QUALITY-OF-SERVICE PROVISIONING AND RESOURCE RESERVATION MECHANISMS FOR MOBILE WIRELESS NETWORKS

THESIS

Presented to the Graduate Council of
University of North Texas in Partial
Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Rajeev Jayaram, B.E.
Denton, Texas
August, 1998

The advent of powerful hand-held computers and the desire for communication on the move, are the driving forces behind an emerging technology called _Mobile Computing_. Today, mobile computing users are being provided with applications such as electronic mail and calendar/diary programs. Observing the growing demands of roaming users, it is predicted by the mobile computing research community that the next generation wireless networks will be burdened with bandwidth-intensive traffic generated by personal multimedia applications such as web browsing and traveler information systems. However, the available bandwidth for supporting these applications is rather limited, and proper management of the bandwidth is necessary to accommodate the envisaged high-bandwidth applications. For multimedia traffic (voice, video, and text) to be supported successfully, it is necessary to provide _Quality-of-Service (QoS) provisioning_ between the end-systems.

In this thesis, a framework for QoS provisioning in next generation wireless access networks is proposed. The framework aims at providing a differentiated service treatment to real-time (delay-sensitive) and non-real-time (delay-tolerant) multimedia traffic flows at the _link layer_. Novel techniques such as _bandwidth compaction, channel reservation, and channel degradation_ are proposed. Using these techniques, we develop a _call admission control algorithm_ and a _call control block_ as part of the QoS framework.

The performance of the framework is captured through analytical modeling and simulation experiments. By analytical modeling, the average carried traffic and the worst case buffer requirements for real-time and non-real-time calls are estimated. Simulation results show a 21% improvement in call admission probability of real-time calls, and a 17% improvement for non-real-time calls, when bandwidth compaction

The advent of powerful hand-held computers and the desire for communication on the move, are the driving forces behind an emerging technology called Mobile Computing. Today, mobile computing users are being provided with applications such as electronic mail and calendar/diary programs. Observing the growing demands of roaming users, it is predicted by the mobile computing research community that the next generation wireless networks will be burdened with bandwidth-intensive traffic generated by personal multimedia applications such as web browsing and traveler information systems. However, the available bandwidth for supporting these applications is rather limited, and proper management of the bandwidth is necessary to accommodate the envisaged high-bandwidth applications. For multimedia traffic (voice, video, and text) to be supported successfully, it is necessary to provide Quality-of-Service (QoS) provisioning between the end-systems.

In this thesis, a framework for QoS provisioning in next generation wireless access networks is proposed. The framework aims at providing a differentiated service treatment to real-time (delay-sensitive) and non-real-time (delay-tolerant) multimedia traffic flows at the link layer. Novel techniques such as bandwidth compaction, channel reservation, and channel degradation are proposed. Using these techniques, we develop a call admission control algorithm and a call control block as part of the QoS framework.

The performance of the framework is captured through analytical modeling and simulation experiments. By analytical modeling, the average carried traffic and the worst case buffer requirements for real-time and non-real-time calls are estimated. Simulation results show a 21% improvement in call admission probability of real-time calls, and a 17% improvement for non-real-time calls, when bandwidth compaction
is employed. The channel reservation technique shows a 12% improvement in call admission probability in comparison with another proposed scheme in the literature.

Of late, the Internet is being used to carry real-time data such as voice and video, for which resource reservation is an important issue. As increased number of mobile computers connect to the Internet, the QoS commitment offered to fixed Internet hosts will need to be offered to mobile users. This thesis proposes a QoS reservation protocol for mobile computers in the Internet. The reservation protocol works at the network layer and can be easily integrated into the existing Internet infrastructure.
QUALITY-OF-SERVICE PROVISIONING AND RESOURCE RESERVATION MECHANISMS FOR MOBILE WIRELESS NETWORKS

THESIS

Presented to the Graduate Council of University of North Texas in Partial Fulfillment of the Requirements

For the Degree of

MASTER OF SCIENCE

By

Rajeev Jayaram, B.E.
Denton, Texas
August, 1998
ACKNOWLEDGMENTS

I am immensely grateful to my advisor Dr. Sajal Das for his encouragement and guidance at every stage of this work.

I am thankful to my committee members, Dr. Tom Jacob and Dr. Armin Mikler, for their constructive suggestions towards the improvement of this thesis.

I would like to thank the Department of Computer Sciences at UNT for providing academic resources and financial support throughout my tenure as a student. I also acknowledge the support from the Texas Advanced Technology Program.

My sincerest thanks to my friends at UNT, Sanjoy Sen and Naveen Kakani, for the many hours of discussions that shaped this thesis.

Finally, this endeavor would have been impossible without the love, support, encouragement, and blessings of my parents. I will forever be indebted to them.
TABLE OF CONTENTS

1 Introduction 1

1.1 Issues in Mobile Computing ................................. 1
   1.1.1 Wireless communications .............................. 1
   1.1.2 Mobility ............................................ 2
   1.1.3 Portability .......................................... 2
1.2 Cellular Wireless Networks .................................. 3
1.3 Quality-of-Service Issues in Cellular Wireless Networks ........ 5
1.4 Quality-of-Service Issues in the Internet .................... 8
1.5 Contributions of this Thesis ............................... 10
1.6 Chapter Organization ...................................... 10

2 Related Work 11

2.1 Previous work on multimedia support in cellular environments ... 11
   2.1.1 Motivation .......................................... 15
2.2 Previous work on resource reservation in the Internet ........... 15
   2.2.1 Motivation .......................................... 18

