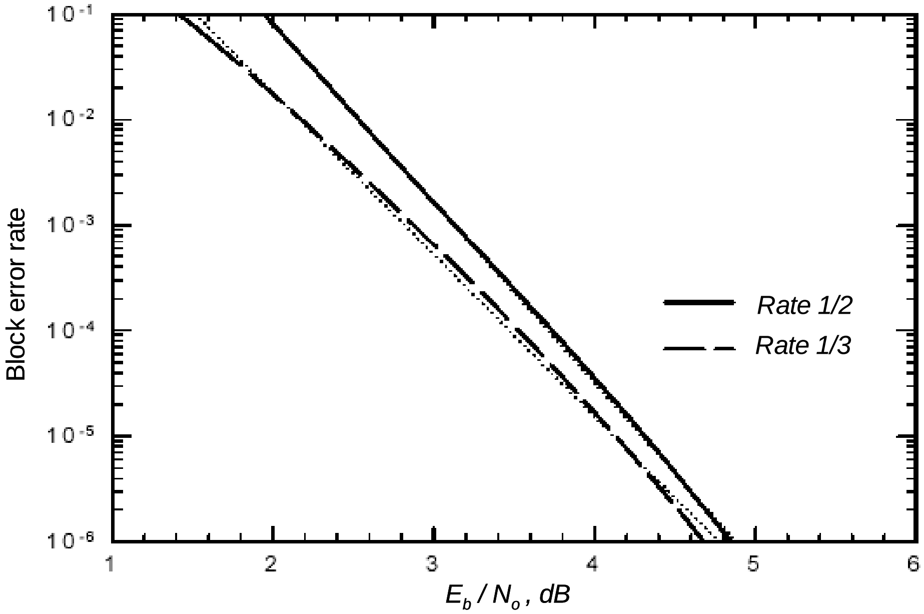


Figure 29. Block error rate vs. E_b/N_o when B=100.

Table 6. Regression coefficients b_0, b_1.

	b_0	b_1
Convolutional coding rate 1/2	2.35	–1.71
Convolutional coding rate 1/3	1.33	–1.54

We can generalize the above results to the P_B , which is a block error rate when a block length is B. P_B is proportional to the block length as shown in the following equation.

$$P_B = B * P_{100} / 100 \tag{11}$$

$$\log P_B = \log B + \log P_{100} - 2$$

$$= \log B + b_0 + b_1 * (E_b/N_0)_{dB} - 2 \qquad (12)$$

Now, the relation between E_b/N_o and block error rate for a block size B can be defined as the following equation.

$$(E_b/N_0)_{dB} \cong \frac{\log B + b_0 - 2 - \log P_B}{-b_1} \qquad (13)$$

As shown in the equation (13), E_b/N_o and block error rate, P_B is inversely proportional. Therefore, to support lower error rate, higher E_b/N_o is required, which causes higher interference. Therefore, extra radio resources are obtained to assign to the other users if higher block error rate is acceptable.

The QoS in the WCDMA system can be controlled by an appropriate selection of transmission power and processing gain as described in this section.

5.5 Management of call dropping rate / call blocking rate

In this research, we focus on an algorithm to reduce a call drop caused by a handoff failure. Dropping an on-going call is more annoying to users than blocking a new call. Therefore, a handoff call has a higher priority than a new call when a cell assigns its resources. When a MH moves within a heterogeneously integrated mobile network, the MH has higher probability of call drop since each network provides different quality of resources. Therefore, a MH and a network should have a resource management scheme that dynamically controls the available resources according to the network condition.

The proposed scheme adopts a rate adaptation scheme based on the predefined QoS priority and employs the queuing scheme as the next policy. If there is no available resource when a handoff is requested, the network tries to downgrade the quality of some ongoing traffic according to the service level agreement (SLA) matrix.

Each user has his or her SLA matrix for each of the traffic types that is defined when he or she subscribes to the service. An SLA matrix should include QoS parameters, such as maximum and minimum data rate, delay, delay jitter, error rate, call dropping rate, call blocking rate, and vulnerability level that defines the possibility of downgrading when network resources are not enough. We call the users who have better performance levels on the SLA the higher priority users. Users are classified by the SLA and have to pay based on the service quality level defined in their SLA. Therefore, we assume that the lower priority users willingly yield their resources to the higher priority users within the specified degree, depending on the network loads.

We assume there are 3 service levels (SLs): SL1, SL 2, and SL 3. Each user belongs to one of these SLs. SL 1 has the highest priority while SL 3 has the lowest priority. SLs can be defined mainly with the vulnerability level. For simplicity, we assume all of the users in the same SL have the same service quality matrix.

Figure 30 shows the proposed adaptive rate control scheme. When a handoff call request arrives, it can be accepted if there is a resource available. If not, the SL of the requested call is checked. If the call is for SL 1 traffic, downgrading the on-going lower SL traffic is attempted by decreasing the data rate of that traffic.

First, an analysis is made to determine whether some of the SL 3 traffic can be downgraded by reducing its data rate. If there are SL 3 users whose service can be downgraded by a predefined SLA, an attempt is made to reduce their data rates. If enough resources can be obtained after this attempt, the call is accepted. If not, the same process is tried onto the ongoing SL 2 traffic. If the cell still cannot get enough resources for the handoff call, the network attempts to downgrade the service quality of the handoff call. Finally, the call is accepted or put into a queue. The same procedure is conducted for the lower priority users.

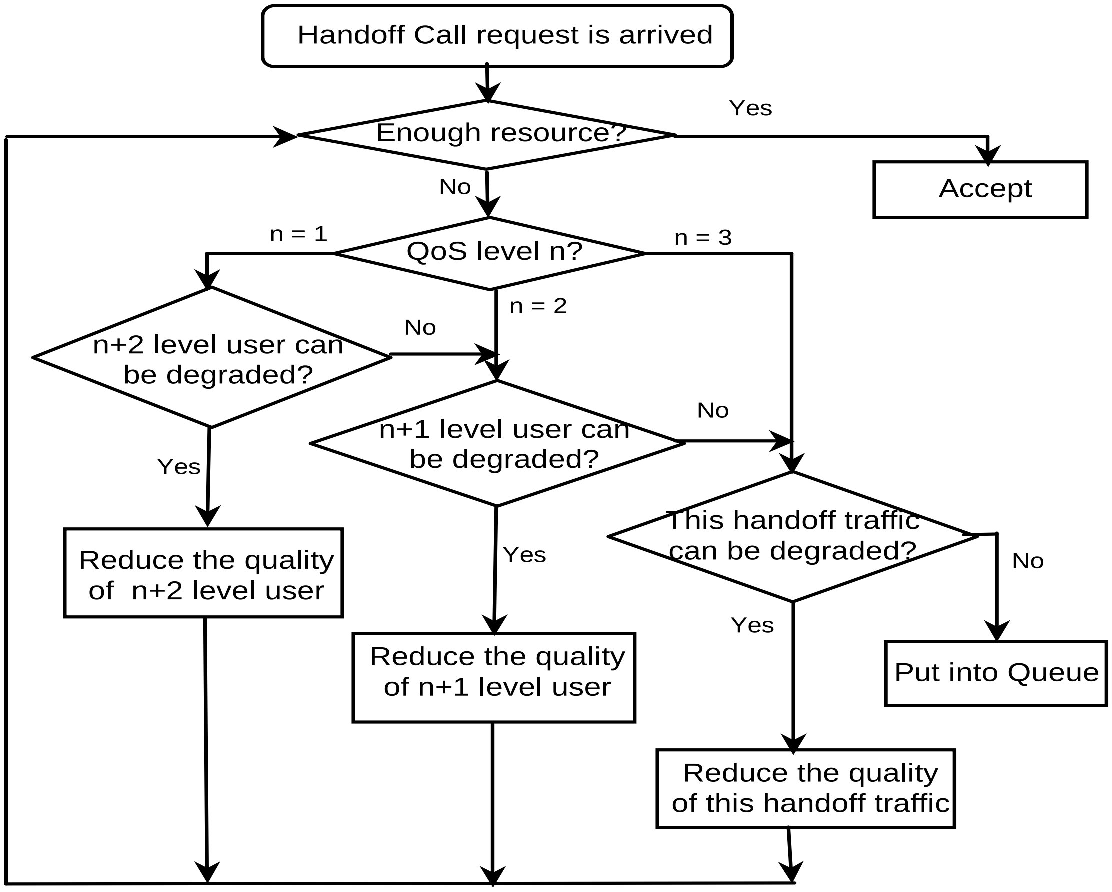

Handoff Call request is arrived
Enough resource?
Yes
Accept
No
QoS level n?
n = 1
n = 2
n = 3
n+2 level user can be degraded?
No
Yes
Reduce the quality of n+2 level user
n+1 level user can be degraded?
No
Yes
Reduce the quality of n+1 level user
This handoff traffic can be degraded?
No
Yes
Put into Queue
Reduce the quality of this handoff traffic

Figure 30. Proposed radio resource allocation scheme.

5.6 Performance

5.6.1 System Capacity

We assume the interference based CAC is used. In this research, we focus on the uplink because the uplink is more critical in defining the network capacity in the CDMA network. We use the equations presented in Section 5.4 to evaluate the variation of cell capacity. Based on the results from the cell capacity analysis, we evaluate the call dropping rate and call blocking rate using the Markov chain analysis. To simplify, we classify traffic as real-time voice traffic and non-real-time data traffic. We assume that the data traffic rate is 64Kbps and the required E_b/N_o is 10dB, while the voice traffic rate is 24Kbps with an E_b/N_o of 7dB. An activity factor is defined as 0.35 and other-to-own cell interference ratio is 0.55, and chip rate is 3.84Mcps.

Figure 31 shows total interference in a cell versus the number of users for voice and data traffic, respectively. The numerical value of total interference is multiplied by *a* based on the initial value. The interference is exponentially increased as the number of users increases. The total interference generated by data traffic increases faster, and the number of data traffic connection that a cell can support is smaller than the number of voice traffic connection.

Figure 32 shows total interference versus the number of users within a cell for mixed voice and data traffic. Table 7 presents values of some system parameters. We assume the interarrival time of voice traffic is twice that of data traffic. The interarrival time of new traffic and handoff traffic, along with service time, are defined to analyze the call blocking rate and call dropping rate in the next section. The interference increases exponentially as the number of users increases, and the interference increased by

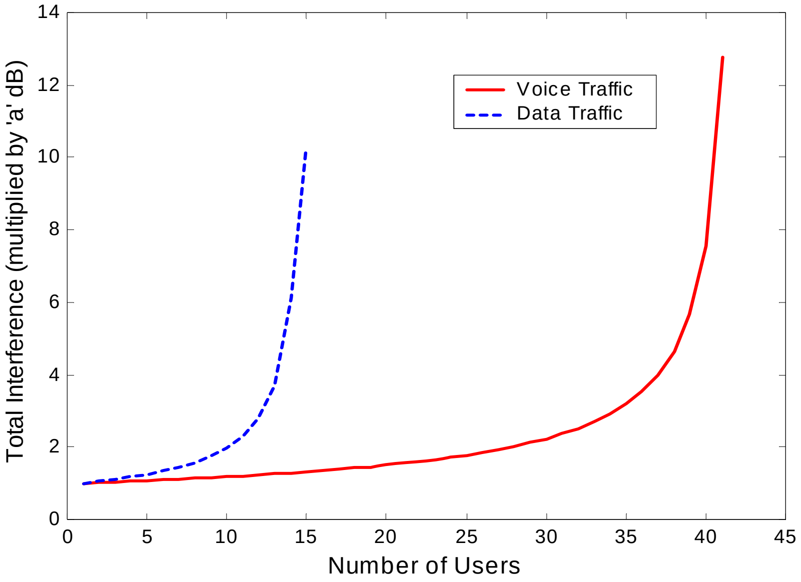


Figure 31. Total interference versus number of users for voice and data traffic.

data traffic is larger than that by voice traffic. The solid line presents the total interference of the original case with the data rate of 64 Kbps. The dashed line presents the 20% degradation case and dotted line presents the 40% degradation case, in that the bit rate of data traffic decreases from 64Kbps to 51.2Kbps and 38.4Kbps, respectively.

A system can support more users by applying the degradation scheme as shown in Figure 32. The system can support an incoming handoff call even when the total interference within a cell has already reached its $I_{\_threshold}$ value by downgrading some of the ongoing traffic that has a lower priority. The results show a cell supports about 10% and 20% more users when it downgrades the data traffic rate by 20% and 40%, respectively. As a consequence, the call dropping probability is reduced. The exact numerical values are dependent on the $I_{\_threshold}$ defined by the network operator.