3 A Quality-of-Service Provisioning Framework 20

3.1 A General Framework for QoS Provisioning ................. 21
   3.1.1 Packet Sorter ...................................... 22
   3.1.2 Scheduling .......................................... 23
   3.1.3 Degradation Policies ............................... 25
   3.1.4 Bandwidth Reservation Based on User Location Prediction .... 25
   3.1.5 Reservation through Channel Borrowing ................ 27
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.6</td>
<td>Bandwidth Compaction</td>
<td>28</td>
</tr>
<tr>
<td>3.1.7</td>
<td>Call Admission Algorithm</td>
<td>30</td>
</tr>
<tr>
<td>3.1.8</td>
<td>Bandwidth Stealing</td>
<td>32</td>
</tr>
<tr>
<td>3.2</td>
<td>Performance Analysis</td>
<td>32</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Call Admission and Successful Hand-off for Real-time Calls</td>
<td>32</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Carried Traffic under Bandwidth Degradation</td>
<td>34</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Average Queue Lengths for RTQ and NRTQ</td>
<td>36</td>
</tr>
<tr>
<td>3.3</td>
<td>Simulation Experiments</td>
<td>38</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Simulation Model</td>
<td>38</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Admission Control</td>
<td>39</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Frame Structure and Scheduling</td>
<td>39</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Compaction Implementation</td>
<td>42</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Simulation Parameters</td>
<td>43</td>
</tr>
<tr>
<td>3.3.6</td>
<td>Simulation Results</td>
<td>43</td>
</tr>
<tr>
<td>3.3.7</td>
<td>Comparison of Reservation Scheme</td>
<td>45</td>
</tr>
<tr>
<td>4</td>
<td>A Reservation Mechanism for Mobile Nodes in the Internet</td>
<td>48</td>
</tr>
<tr>
<td>4.1</td>
<td>Internet Architecture</td>
<td>48</td>
</tr>
<tr>
<td>4.1.1</td>
<td>Scope of Resource Reservation</td>
<td>49</td>
</tr>
<tr>
<td>4.2</td>
<td>Neighbor Mobility Agent Discovery Protocol (NMADP)</td>
<td>50</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Direct Method</td>
<td>50</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Indirect Method</td>
<td>52</td>
</tr>
<tr>
<td>4.3</td>
<td>Mobile Reservation Update Protocol (MRUP)</td>
<td>54</td>
</tr>
<tr>
<td>4.4</td>
<td>Analytical Model for Resource Reservation</td>
<td>57</td>
</tr>
<tr>
<td>4.5</td>
<td>Simulation Model for Reservation Mechanism</td>
<td>62</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Assumptions and Model</td>
<td>62</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Experiments and Results</td>
<td>62</td>
</tr>
</tbody>
</table>
5 Conclusions

5.1 Research Contributions .................................................. 65
5.2 Future Work ................................................................. 67

REFERENCES ................................................................. 68
LIST OF TABLES

2.1 Summary of QoS provisioning activities at various layers ........... 12

3.1 Example of traffic mix .............................................. 41
LIST OF FIGURES

1.1 Cellular Wireless Network Architecture .................................. 4
1.2 Characteristics of multimedia traffic ................................. 7

3.1 A Framework for Link Layer QoS Provisioning in Wireless Multimedia Systems ................................................................. 21
3.2 Classification of users in a cell ........................................... 26
3.3 Channel borrowing in destination cells ............................... 28
3.4 Bandwidth allocation pattern showing segments and pages ...... 29
3.5 Flowchart of the call admission algorithm ......................... 31
3.6 $r$ versus $P_{av}$ .......................................................... 35
3.7 Admission Control Messaging ........................................... 39
3.8 Segmenting packets to FMA1 frame ................................. 42
3.9 Call admission probability for real-time users versus call arrival rate for queue size = 100 and service rate $\mu = 0.03$ .......... 44
3.10 Call admission probability for non-real-time users versus call arrival rate for queue size = 100 and service rate $\mu = 0.03$ .......... 45
3.11 Call admission probability for non-real-time users versus queue size for arrival rate $\lambda = 0.08$ and service rate $\mu = 0.03$ .......... 46
3.12 Call admission probability for non-real-time users versus call arrival rate for queue size = 100 ................................. 46
3.13 Hand-off drop probability versus hand-off rate for queue size = 100, service rate $\mu = 0.01$ and call arrival rate $\lambda = 0.1$ .......... 47
3.14 Call admission and successful hand-off probability versus traffic load ................................. 47

4.1 Subnet Architecture .......................................................... 49
4.2 NFA Discovery Message and Routing Table .......................... 51
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3 Illustration of Distributed and Centralized NMA Discovery</td>
<td>53</td>
</tr>
<tr>
<td>4.4 Format of Reserve Resource message</td>
<td>56</td>
</tr>
<tr>
<td>4.5 Markov Model for State of a Mobility Agent</td>
<td>59</td>
</tr>
<tr>
<td>4.6 Arrival rate vs. Call dropping probability for service rate = 0.1</td>
<td>63</td>
</tr>
</tbody>
</table>
CHAPTER 1

Introduction

Since the introduction of the cellular telephone over a decade ago, the growth of wireless communications has been rapid. A host of new "untethered" communications services, under the general heading of Personal Communications Services (PCS), have recently been deployed. As computing devices have shrunk in size and weight (e.g., laptop and computers, personal digital assistants), and their processing capabilities are increasing by the day, untethering the communications link will make possible the next step in the evolution of computing environments. This emerging computing environment is being called Mobile Computing. Mobile computing promises users with more flexible use of computing resources, because users can be freed from being physically connected to the underlying network. The ultimate goal of mobile computing is to enable a multitude of users at any place, access information from anywhere, at any time.

1.1. Issues in Mobile Computing

The technical challenges to establishing mobile computing are non-trivial. There are three distinct issues to be addressed, each stemming from an essential property of mobile computing [15].

1.1.1. Wireless communications

Untethered access to network resources means mobile computers need wireless communications support. There are distinct differences between wired and wireless communications, that make wireless communications harder to achieve. Wireless communications are of much lower quality than wired connections. Wireless links are
characterized by lower bandwidths, higher error rates, frequent spurious disconnections, and transmission security problems. Wireless communications may also be degraded by mobility. Users may go out of the range of network transceivers, or there could be too many users in a particular area resulting in network overload [11, 13].

1.1.2. Mobility

The ability to change locations while connected to the network increases the volatility of data. For example, a fixed client computer can be configured manually on which server to use for a particular service, whereas a mobile computer will need to find servers dynamically. The mobility related issues are data management and location management of the mobile user.

1.1.3. Portability

The portability issues are concerned with the design of lightweight terminals with high storage capacity, low power consumption, and effective user interfaces. Design of portable computers for wireless applications entails consideration of mobility related issues like disconnections due to unreliable links, as well as communication aspects like low bandwidth availability.

There are significant overlaps between mobility and wireless communication domains, and none of the above three mobile computing issues can be considered independent of the other.