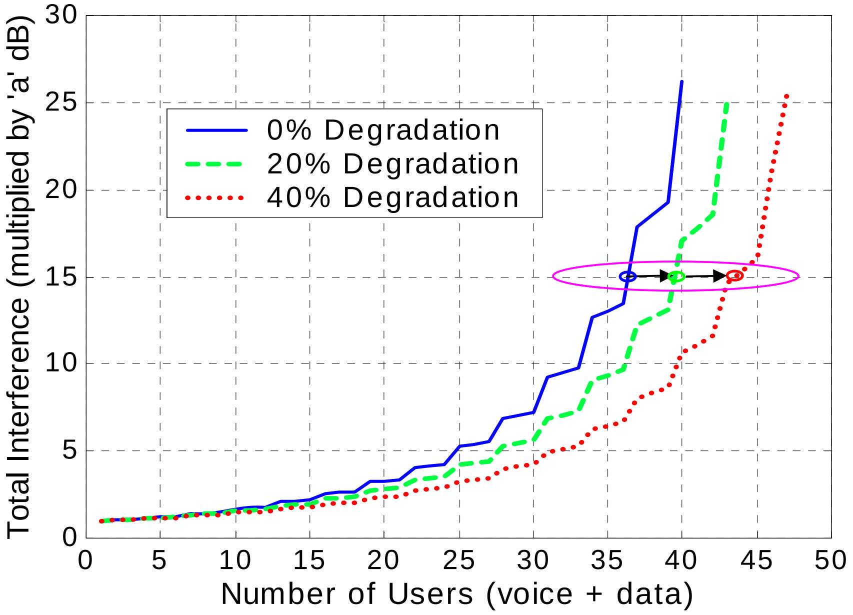


Figure 32. Total Interference variation after downgrading.

Table 7. System parameters.

Chip rate		3.84Mcps
Other-to-own cell interference ratio		0.55
Voice user	bit rate	24Kbps
	Eb/No	7dB
	activity factor	0.35
Data user	bit rate	64Kbps
	Eb/No	10dB
	activity factor	1.0
No of data users / no. of voice users		1/2
λ handoff / λ new		1/2
μ (service time)		3 min

5.6.2 Call dropping probability

Guaranteeing a handoff success for a voice call is more important than that for a data call because a dropped voice call is more annoying to users, and a voice call is narrowband traffic that is more suitable for applying the proposed scheme. Therefore, we focus only on the call dropping probability of voice calls.

To simulate the call dropping probability, the Markov chain analysis is used. Based on the CAC algorithm, the maximum number of voice users and the maximum number of data users in a cell are predefined as K_v and K_d, respectively. The number of voice users can be modeled by the M/M/K/K queuing model where K=Kv. The steady state probability that k voice users are in a cell is obtained. The call dropping probability can be expressed by inputting $k = Kv$ as shown in the following equation:

$$P_{kv} = \frac{1}{k_v!}\left(\frac{\lambda_v + \lambda_h}{\mu}\right)^{kv} P_0 \tag{14}$$

$$P_0 = \left[\sum_{i=0}^{kv}\left(\frac{\lambda_v + \lambda_h}{\mu}\right)^{i}\frac{1}{i!}\right]^{-1}$$

Figure 33 (a) shows the state transition diagram without the proposed adaptive rate control scheme and is modeled by M/M/Kv/Kv. Figure 33 (b) shows the diagram with the adaptive rate control scheme and is modeled by M/M/Kv+α/Kv+α, where α is the number of extra channels generated by downgrading some ongoing calls. Figure 33 (c) shows the model of M/M/Kv+α/Kv+α+B by the adaptive rate control scheme and queuing scheme, in which the handoff request is buffered in a queue with a size of B if it cannot be accepted even after downgrading some lower priority traffic. Before the number of voice users reaches *Kv*, the state is changed according to the arrival rate of a new voice call, λ_v, and handoff call, λ_h. After the number of voice calls in the cell reaches its maximum value, *Kv*, the cell only accepts a handoff call until the maximum number of voice users reaches $kv+\alpha$. The value of α is determined by the total interference threshold in a cell and the degree of downgrading traffic as shown in Figure 32. The value of μ is defined by the sum of service time and dwelling time, which means the call is handed off to the neighbor cell.

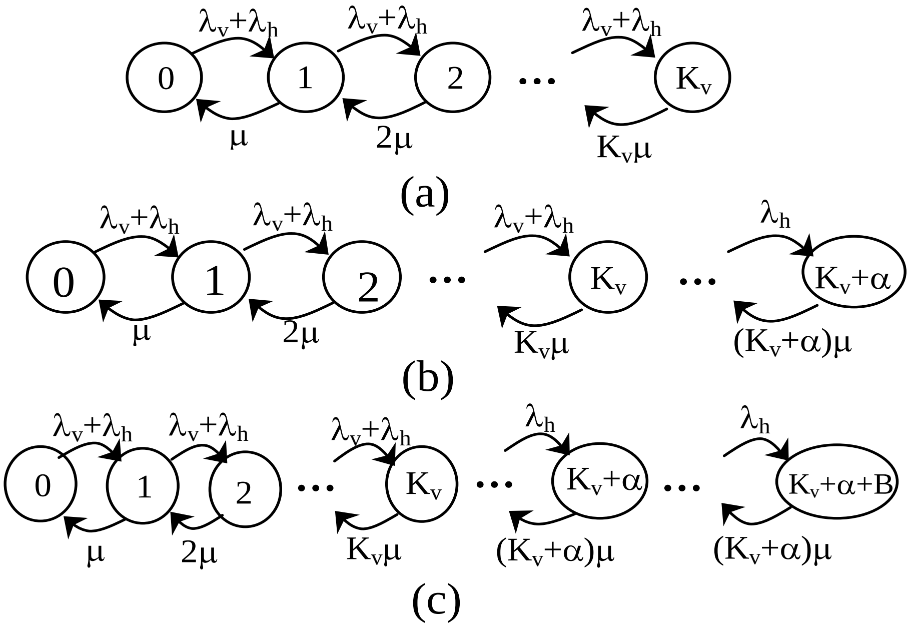


Figure 33. State Transition Diagram. (a) Without QoS aware adaptive rate control, (b) With Qos aware adaptive rate control, (c) With QoS aware adaptive rate control and queuing.

For the case of Figure 33 (b), the call blocking probability can be calculated by the following formula:

$$P_{kv+\alpha} = \frac{1}{(k_v+\alpha)!}\left(\frac{(\lambda_v+\lambda_h)^{kv}(\lambda_h)^{\alpha}}{\mu^{kv+\alpha}}\right)P_0 \qquad (15)$$

$$P_0 = \left[\sum_{i=0}^{kv}\left(\frac{\lambda_v+\lambda_h}{\mu}\right)^{i}\frac{1}{i!}+\sum_{i=kv+1}^{Kv+\alpha}\left(\frac{(\lambda_v+\lambda_h)^{kv}(\lambda_h)^{i-kv}}{\mu^{i}}\right)\frac{1}{i!}\right]^{-1}$$

For the case of Figure 33 (c), the following formulas are used to calculate the call blocking probability:

$$(\lambda_v+\lambda_h)P_0=\mu P_{1}, \qquad \text{for } k=0$$

$$(k\mu+\lambda_v+\lambda_h)P_k=(\lambda_v+\lambda_h)P_{k-1}+(k+1)\mu P_{k+1}, \qquad \text{for } 1\leq k < k_v$$

$$(k\mu+\lambda_h)P_k=(\lambda_v+\lambda_h)P_{k-1}+(k+1)\mu P_{k+1}, \qquad \text{for } k = k_v$$

$$(k\mu+\lambda_h)P_k=(\lambda_h)P_{k-1}+(k+1)\mu P_{k+1}, \qquad \text{for } k_v+1\leq k < k_v+\alpha$$

$$((k_v+\alpha)\mu+\lambda_h)P_k=(\lambda_h)P_{k-1}+(k_v+\alpha)\mu P_{k+1}, \qquad \text{for } k_v+\alpha\leq k < k_v+\alpha+B$$

$$((k_v+\alpha)\mu)P_k=(\lambda_h)P_{k-1}, \qquad \text{for } k = k_v+\alpha+B$$

$$\sum_{k=0}^{Kv+\alpha+B} Pk = 1 \qquad (16)$$

We assume the maximum number of voice users is 20 while the maximum number of data users is 10, the average service time is 3 min, and dwelling time is 60 min. The interarrival time of a voice handoff call, λ_h, is half of the interarrival time of a new voice call, λ_v.

Figure 34 shows call dropping probability versus call interarrival time. The proposed adaptive rate control scheme has a smaller call dropping probability. The dropping probability decreases as the degree of downgrade increases, and the probability is smaller when buffering is used.

If the GC scheme reserves 2 channels, which is roughly 10% of the total capacity, for handoff, the call dropping probability is same as the 20% downgrade case of the proposed scheme. However, the call blocking probability is higher, as shown in Figure 35. The proposed scheme has a smaller call blocking probability because the GC scheme always reserves some channels for handoff, causing fewer available channels for a new call. However, the proposed scheme does not increase the call blocking rate because it can have some extra channels by borrowing resources from ongoing traffic. Another advantage of the proposed scheme is that it does not need overhead information from the neighbor cells to decide the number of GCs. However, the performance degradation of the sacrificed ongoing traffic should be considered as a tradeoff.

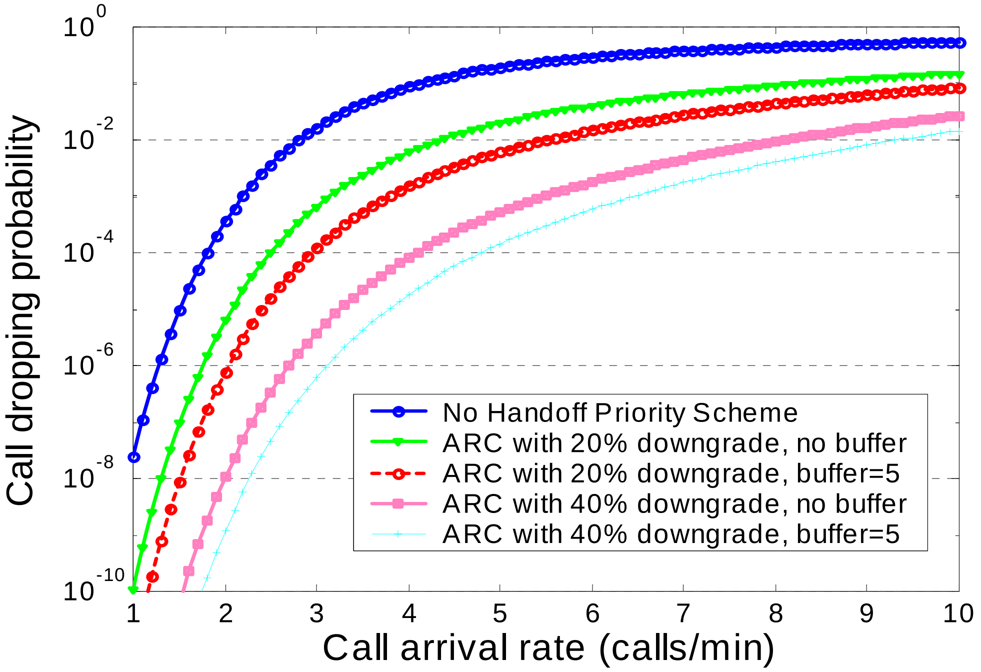


Figure 34. Call dropping probability versus call interarrival time when Kv=20.

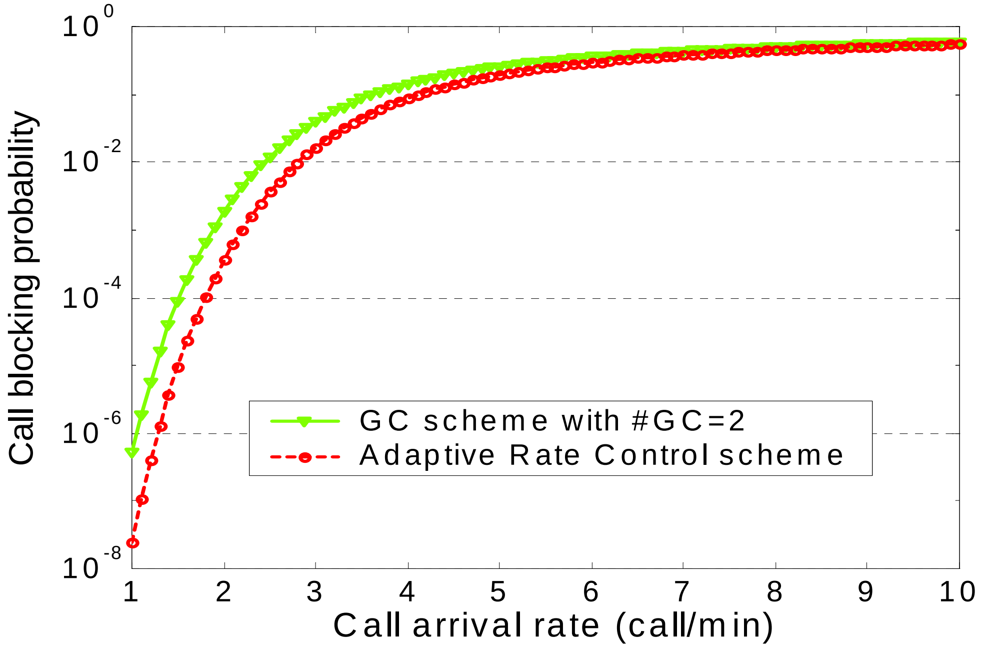


Figure 35. Call blocking probability versus call interarrival time when Kv=20 and GC scheme has 2 GCs.