Mobile Computing is a general term that encompasses a number of specialized areas, including wireless computing, ubiquitous computing, nomadic computing and decoupled computing. Wireless computing refers to computing systems that are connected to their working environment via wireless links, such as radio frequency (RF) or infrared (IR) [22]. Ubiquitous computing [43] envisions an environment in which computing devices are so inexpensive and readily available that a wireless communications in-
Infrastructure is built around these devices. *Nomadic computing* refers to the ability to compute as the user relocates from one support environment to another. *Decoupled computing* refers to the ability to compute when detached from existing computing and communications infrastructure.

Realization of truly seamless mobile computing is still quite a few years away. Today's research and prototyping efforts are directed towards various components of mobile computing. A sample of areas being investigated includes research in wireless communications, networking technologies for integration of mobile laptop computers into existing computer and communication networks, networking technologies to provide broadband access for mobile computers, and application design for mobile computers. In this thesis, the focus is on wireless networks and mobility aspects.

1.2. Cellular Wireless Networks

There is no industry-wide consensus on what the next generation broadband access backbone network(s) should be in order to support multimedia services. *Asynchronous Transfer Mode* (ATM) technology [4] switches are being hailed as the backbone of the future for providing packet switched services. The large infrastructure that exists in the form of the *Public Switched Telephone Network* (PSTN), and the *Integrated Services Digital Network* (ISDN) are also being used to foray into the multimedia services market. Irrespective of what the backbone network of the future will be, it can be said with certainty that the last hop to providing multimedia services to the roaming user will be through the *cellular wireless network*, at least over large geographic areas.

Although radio technology is quite old, the possibility of untethered communication was made possible by the birth of *cellular wireless networks*. Wireless communications is an integral part of cellular technology. Figure 1.1 shows the schematic diagram of the *Global System for Mobile Communications* (GSM) cellular wireless network. Without the box labeled BSC, this architecture is also referred to as the
Public Land Mobile Network (PLMN), in North America.

A geographic area is partitioned into cells\(^1\). Each cell is serviced by a Base Station (BS). The Mobile Station (MS), or the mobile user communicates with the BS through radio frequencies (wireless link or air interface). Several BS's are connected to a Base Station Controller (BSC), and several BSC’s are connected to a Mobile Switching Center (MSC). The MSC handles switching of calls between MS’s in different cells through the BSC’s and BS’s. The MSC is also responsible for providing services to the MS. The MS (or subscriber) information is stored on a permanent basis at the Home Location Register (HLR), and on a transient basis at the Visitor Location Register (VLR). The HLR communicates with an Authentication Center (AC) for MS validation, before services can be provided to the MS. The MSC also has interfaces with the PSTN, which is the wired telephone network, the Service Control

\(^1\)Cells are widely represented as hexagons in the literature, although in reality they could be arbitrarily shaped.
Point (SCP), for providing intelligent network services to the MS, and with the ISDN network for future provisioning of broadband services.

1.3. Quality-of-Service Issues in Cellular Wireless Networks

In spite of the tremendous growth of mobile communication users, the frequency spectrum allocated by the FCC (Federal Communications Commission) for servicing mobile users is very limited. Recently, more frequency bands were auctioned to support Personal Communication Services (PCS) by the FCC, and as a result of this, applications such as wireless email\(^2\), and wireless-fax are being offered to mobile users today. Even with the latest increase in frequency allocations, the demand for bandwidth is expected to become increasingly prominent with the emergence of, and the need to support next generation wireless multimedia applications for roaming users with portable computers. Typical next generation applications include video-on-demand, news-on-demand, web browsing and traveler information systems. Proper management of available spectrum is necessary to accommodate these high bandwidth applications.

The dominating topic in today's efforts on the development of future communication networks is how to bring about an improvement in the quality of communication. For multimedia traffic (voice, video, and data) to be supported successfully, it is necessary to provide *Quality-of-Service* (QoS) guarantees between the end-systems.

QoS provisioning means that the multimedia traffic should get predictable service from resources in the communication system. Typical resources are CPU time (for the communication software to execute) and network bandwidth. The communication software must also guarantee an acceptable end-to-end delay and maximum delay jitter, i.e., maximum allowed variance in the arrival of data at the destination. In most cases, QoS requirements are specified by the 3-tuple: \((\text{bandwidth}, \text{delay}, \text{reliability})\).

The QoS provisioning problem for multimedia traffic in non-wireless networks such as

\(^2\)AT&T's recently introduced PocketNet email service.
broadband wire-line networks (e.g., B-ISDN) has been extensively studied (for a good introduction to the relevant issues, refer to [23]). The ongoing work in the wire-line field mainly concentrates on the problems of bandwidth management, and switch-based scheduling to provide deterministic guarantees on end-to-end delay, throughput and packet losses.

The preceding QoS aspects will cause problems when including wireless environments, due to the following major differences between wire-line and wireless networks.

1. **Link characteristics**: The transmission quality is generally lower and time variant (due to factors such as fading influences), and the communication bandwidth is limited.

2. **Mobility**: The location of subscribers is not known a priori and may change during a connection.

The broadband wire-line network transmission links are characterized by high transmission rates (in the order of Gbps) and very low error rates \(10^{-8}\). In contrast, wireless links have a much smaller transmission rate (Kbps-Mbps) and a much higher error rate \(10^{-3}\). For example, the commercially available wide area wireless data networks such as ARDIS or Mobitex offer a channel rate of 8Kbps to 2Mbps and similar local area wireless networks such as Motorola’s Altair-II offer about 6 Mbps, in comparison to a thin Ethernet that offers 10 Mbps or an ATM network that offers several Gbps. Additionally, wireless links experience losses due to multi-path dispersion and Rayleigh fading. The second major difference between the two networks is user mobility. In wire-line networks, the user-network-interface (UNI) remains fixed throughout the duration of a connection whereas the UNI in a wireless environment can keep changing throughout the connection. Therefore, it is necessary to re-design or revise the usual QoS provisioning techniques for wireless networks.

Figure 1.2 shows the characteristics of various traffic types in wireless networks, in terms of the bandwidth usage and typical tolerable delay (the bandwidth axis is
in bps and the delay axis in seconds). Since the traffic varies significantly within a wide range of parameters, the QoS provisioning becomes even more challenging. From this viewpoint, multimedia traffic can be broadly classified as real-time and non-real-time [1]. Real-time traffic (e.g., video and voice) is highly delay sensitive, while non-real-time traffic (e.g., Transmission Control Protocol (TCP) packets and text data transfers) can tolerate large delays.

Multimedia traffic is also characterized by its burstiness and stream-oriented nature, and for keeping a long and continuous load on the network. Existing public data networks such as Cellular Digital Packet Data (CDPD), General Packet Radio Service (GPRS), and High Speed Circuit Switched Data (HSCSD) utilize the unused voice capacity to support low-priority, non-real-time data. In case of scarcity of available bandwidth, the transmitted data packets are buffered or suitable flow control techniques are used leading to an increase in the transmission delay.

Despite the recent auction of 1850-2000 MHz band by the FCC for personal com-
communication services (PCS) users, bandwidth is still the major bottleneck in most
real-time multimedia services. Such services can substantially differ in bandwidth
requirements, e.g. 9.6 Kbps for voice service and 76.8 Kbps for video. Most of the
earlier research on wireless bandwidth allocation concentrated on the problem of op-
timizing frequency reuse, and hence the carried traffic, for only one class of service,
namely voice. For providing multiple classes of wireless services, the carried traffic for
each class in the system has to be considered individually. Also, existing literature
deal separately with resource reservation approaches and call admission control in
wireless multimedia networks.

1.4. Quality-of-Service Issues in the Internet

The Internet has enjoyed an explosive growth that began in the early 1990’s and is
expected to continue well into the next millennium. As more and more users, and
new applications access the Information Superhighway, the Internet has to evolve at
a faster rate than ever before to meet the increasing demands being placed on it.

Voice over Internet is the hottest telecommunications buzz word today. Of late,
the Internet is being used to carry real-time data such as voice and video. The
economics behind this trend are evident -- lower cost of information delivery when
compared to traditional telecommunication networks, scope for introduction of in-
novative user interfaces for applications like Internet Telephony, and integration of
basic communications (such as telephony, fax) with software such as web browsers
and calendar managers.

There are several barriers in adapting the Internet to real-time uses. The most
prominent problem is in the poor quality of connections on the Internet. The Internet
data delivery model is often referred to as best-effort. There are no firm guarantees
offered to data flows\(^3\) in the Internet. For traditional non-real-time Internet traffic

\(^3\)A data flow is a set of packets belonging to a traffic stream from a source to destination in the
network
such as File Transfer Protocol (FTP) data, this best-effort delivery model has not been a problem. However, multimedia applications are very sensitive to the quality of service their packets receive. As multimedia applications become ubiquitous on the Internet, the need for "better than best-effort" network services will become inevitable. Recognizing this growing scope of the Internet, the Internet Engineering Task Force (IETF) [19] is developing a wide range of protocols and architectures to support real-time and multimedia traffic over the Internet.

Soon, it is envisaged that mobile computers will connect to the Internet using wireless communications technologies such as radio or infrared. Initially this will most likely be in environments such as those of university campuses or office buildings where LAN connectivity is abundant, and geographic areas are not very widespread. For example, a user can take his laptop computer from his office to a conference room if the entire office was connected via wireless LANs. Mobile nodes can change their point of attachment from one network or subnetwork to another. A mobile node does not lose its IP (Internet Protocol is the network layer protocol) address when it changes its point of attachment, and may continue to communicate with other Internet nodes using its (constant) IP address. To accommodate mobile nodes with wireless access into the Internet, the IETF has defined Mobile IP [32] and Mobile IPv6 [20]. As the number of mobile computers grows, the QoS commitments offered to fixed users on a LAN will need to be offered to mobile users. As mentioned earlier, mobile nodes bring along with them two distinct differences - link characteristics and mobility. Wireless links have significantly lower transmission rates than their wired counterparts. Several efforts have been recently made to address the QoS-deliverance problem over wireless links.

Node mobility implies that reservation has to be performed at all places a mobile node may visit, during the lifetime of a data flow connection. This is a significant problem since the movement of a mobile node is unpredictable to a large extent.
1.5. Contributions of this Thesis

The most important contribution of this thesis is the development of an integrated framework for QoS provisioning at a lower layer (e.g., the radio link layer) combining various novel approaches to call admission control, channel reservation, bandwidth degradation and a new technique of improving spectrum utilization called bandwidth compaction. The dynamic, error-prone behavior of the wireless physical link necessitates such low level QoS control.

The second major contribution of this thesis is addressing the problem of resource reservation at the Network layer for mobile nodes. We propose a simple and scalable resource reservation mechanism that allocates resources to a currently ongoing data flow of the mobile node in all neighboring subnets.

1.6. Chapter Organization

The rest of this thesis is organized as follows. Chapter 2 gives an overview of related research in QoS provisioning for next generation wireless networks and QoS provisioning architectures for the Internet. Chapter 3 presents a QoS provisioning framework for multimedia traffic at lower layers. We discuss in depth our traffic classification algorithm, and the proposed bandwidth management algorithms. Chapter 4 presents two protocols for guaranteeing resource reservations for mobile nodes in the Internet at the network layer. Finally, chapter 5 concludes this thesis, with a summary and directions for future work.
CHAPTER 2

Related Work

Due to the increasing demands of multimedia applications, efficient and effective QoS provisioning has become increasingly important. To support QoS requirements, communication systems and end-systems must provide bandwidth and latency characteristics that allow timely transmission of information. Successful provisioning of end-to-end service quality is a distributed function to be realized in the link, network, transport, and application layers. In this chapter, we review the ongoing efforts at the lower communication layers. In particular, we restrict our attention to related work at the cellular link layer and the Internet network layers. Table 2.1 gives an overview of the QoS-related issues that are being studied at the various layers for mobile computing. We have included key references for the interested reader.

2.1. Previous work on multimedia support in cellular environments

Recently, some work has been proposed to guarantee QoS for multimedia traffic in wireless cellular networks [30, 1, 2, 25, 40, 39, 31, 33, 38].

Rappaport and Purzynski [33] have developed analytical models for a cellular mobile environment consisting of mixed platform types with different classes of channel and resource requirements. The different platform types are pedestrians, automobiles etc. Different call types require different types and amounts of resources. The authors have considered two broad cases - with hand-off queuing and without. In both cases, hand-off calls are given priority over ordinary calls with the former given a certain cut-off resource priority over the latter. Also, quotas for the usage of each type of resource is implemented. The problem is mapped into multi-dimensional Markov chains, with permissible states determined by the resource usage constraints. The
underlying driving processes of the model are call arrivals, call completions, hand-off arrivals into and departures from a cell. Numerical algorithms are devised to solve for the steady state probabilities. Various performance measures like carried traffic, blocking and forced termination probabilities for each platform and call type are numerically computed from the analytical models.