5.6.3 Performance of downgraded data traffic

In this section, performance variations of the downgraded traffic are presented. The possible types of downgraded traffic are real-time multimedia traffic that is coded by multiple layers and non-real-time data traffic.

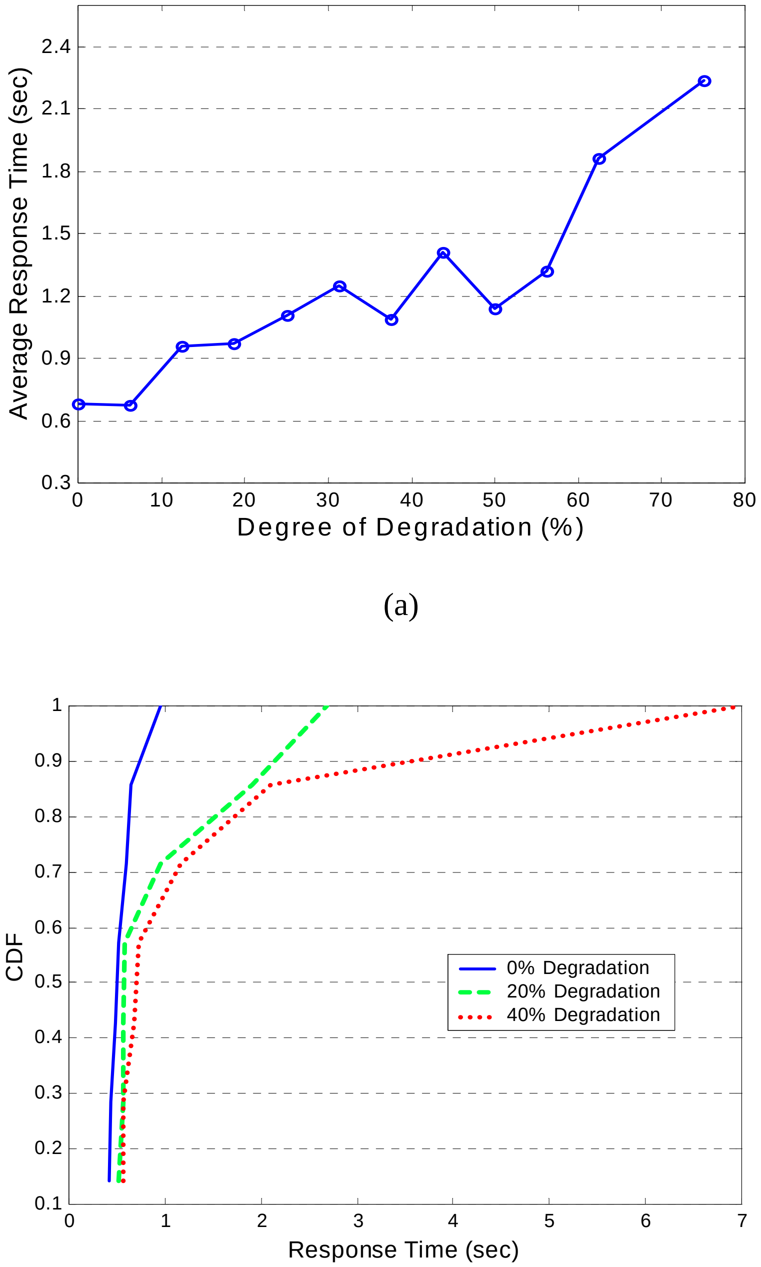
For real-time multimedia traffic, a user experiences a low quality service by downgrading. For example, the streaming service data are coded with a multi-layer coding technique and each layer may have a different importance in decoding process. A system within a network, such as BS can filter out some layers of the coded data according to the network condition. As a result of this process, the data rate can be changes because the number of layers changes. And the quality of the streaming data also changes.

For non-real-time traffic, we are especially interested in the response time of web traffic because it is the major traffic in data packet networks. The work that mathematically analyzes the performance of web traffic is too complex since web traffic has a heavy-tailed distribution. Therefore, event driven simulations using OPNET [68] are performed. Voice traffic is modeled with an AMR codec with silence suppression, and talk spurts are exponentially distributed with a mean value of 0.35 sec. Web traffic is modeled with alternating ON/OFF periods that have a heavy-tailed distribution. It is known that the Poisson process is not well suited to web traffic. The ON period depends on the size of the object. We define the object size to follow a Pareto distribution with a minimum size of 500bytes and the shape parameter α of 1.2. Page interarrival time, which is dependent on OFF time, is modeled by a Pareto distribution with a mean of 300sec and an α of 1.2 so that the reading time is also modeled with a heavy-tailed distribution. We assume these services are uniformly distributed in a cell. We simulate the impact upon the response time of web traffic depending on the degree of service downgrade. Table 8 summarizes the simulation parameters for voice and web traffic.

Table 8. Simulation parameters

Voice Traffic	Bit rate	24 Kbps (AMR codec)
	Activity factor	0.35
Web traffic	Bit rate	64 Kbps
	Page size	Pareta dist. Mean 500 byte Shape parameter 1.2
	Page interarrival time	Pareta dist. Mean 300 sec Shape parameter 1.2

Figure 36 (a) presents the amount of increase in the delay according to the service degradation of the web traffic. Average response time increases as the service level is downgraded, but occasionally fluctuates because the object size has a very large variance. Since we define the heavy-tail distribution for the object size and the think time, the amount of total data transferred for each user has also heavy-tailed characteristics. Therefore, a longer response time might be obtained even though the link has higher bandwidth.



(a)

(b)

Figure 36. Performance of downgraded web traffic (a) Response time, (b) CDF of response time

Figure 36 (b) shows the cumulative distribution function (CDF) of response time. Response time is increased as the service level is downgraded. It can also have some cross-points above a large CDF because of a heavy-tailed object size. The response time of web traffic is largely dependent on page size as well as the transmission bit rate. The impact of downgrading data rate is less serious to each web service user in terms of average response time, while it is more serious to the real-time service users.

Another attribute, which can be downgraded for non-real-time traffic, is a block error rate. High block error rate causes more retransmissions and longer delay in file transfer, however needs less E_b/N_o *,* as seen in Section 5.4. And this smaller E_b/N_o value can make extra resource when the resource is not enough for accepting an handoff call. To evaluate the performance variation of the downgraded background traffic, we simulated by changing the block error rate from 10^{-2} to 10^{-1}. Table 9 shows the file transfer time for a 1MB file.