The carried traffic in a wireless network can be increased by the graceful degradation of some or all of the existing services in the system. Seal and Singh [40, 39] have identified two QoS parameters namely, graceful degradation of service and guarantee of seamless service. Graceful degradation of service refers to reducing allocated bandwidth to the existing calls. The quality of each connection deteriorates as data is discarded by the base station transmitter to adjust to the reduced allocated bandwidth. It is highly possible that discarding of data results in the loss of some critical

Table 2.1: Summary of QoS provisioning activities at various layers

<table>
<thead>
<tr>
<th>Layer</th>
<th>Activity</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application</td>
<td>Disconnected operation</td>
<td>[36]</td>
</tr>
<tr>
<td></td>
<td>Application adaptation</td>
<td>[21]</td>
</tr>
<tr>
<td></td>
<td>Caching, Replication</td>
<td></td>
</tr>
<tr>
<td>Transport</td>
<td>Indirect-TCP</td>
<td>[6]</td>
</tr>
<tr>
<td></td>
<td>NFS over wireless</td>
<td>[14]</td>
</tr>
<tr>
<td></td>
<td>Wireless ATM</td>
<td>[34]</td>
</tr>
<tr>
<td>Network</td>
<td>Routing</td>
<td>[32]</td>
</tr>
<tr>
<td></td>
<td>Address translation mechanisms</td>
<td>[20]</td>
</tr>
<tr>
<td></td>
<td>Survey of proposals</td>
<td>[5]</td>
</tr>
<tr>
<td>Link</td>
<td>Multiple access techniques</td>
<td>[16, 3]</td>
</tr>
<tr>
<td></td>
<td>Scheduling algorithms</td>
<td>[10, 8]</td>
</tr>
<tr>
<td>Physical</td>
<td>Spreading, Coding, Framing</td>
<td>[17, 18, 28]</td>
</tr>
</tbody>
</table>
portions of data which may not be recoverable. With the help of user-supplied loss profiles which tell the system the user-preferred way to lose data, bandwidth usage of applications that can sustain loss is degraded in situations where user demands exceed the network's capacity to satisfy them. A new transport sub-layer called, loss profile transport sub-layer (LPTSL), is proposed to implement loss profiles by selectively discarding data out of special applications like a compressed video stream. This is implemented as a library of discarding functions which discard data in various manner (e.g., clustered loss, uniform loss etc.), and the user chooses the most appropriate function according to his needs. The algorithms have been incorporated in the Multi-Stream Protocol (MSP) and MPEG-2 transport systems.

Based on the minimum requirement criteria provided by the user, Oliveira, Kim and Suda [30] proposed a bandwidth reservation algorithm for guaranteeing QoS to multimedia traffic. Two main types of multimedia traffic was considered – real-time and non-real-time. For real-time traffic, the call is admitted only if the requested bandwidth can be reserved in the call-originating cell and all its neighbors. For a non-real-time call, the requested bandwidth is reserved only in the originating cell. Various approaches for bandwidth reservation for real-time traffic in neighboring cells were suggested. The amount of bandwidth reserved is either a function of the number of real-time calls or the requested bandwidth of all real-time connections in the cell. Detailed simulation experiments were performed to compare this scheme with two other variants. In one variant, the incoming real-time call is accepted if the requested bandwidth is available. If not, the algorithm attempts to allocate the minimum required bandwidth (provided by the call) only if it is a hand-off request. In the second variant, if this minimum required bandwidth is not available, bandwidth is “stolen” from the ongoing non-real-time calls and allocated to the hand-off request. Although this scheme guarantees QoS, the main drawbacks are: (i) bandwidth is reserved redundantly, since the user moves only to one of the six neighboring cells (assuming hexagonal cell geometry), and (ii) the stringent call admission procedure
might not admit many real-time requests in a highly overloaded system.

Acampora and Naghshineh [1] proposed a virtual connection tree (a cluster of base stations) approach which is used for call routing, call admission and resource allocation. This concept is used to reduce the call set-up and routing load on the network call processor such that a large number of mobile connections can be supported. A virtual connection tree is a collection of base stations and wire-line network switching nodes and links, with the root being a fixed switching node. For setting up a call for a mobile terminal, the call processor allocates two sets of virtual connection numbers (one in each direction), with each member pair of the set defining a path between the root and a base station in the virtual connection tree. When the mobile user wishes to hand-off to another BS in the same tree, it simply begins to transmit packets with its allocated connection number for the new BS. In the reverse direction, a hand-off is identified by an arrival packet bearing a new connection number. In both cases, the routing table at the root of the tree is updated accordingly. Whenever the mobile user reaches the boundary of a virtual connection tree, it seeks hand-off to a new tree, and the network call processor is invoked to perform an inter-tree hand-off. The network traffic is divided into three classes in decreasing order of priority. Class I comprises of real-time traffic like voice/video and has hand-off dropping probability as its QoS metric. Class II connections can be “put on hold” and will suffer packet losses and delay when the system enters an overloaded state. Class III connections utilize the leftover bandwidth of the other two classes of traffic and its QoS metric is average queuing delay. Call admission decision is based on the number of ongoing calls of each class, hand-off rate, and call duration statistics.

Acampora and Naghshineh [2] also proposed a call admission control algorithm for QoS provisioning for multimedia traffic, based on an adaptive resource sharing policy among real-time and non-real-time classes of traffic, with the former class having preemptive priority over the latter. A simple analytical model (based on Poisson call arrivals and departures, and threshold rules for each class of traffic) is proposed
to capture the various QoS parameters like call blocking and dropping probabilities and the probability for a minimum bandwidth availability for non-real-time requests which share the available bandwidth equally among themselves.

2.1.1. Motivation

One of the motivations behind our proposed QoS framework is to provide differential treatment at a low layer to the two important classes of wireless multimedia traffic - those generated respectively by real-time (delay-sensitive) and non-real-time (delay-tolerant) applications. This is achieved through sorting the real-time and non-real-time packets by a packet sorter and providing for packet marking and various priority scheduling techniques, as required. Driven by the industry's policy-driven networking paradigm, we keep our framework flexible enough to incorporate various QoS provisioning policies at the discretion of the network operator.

2.2. Previous work on resource reservation in the Internet

Recognizing the growing scope of the Internet, the Internet Engineering Task Force (IETF) [19] is developing a wide range of protocols and architectures to support real-time and multimedia traffic.

The Integrated Services (intserv) working group is defining new service classes for Internet traffic. When these service classes are supported by routers in the Internet, data flows can be provided with certain QoS commitments. The two important service classes proposed by the intserv working group are (a) Guaranteed Service [37], and (b) Controlled-Load Service [44]. Guaranteed service provides a data flow, an assured level of bandwidth, a firm end-to-end delay bound, and no queuing loss provided that the data flows traffic conforms to its specified traffic parameters. Guaranteed service is intended for applications with stringent real-time requirements. Unlike guaranteed service, Controlled-Load service offers no firm guarantees. If a flow requesting Controlled-Load service is accepted, the network commits that the flow will be offered
a service equivalent to that offered to a best-effort flow when the network is unloaded.

To avail these services, the QoS requests generated by applications are passed to the routers by a reservation protocol. The Resource ReSerVation Protocol (RSVP) [45] is by far the most popular of such protocols. RSVP addresses the needs of applications requiring QoS, promising per-flow service. RSVP has positioned itself as a purely resource reservation protocol and relies on input from the routing protocol to route its own messages, and setup reservations. The sender specifies its traffic characteristics via a Sender TSpec. The Sender TSpec is carried by the RSVP protocol in its PATH message. A PATH message is routed to a destination (a receiver) using information obtained from the routing protocol. On receiving a PATH message from a source, a receiver sends a reservation RESV message. The RESV message contains the flowspec, which is the traffic profile the receiver is willing to receive. RESV messages are routed back to the sender on the reverse path of PATH messages. Each intermediate router commits resources to the flow by examining the RESV message, and thus a reservation is established.

The reliance of RSVP on a per-flow state and per-flow processing has scalability problems during deployment. This has led to an attention shift to Differentiated Services (diffserv), which offers a significantly simpler alternative by defining only a few service classes, instead of supporting each individual flow. This greatly reduces scalability problems which make diffserv architectures deployable.

The Two Bit Differentiated Services Architecture [27] encompasses two types of services. (a) Assured Service [9], and (b) Premium Service [27]. Assured service assures that the user’s traffic is unlikely to be dropped as long as it stays within the expected capacity profile. An Assured service traffic flow may exceed its Profile, but the excess traffic flow is not given the same assurance level. Premium service is provisioned according to peak capacity profiles that are strictly not oversubscribed and that is given its own high priority queue in routers. A Premium service traffic flow is shaped and hard-limited to its provisioned peak-rate and shaped so that bursts
are not injected into the network.

The current routing protocols used in the Internet are transparent to any particular QoS requirements that data packets may have. This implies routing decisions are made by routers without any knowledge about resource availability at other routers. As a result, flows are often routed over paths that are unable to support their requirements, while alternate paths with sufficient resources may be available. QoS Routing attempts to alleviate the above problem by identifying paths with sufficient resources.

To accommodate mobile nodes with wireless access into the Internet, the IETF has defined Mobile IP [32] and Mobile IPv6 [20]. As the number of mobile computers grows, the QoS commitments offered to fixed users will need to be offered to mobile users. However, as we have seen before, mobile computers bring with them two distinct differences—link characteristics and mobility. Wireless links have significantly lower transmission rates than their wired counterparts. Several efforts exist that address the QoS-deliverance problem over wireless links [16, 3, 8, 10]. In this thesis, as mentioned earlier, we present a QoS framework for lower layers.

Node mobility implies that reservation has to be performed at all places a mobile node may visit, during the lifetime of a data flow connection. This is a significant problem since the movement of a mobile node is unpredictable to a large extent. There has been very scant research done in the area of providing mobile nodes on the Internet with resource reservation, although a significant body of literature exists for fixed nodes and several efforts are underway.

Recently, recognizing that RSVP is inadequate to provide QoS guarantees to mobile nodes, Talukdar, et al.[41] have proposed Mobile RSVP (MRSVP), a protocol with extensions to RSVP to support Integrated Services to mobile nodes. MRSVP extends RSVP by allowing a mobile node to make passive reservations from all possible subnets it may visit during the lifetime of a connection. An active reservation is made from the sender host to the subnet in which the mobile node is currently residing. MRSVP identifies the foreign agent of the subnet the mobile is currently in,
as a local proxy agent and the foreign agents' of the subnets the mobile may possibly visit, as remote proxy agents. MRSVP defines several messages to handle mobility in addition to RSVP's messages. The Join-group message is sent by a mobile node to its proxy agent directing the proxy agent to join a multicast group. The Spec message is used by the mobile to provide its flow specification to its remote proxy agents. The MSpec message is provided by the mobile to a local proxy agent and contains the possible subnets the mobile node may visit. The Terminate message is used by the mobile node to request its remote proxy agents to terminate a reservation.

2.2.1. Motivation

It has been recently realized that promising per-flow service and maintenance of per-flow state is not scalable. This is evident in the way different architectures are gaining popularity.

A critical drawback of MRSVP is that MRSVP relies on the mobile node to supply its mobility specification (MSpec). The MSpec is a list of all care-of-addresses of the mobile node in the foreign subnets it may visit. There are two issues that are not clearly addressed in this approach: (a) Many a time the care-of-address will not be known until the mobile node attaches itself to the foreign subnet. This is usually the case since mobile nodes are likely to acquire care-of-addresses dynamically using a protocol such as Dynamic Host Configuration Protocol (DHCP) [7]; (b) There is no guarantee that the MSpec will be complete. In spite of reservations being setup, the mobile node may not even use the passive reservation, if it moves to a subnet that does not belong to the MSpec. Another drawback of MRSVP is that it uses RSVP signaling from multiple points. This means a tremendous amount of network signaling occurs, which is costly, especially when it serves a single user.

This has prompted us to address simpler mechanisms of promising resource reservations for mobile nodes in the Internet. We propose a simple and scalable resource reservation mechanism that allocates resources to a currently ongoing data flow of
the mobile node in all neighboring subnets. We propose two simple protocols, *The Neighbor Mobility Agent Discovery Protocol* (NMADP), and the *Mobile Reservation Update Protocol* (MRUP) to achieve the reservation setup and maintenance.