Table 9. Performance of downgraded FTP traffic

Block error rate	10^{-2}	10^{-1}
File transfer time	135.6 sec	169.7 sec

The results show that the file tranfer time increases by about 25% when the block error rate increases from 10^{-2} to 10^{-1}.

5.7 Summary

In this chapter, an adaptive rate control scheme based on the QoS priority for handoff success in WCDMA network is proposed. The proposed scheme decreases the call drop rate by adaptively managing the data rates of existing lower priority users' services when resource is insufficient. We describe a technique that a resource can be dynamically allocated to a user within a 3G cellular network. We evaluate the performance variation in terms of a system capacity and a call dropping probability and the delay of the downgraded traffic by the proposed scheme. The results show that the proposed scheme increases system capacity and decreases the call dropping probability at the cost of small delay of the downgraded data traffic. The proposed scheme has a better system utilization than GC algorithm and it can be implemented independently of GC scheme for maximizing system performance. The proposed scheme is also applicable for CAC to decrease the call blocking rate, and to guarantee the service quality of higher priority users within networks that support many differentiated services.

CHAPTER VI

CONCLUSIONS

In this dissertation, we researched and developed a transport protocol for efficient and seamless data transmission within a heterogeneous mobile network.

The main contributions of the researches are listed as follows:

- We researched the heterogeneous mobile network and presented the technical challenges for developing the network. We described a network architecture and presented the types of handoffs that a MH would support within the network. Loosely coupled integration is the most feasible solution for the network architecture, since it allows independent deployment of different types of wireless networks. However, it increases handoff latency and packet losses. Also a vertical handoff, which should be implemented in the heterogeneous mobile network, causes some problems that the current TCP does not experience. Therefore, a transport protocol that works efficiently in this type of network should be developed.
- We analyzed RTO variation during a vertical handoff and the results revealed the RTO incorrectly increased by an instant RTT_VAR. We also analyzed the TCP throughput degradation caused by the increased RTO variations. The RTO is an essential parameter to define a TCP throughput, especially over unreliable wireless links, and an unnecessarily large RTO seriously decreases the network performance. Therefore, TCP has to know the impending handoff event, type of the handoff, and the completion of the handoff for supporting a seamless data transfer in this environment.

- We developed a transport protocol to be tailored to the heterogeneous mobile network to support seamless data transmission. In a heterogeneous network, different types of networks are integrated and a MH experiences drastic changes in network condition during a session. Traditional TCP struggles due to abrupt network changes caused by vertical handoffs and cannot work efficiently in this environment. In the proposed algorithm, the TCP is informed the impending handoff events and the type. We proposed to use two bits out of the TCP header's option field to recognize the handoff-related events. During the handoff, the proposed TCP halts its data transmission to prevent packet drops and a backed-off RTO value during a handoff. The TCP adjusts its data rate according to the type of the handoff as soon as the handoff completes. If it was a vertical handoff, the TCP re-estimates the network capacity, updates its RTO, and transfers its data. If it was a horizontal handoff, the TCP resumes its data transmission with the same rate and RTO. From simulation results, we demonstrated the proposed algorithm improves throughput and reaches a stable condition immediately and rapidly after the handoff, so provides a seamless data transfer. In addition, it is proposed with a minimum modification without an extra processing load, and it works compatibly with the currently working TCPs even though they are not modified to use the proposed scheme, since it only uses an optional field.
- We proposed an adaptive resource management scheme within a WCDMA network based on a user's priority level to reduce the call dropping and blocking rates.

In a heterogeneous network, each network provides different bandwidth to a user. A network that provides smaller bandwidth to a user may struggle with handed-off calls being served with a higher bandwidth. Therefore, an adaptive resource management algorithm should be defined in this network. We proposed a vulnerability level in the QoS attributes, which defines the possibility of downgrading quality, when network resources are not enough. If a cell does not have enough resource when a handoff call arrives, the cell tries to downgrade the quality of some existing services based on their vulnerability level. We analyzed the system capacity, call blocking rate and call dropping rate of the proposed algorithm, and simulated the performance variation of the downgraded traffic. The results showed the proposed scheme increased system capacity, and decreased the call dropping probability at the cost of small delay of the downgraded data traffic.

The research presented in this dissertation suggests future directions of interest. The Transport Control Protocol performs an important role to enhance network performance, and various transport protocols can be optimized for different types of traffic and QoS requirements. Other transport protocols tailored for specific types of multimedia services within the heterogeneous mobile network are promising areas for future research.

REFERENCES

[1] A. K. Salkintzis, “Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks,” *IEEE Wireless Commun.*, pp. 50-61, Jun., 2004.

[2] M. Buddhikot, G. Chandranmenon, S. Han, Y.W. Lee, S. Miller, and L. Salgareli, “Integration of 802.11 and third-generation wireless data networks,” *in Proc. IEEE INFOCOM*., pp.503-512, 2003.

[3] M. Stemm, and R. H. Katz, “Vertical handoffs in wireless overlay networks,” *ACM Mobile Networking and Applications (MONET)*, vol. 3, no. 4, pp.335-350, 1998.

[4] A. K. Salkintzis, C. Fors, and R. Pazhyannur, “WLAN-GPRS integration for next-generation mobile data networks,” *IEEE Trans. Wireless Commun.*, pp.112-124, Oct., 2002.

[5] T. Goff, J. Moronski, D. S. Phatak, and V. Gupta, “Freeze-TCP : A true end-to-end TCP enhancement mechanism for mobile environments,” in *Proc. IEEE INFOCOM*., pp.1537-1545, 2000.

[6] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttila, J. Mareka, R. Pichna, and J. Vallstrom, “Handoff in hybrid mobile data networks,” *IEEE Personal Commun*., vol. 7, issue 2, pp.34-47, Apr., 2000.

[7] J. W. Floroiu, R. Ruppelt, D. Sisalem, and J. V. Stephanopoli, “Seamless handover in terrestrial radio access networks : a case study,” *IEEE Commun. Magazine*, pp. 110 – 116, Nov., 2003.

[8] H. Balakrishnan, V. N. Padmanabhan, S. Seshan, and R. H. Katz, “A comparison of mechanisms for improving TCP performance over wireless links,” *IEEE/ACM Trans. Networking,* vol. 5, no 6, pp.756-769, Dec. 1997.

[9] A. Bakre, and B. Badrinath, “I-TCP : Indirect TCP for mobile hosts,” *in Proc. IEEE ICDCS*, pp.136-143, 1995.

[10] K. Brown, and S. Singh, “M-TCP : TCP for mobile cellular networks,” *ACM Computer Commun. Review,* vol. 27, no. 5, 1997.

[11] P. Sinha, N. Venkitaraman, R. Sivakumar, and V. Bharghavan, “WTCP : A reliable transport protocol for wireless wide-area networks,” *in Proc. ACM MOBICOM.*, pp. 231-241, 1999.

[12] S. Mascolo, C. Casetti, M. Gerla, M.Y. Sanadidi, and R. Wang, “TCP Westwood : bandwidth estimation for enhanced transport over wireless links,” *in Proc. ACM MOBICOM.*, pp. 287-297, 2001.

[13] M. Mathis, J. Mahdavi, S. Floyd, and A. Romanow, “TCP selective acknowledgement options,” *IETF RFC 2018*, Apr., 1996.

[14] K. Pentikousis, “TCP in wired-cum-wireless environments,” *IEEE Commun. Survey*, pp.2-14, Fourth quarter, 2000.

[15] V. Tsaoussidis, and I. Matta, “Open issues on TCP for mobile computing*,” Journal on Wireless Commun. and Mobile Computing,* 2002.

[16] 3GPP, “3rd generation partnership project; Technical specification group services and system aspects; Feasibility study on 3GPP system to wireless local area network (WLAN) interworking,” *3GPP TR 22.934*, 2004.

[17] V. Jacobson, and M. Karels, “Congestion avoidance and control,” *ACM Computer Commun. Review*, vol. 18, no. 4, pp.314-329, Aug., 1988.

[18] “Transmission control protocol,” IETF RFC 793, Sep., 1981.

[19] V. Paxson, and M. dAllman, “Computing TCP’s retransmission timer,” IETF RFC 2988, Nov., 2000.

[20] R. Ludwig, and R. H. Katz, “The Eifel algorithm : making TCP robust against spurious retransmissions,” *ACM Computer Commun. Review*, Vol. 30, No. 1, Jan., 2000.

[21] A. Gurtov, and R. Ludwig, “Responding to spurious timeouts in TCP,” *in Proc. IEEE INFOCOM.,* 2003.

[22] H. Inamura, G. Montenegro, R. Ludwig, A. Gurtov, and F. Khafizov, "TCP over second (2.5G) and third (3G) generation wireless networks," IETF RFC 3481, Feb., 2003.

[23] M. Yavuz, and F. Khafizov, "TCP over wireless links with variable bandwidth," *in Proc. IEEE VTC*, pp.1322-1327, 2002.

[24] M. C. Chan, and R. Ramjee, "TCP/IP performance over 3G wireless links with rate and delay variation," *in Proc. ACM Mobicom*, pp.71-82, 2002.

[25] E. Chaponniere, S. Kandukuri, and W. Hamdy, "Effect of physical layer bandwidth variation on TCP performance in CDMA 2000," *in Proc. IEEE VTC*, 2003.

[26] R. Wang, M. Valla, M. Sanadidi, and M. Gerla, "Using adaptive rate estimation to provide enhanced and robust transport over heterogeneous networks," *in Proc. IEEE ICNP*, 2002.

[27] P. Manzoni, D. Ghosal, and G. Serazzi, "Impact on mobility on TCP/IP : An integrated performance study," *IEEE Jour. on Selected Areas in Commun.*, Vol. 13, No. 5, pp.858-867, Jun. 1995.

[28] W. Zhuang, Y. Gan, K. Loh, and K. Chua, "Policy-based QoS management architecture in an integrated UMTS and WLAN environment," *IEEE Commun. Magazine*, pp. 118 – 125, Nov., 2003.

[29] Q. Zhang, C. Guo, Z. Guo, and W. Zhu, "Efficient mobility management for efficient handoff between WWAN and WLAN," *IEEE Commun. Magazine*, pp. 102 – 108, Nov., 2003.

[30] K. Pahlavan, P. Krishnamurthy, A. Hatami, M. Ylianttia, J. Pekka, R. Pichna, and J. Vallstrom, "Handoff in hybrid mobile data networks," *IEEE Personal Commun.*, pp. 34-47, Apr., 2000.

[31] R. Caceres, and L. Iftode, "Improving the performance of reliable transport protocols in mobile computing environments," *IEEE Jour. on Selected Areas in Commun.*, Vol. 13, No. 5, pp.850-857, Jun., 1995.

[32] C. Perkins, "Mobile IP," *IEEE Commun. Magazine*, pp.66-82, May, 2002.

[33] R. Hsieh, Z. G. Zhou, and A. Seneviratne, “S-MIP : a seamless handoff architecture for Mobile IP,” in *Proc. IEEE INFOCOM.*, pp.1774-1784, 2003.

[34] S. A. Ghorash, H. K. Cheung, F. Said, and A.H. Aghvami, “Performance of a CDMA-based HCS network with hybrid speed/overflow-sensitive handover strategy,” *IEE Proceedings. Commun,* Vol. 150, No. 4, pp.293-206, Aug., 2003.

[35] A. Pang, J. Chen, Y.Chen, and P. Agrawal, “Mobility and session management : UMTS vs. CDMA 2000,” *IEEE Wireless Commun.*, pp. 30-43, Aug., 2004.

[36] I. Akyildiz, J. Xie, and S. Mohanty, “A survey of mobility management in next-generation all-IP-based wireless systems,” *IEEE Wireless Commun.*, pp. 16-28, Aug., 2004.

[37] V. Gupta, “Media independent handover service,” Draft technical requirements 21-04-0087-08-0000, IEEE 802.21, Sep., 2004.

[38] X. Gao, G. and Wu, T. Miki, “End-to-end QoS provisioning in mobile heterogeneous networks,” *IEEE Wireless Commun.*, pp. 24-34, Jun., 2004.

[39] G. Xylomenos, G. Polyzos, P. Mahonen, and M. Saaranen, “TCP performance issues over wireless links,” *IEEE Commun.*, pp. 52 -58, Apr., 2001.

[40] A. Zanella, G. Procissi, M. Gerla, and M. Sanadidi, “TCP Westwood : analytic model and performance evaluation,” *in Proc. IEEE GLOBECOM,* pp.1703-1707, 2001.

[41] C. Casetti, M. Gerla, S. Mascolo, M. Sansadidi, and R. Wang, “ TCP Westwood : End-to-end congestion control for wired/wireless networks,” Wireless Networks Journal 8(4), pp. 467-479, 2002.

[42] M. Gerla, M. Y. Sanadidi, R. Wang, A. Zanella, C. Casetti, and S. Mascolo, “TCP Westwood: congestion window control using bandwidth estimation,”, in *Proc. GLOBECOM,* pp.1698 – 1702, 2001.