We conclude this chapter with the important note that the 'product' of service quality and mobility is nearly constant, i.e. a maximum degree of QoS will imply restrictions on mobility. There is bound to be degradation of QoS with unlimited mobility. The research strive is for an acceptable level of service in large cells and higher levels of QoS in smaller dedicated areas.
A Quality-of-Service Provisioning Framework

In this chapter, we propose a framework for QoS provisioning in next generation wireless networks, with a goal to guarantee multimedia traffic the best possible service from the lower layers of the network. Multimedia services can be broadly classified as real-time (or delay-sensitive) and non-real-time (or delay-tolerant). The proposed framework aims at providing a differentiated service treatment to real-time and non-real-time traffic flows at the link layer by classifying and queuing the packets arriving at the link layer in two separate queues — real-time packet queue (RTQ) and non-real-time packet queue (NRTQ). Various novel strategies are proposed to support differential treatment and guarantee QoS, namely, bandwidth compaction, channel reservation based on user location prediction and degradation. Using these strategies, we develop a call admission control algorithm and a call control block as part of the QoS framework.

The performance of the QoS provisioning framework is captured through analytical modeling, as well as simulation experiments. The analytical model compares our reservation and call admission schemes with an existing proposal due to Oliveira, et al [30]. Significant improvement is predicted, which is then validated by simulation experiments. By analytical modeling, we also estimate the carried traffic and the worst case buffer requirements for real-time and non-real-time calls. Simulation studies show significant improvement of the system performance using some of the functionalities proposed within our QoS provisioning framework.
3.1. A General Framework for QoS Provisioning

Figure 3.1 depicts a framework for low layer QoS provisioning in wireless multimedia systems. During the call setup period, the higher layers are required to provide the following input parameters to our link-layer QoS framework: (i) average bandwidth required, (ii) minimum bandwidth required, and (iii) whether the application is delay-tolerant (non-real-time) or delay-sensitive (real-time). We call this 3-tuple input a Requirement Profile. The higher layers can obtain the requirement profile information by mandating an RSVP-style [45] of reservation setup signaling, before accepting a data flow into the network.

The Admission Controller decides whether to admit a flow or not based on the input requirement profile, and existing traffic-related conditions in the system. The admission controller algorithm is described in detail in section 3.1.7.

![Diagram](image)

Figure 3.1: A Framework for Link Layer QoS Provisioning in Wireless Multimedia Systems

Once the admission controller admits the call in the cell, it passes the admission decision and the requirement profile of the user to the Packet Sorter. The packet sorter
sorts and classifies traffic packets of admitted flows. Its functionality is detailed in section 3.1.1

The Call Control Block in Figure 3.1 is mainly responsible for various policy-driven schemes like (i) scheduling different classes of packet, (ii) call degradation or reducing bandwidth allocation to degradable applications in face of scarcity of available radio channels, (iii) bandwidth reservation for highly delay-sensitive high priority applications, (iv) bandwidth compaction to maximize utilization of available channel resources, and (v) radio resource usage monitoring.

The call control block also contains the supervisory algorithms that manage and monitor the progress of the ongoing sessions. To enforce the degradation and compaction policies, the call control block interacts with the radio resource manager of the local system, while the channel reservation policy necessitates communication with a remote system (e.g., another BSC).

3.1.1. Packet Sorter

Our proposed framework differentiates between the packets generated by real-time and non-real-time traffic sources at the link layer. These packets are directed towards the mobile user from the base station controller (BSC) on the forward link. We emphasize such service differentiation in the forward link in the proposed architecture because, the next generation wireless applications are expected to be likewise asymmetric. Therefore, we do not consider traffic flows on the reverse link in our framework.

At the link layer, the upper layer packets, depending on their size, may have to be segmented into link layer packets. The segmentation size would be implementation-dependent. The packet sorter interfaces with a two-level priority queue in the system – the real-time packet queue (RTQ) and the non-real-time packet queue (NRTQ), the former having a higher scheduling priority over the latter. Note that both RTQ and NRTQ are abstractions for actual physical buffer implementations, of which there can
be multiple instances under each type. For example, the system could implement a multi-level queue for real-time traffic, based on delay jitter or reliability requirements for further differentiating the real-time traffic packets.

The packet sorter has the following main functionalities:

1. **Identify** traffic packets entering the QoS sub-layer based on a flow id.

2. **Sort** the incoming packets based on the *packet type* information in the requirement profile.

3. **Segment** the packets and append the flow id to the segmented packets.

4. **Mark** the packets by setting a certain bit pattern in their header to distinguish them further, so that the scheduler (and also the radio resource manager in the system) can treat them differentially while allocating radio resources (e.g., time slot) or while faced with resource constraints. One example of this service differentiation is the priority of real-time packets over non-real-time, hence, in case of buffer overflow, the non-real-time packets are discarded first. Another example might be to allocate smaller rate time slots to packets generated by degradable applications, and allocate larger time slots to non-degradable ones. This concept leads to more efficient bandwidth usage.

5. **Route** the packets into the appropriate queue.

3.1.2. Scheduling

Real-time multimedia traffic such as those generated by video or voice telephony is highly delay-sensitive, while traffic such as text e-mail or fax can tolerate large delays. Due to the delay tolerant nature of non-real-time traffic, such traffic packets can be *buffered* in case of bandwidth unavailability. If bandwidth is available, the scheduler allocates a fixed amount, which we call a *bandwidth page*, is allocated to non-real-time traffic packets on a time-sharing basis. Real-time traffic, on the other hand,
cannot be delayed beyond a certain duration and is assumed blocked or dropped if the minimum specified bandwidth, which we call a bandwidth segment, is not available during scheduling. The usage of the terms bandwidth segment and page will be justified when we describe our bandwidth compaction algorithm in section 3.1.6.

When a real-time call request arrives and finds all channels occupied, it may, under certain circumstances, force one (or more) ongoing non-real-time calls to be temporarily buffered so that the released channel can be used to admit the real-time request. Therefore, the scheduler offers real-time traffic preemptive priority over non-real-time traffic.