[43] M. Shi, X. Shen, and J. Mark, “IEEE 802.11 roaming and authentication in wireless LAN/cellular mobile networks,” *IEEE Wireless Commun.*, pp. 66 – 75, Aug., 2004.

[44] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, “Modeling TCP throughput : a simple model and its empirical validation,” *ACM Computer Commun. Review*, 28: pp.303-314, 1998.

[45] J. Padhye, V. Firoiu, D. Towsley, and J. Kurose, “Modeling TCP Reno performance : a simple model and its empirical validation,” *IEEE/ACM Trans. on Networking,* Vol.8, Issue 2, pp.133-145, Apr., 2000.

[46] R. Braden, “Requirements for Internet Hosts – Communication Layers,” *IETF RFC 1122*, Oct., 1989

[47] H. Balakrishnan, S. Seshan, and R.H. Katz, “Improving reliable transport and handoff performance in cellular wireless networks, “ *ACM Wireless Networks*, vol. 1, Dec., 1995.

[48] S. Floyd, and V. Jacobson, “Random early detection gateways for congestion avoidance,” *IEEE/ACM Trans. on Networking*, V.1 N.4, Aug. 1993, pp. 397-413.

[49] K.K. Ramakrishnan, S. Floyd, and D. Black, “The addition of explicit congestion notification (ECN) to IP,” *IETF RFC 3168,* Sep., 2001.

[50] S. Floyd, “TCP and explicit congestion notification,” *ACM Computer Commun., Review*, V24 N. 5, Oct. 1994, pp. 10-23.

[51] V. Jacobson, R. Braden, and D. Borman, “TCP extension for high performance,” *IETF RFC 1323*, May, 1992.

[52] K. Ramkrishnan, and S. Floyd, “A proposal to add Explicit congestion notification (ECN) to IP,” *IETF RFC 2481*, Jan., 1999.

[53] F. Khafizov, and M. Yavuz, “Running TCP over IS-2000,” in *Proc. of ICC*, pp.3444-3448, 2002.

[54] W. Richard Stevens, “TCP/IP Illustrated, Vol 1 : The protocols”, *Addison-Wesley,* 1994.

[55] M. Allman, V. Paxson, and W. Stevens, “TCP congestion Control,” *IETF RFC 2581*, Apr., 1999.

[56] NS-2 network simulator, http://www.isi.edu/nsnam/, March 2006.

[57] S. Choi, and K.G. Shin, “Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks”, *in Proc. ACM SIGCOMM,* pp. 155-166, 1998.

[58] C. Oliveira, J. B. Kim, and T. Suda, “An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks”, *IEEE Journal on Selected Areas in Commun..*, vol. 16, pp. 858-874, Aug., 1998.

[59] X. Luo, B. Li, I. L. Thng, Y. Lin, and I. Chlamtac, “An adaptive measured-based preassignment scheme with connection-level QoS support for mobile networks”, *IEEE Trans. on Wireless Commun*., vol. 1, no. 3 , pp. 521-529, Jul., 2002.

[60] Y. Lin, A. R. Noerpel, and D. J. Harasty, “The sub-rating channel assignment strategy for PCS hand-offs,” *IEEE Trans. on Vehicular Technology*, vol. 45, no 1, pp. 122-130, Feb., 1996.

[61] W. Zhuang, B. Bensaou, and K. Chua, “Adaptive Quality of service handoff priority scheme for mobile multimedia networks,” *IEEE Trans. on Vehicular Technol*ogy, vol. 49, pp. 494-505, Mar., 2000.

[62] S. Lu, and V. Bharghavan, “Adaptive resource management algorithms for indoor mobile computing environments”, *in Proc. ACM SIGCOMM,* 1996.

[63] V. Bharghavan, K. Lee, S. Lu, S. Ha, J. Li, and D. Dwyer, “The TIMELY adaptive resource management architecture”, *IEEE Personal Commun*. Vol 5. Issue 4, pp.20-31, Aug., 1998.

[64] H. Holma, and A.Toskala, “WCDMA for UMTS: Radio access for third generation mobile communications”, Third edition, *Wiley, John & Sons*, 2004.

[65] K.Sipila, Z. Honkasalo, J. Laiho-Steffens, and A. Wacker, “Estimation of Capacity and Required Transmission Power of WCDMA Downlink Based on a Downlink Pole Equation,” *in Proc. IEEE VTC spring*, pp.1002-1005, 2000.

[66] J. Zhang, J. Huai, R. Xiao, and B. Li, “Resource management in the next-generation DS-CDMA cellular networks,” *IEEE Wireless Commun.*, pp. 52–58, Aug., 2004.

[67] Third generation partnership project, http://3gpp/org, Dec., 2005.

[68] 3GPP, "Quality of Service (QoS) concept and architecture", *3GPP TS 23.107 3.9.0*, September 2002.

[69] OPNET Simulator, http://www.opnet.com, March, 2006.

[70] UMTS model user guide, 2004, *OPNET Technologies.*

[71] S. Kim, and J. Copeland, “Enhancing TCP performance for intersystem handoff within heterogeneous mobile networks,” in *Proc. of IEEE Vehicular Technology Conference (VTC*), pp. 2276-2280, May, 2004.

[72] S. Kim, J. Copeland, “Interworking between WLANs and 3G wireless networks : TCP challenges,” in *Proc. of IEEE Wireless Communications and Networking Conference (WCNC),* pp.1252-1257, Mar., 2004.

[73] S. Kim, J. Copeland, “TCP for seamless vertical handoff in hybrid mobile data networks”, in *Proc. of IEEE Globecom Conference,* pp.661-665, Dec., 2003.

[74] S. Kim, H. Kim, and J. Copeland, “Simulation Based Study of Adaptive Rate Control Scheme in UMTS Network Using OPNET,” in *Proc. of OPNETWORK 2003*, Aug., 2003.

[75] S. Kim, H. Kim, and J. Copeland, “Adaptive Rate Control Scheme for Handoff and its Performance Evaluation in Mobile Multimedia Networks”, in *Proc. of 57th IEEE Vehicular Technology Conference(VTC),* pp.1445-1449, Apr., 2003.

[76] J. Chung, S. Kim, and J. Copeland, “Reliable Wireless Multicast Using Fast Low-Density Erasure Codes”, in *Proc. of 57th IEEE Vehicular Technology Conference (VTC*), pp.1218-1222, Apr., 2003.

[77] S. Kim, H. Kim, J. Copeland, “Dynamic Radio Resource Allocation Considering QoS in UMTS Network”, in *Proc. of 4th IEEE Conference on Mobile and Wireless Communications Networks(MWCN),* pp.636-640, Sep., 2002.