Recall that a channel is a fixed block of communication medium such as a 2-tuple (time slot, carrier frequency) in TDMA (Time-Division Multiple Access) systems, or simply a fixed radio frequency as in FDMA (Frequency Division Multiple Access) systems. Multiple channels can be allocated to a single user to satisfy higher bandwidth requirements, which is equivalent to a bandwidth segment by our definition. Otherwise, slots larger than a certain minimum duration are defined to support higher bandwidth traffic. For example, in the FRAMES [28] multiple access proposal for UMTS (Universal Mobile Telecommunications System), the GSM evolution for 3G wireless systems, three kinds of slot structure are defined. These are $\frac{1}{16}$th and $\frac{1}{8}$th rate slots (which correspond to bandwidth segments), and a $\frac{1}{64}$th rate slot (corresponding to bandwidth page). IS-136 [42] is another cellular radio interface standard which specifies allocation of multiple time slots to high bandwidth users.

In our framework, we propose the simplest packet scheduling approach, which is to schedule packets based on the queue priority, where a packet from a non-real-time queue (NRTQ) is never scheduled until the real-time queue (RTQ) is empty. This might lead to starvation of the non-real-time packets (leading to queue build-up and ultimately, buffer overflow) if the real-time applications continue to proliferate packets. A weighted fair queuing or a multi-level feedback queue based approach may be more suitable for implementation.
3.1.3. Degradation Policies

Degradation policies are incorporated in our framework in a distributed manner. The packet sorter marks packets of flows that are degradable and the scheduler provides treatment accordingly. For example, degradable flows may not be scheduled at times of bandwidth shortage, or may even be discarded. For this framework, the following two types of degraded behavior are considered.

*Bandwidth degradation* [11, 40] implies that an application occupying multiple channels releases some of them to enter a degraded mode. If the average rate specified in the requirement profile equals the minimum rate for the application, then the application is not degradable. For example, some forms of compressed video may exhibit this behavior. Bandwidth degradation may be precipitated by various circumstances, e.g., when the mobile user faces sudden resource constraint while trying to hand over to a heavily loaded cell. Degradation can be achieved using various flow control techniques, e.g., packet discard at link and transport layers, or by applying hierarchical coding schemes [38].

*Delay degradation* implies an increase in the delay jitter or delay variance of a session. Typically, non-real-time applications are expected to exhibit a large amount of delay degradation if the number of higher priority real-time applications grow in the system. An analysis of queue build-up and delay degradation for non-real-time packets in a mixed traffic scenario is presented in [11].

3.1.4. Bandwidth Reservation Based on User Location Prediction

For frequent and arbitrary movement of the user in a cellular system, the problem of guaranteeing QoS in case of hand-off becomes a critical issue. This is particularly true for high bandwidth applications using multiple channels (bandwidth segment) in a cellular environment. A plethora of reservation schemes have been proposed in the literature in various forms [30, 1, 40], to guarantee user connectivity and service quality in case of frequent hand-off. In order to guarantee QoS, most of these schemes
reserve resources in all the neighboring cells of the current cell of the user. This over-engineering has its price not only in blocking scarce resources in neighboring cells, but also results in an increased cost to the user. We propose a resource reservation technique based on predicting the user mobility in a cell which can significantly reduce the cost of reservation.

The mobile users in a cell are classified as departing or local depending on whether or not they are located within the shaded region in Figure 3.2. This entails an accurate location prediction of the user within the cell boundary (error margin of less than 200 sq. feet for a typical cell of 5 sq. miles). A simple way to predict user location is based on the received signal strength (RSS) from the mobile equipment to the base station. Another technique is based on the phase delay of the received signal. Both the schemes can lead to severe inaccuracies due to obstacles in the signal propagation path and various radio propagation conditions like shadowing or multi-path fading. Hence, it is likely that a scheme which predicts user location by combining the received signal strength measurement (by the BS) with a knowledge-base of the terrain conditions in that cell, will outperform both of the above schemes. From field measurements, a few points (typically, a hundred of them) with known values of RSS are chosen as anchor points within the cell. Based on the RSS from the user and the RSS at the user from neighboring base stations, the anchor point closest to the user is computed. Then, the location of the user is predicted relative
to its closest neighboring anchor point whose coordinates are assumed to be known. As a byproduct of this algorithm, three neighboring cells (or sectors) closest to the current location of the user can also be determined. Let these cells be called the set of destination cells of the user (e.g., cells $D_1, D_2$ and $D_3$ for the cell $A$ in Figure 3.3).

If an error in user location prediction is tolerable, then the destination cell(s) for the user can also be computed from only RSS measurements as follows. The neighboring base station from which the RSS is the maximum, is most likely the closest neighboring cell of the user. After the base station detects that a user has entered the peripheral region (by comparing its RSS with certain threshold), it instructs the mobile station to send the RSS measurements from all the neighboring BS's and the corresponding cell ids. The three closest neighboring cells of the user are predicted. In this prediction, however, we do not consider the effects of terrain conditions of the cell.

Once a real-time user enters the departing region (Figure 3.2), the system initiates a procedure by which the same amount of average bandwidth requested in the requirement profile at call setup time, is attempted to be reserved in its destination cells to reduce the probability of reservation (hence hand-off) failure.

3.1.5. Reservation through Channel Borrowing

Referring to Figure 3.3, a departing user in cell $A$ has the set $D = \{D_1, D_2, D_3\}$ of three destination cells. If the user requested bandwidth is not available in any one of these destination cells, bandwidth is borrowed from one of their neighboring cells selected according to a channel borrowing function. We impose the restriction that a cell in $D$ cannot borrow channels from another cell in the same set or the current cell of the user.

To make the set of lenders for each destination cell uniform and disjoint (implying better load balancing among these cells), the set of possible lenders for the destination cell $D_i$ is restricted to the set of three cells marked as $B_i$, for each $i = 1, 2, 3$. Let us
call this the lender set of $D_i$. The selection of the lender cell from the lender set is determined as follows.

Let $T$ and $N_{bw}$ be the traffic load in Erlangs and the amount of available bandwidth respectively in a probable lender cell. We select that cell as the lender whose parameters maximize the value of the function

$$F = \frac{N_{bw}}{T}$$

(3.1)

Thus the objective is to select a cell with a large amount of available bandwidth and a low average traffic load.

3.1.6. Bandwidth Compaction

We assume that a high bandwidth application will be assigned contiguous multiple channels (e.g., time-slots, frequency band) by the system. This is usually the case for a real system employing multi-rate channels like GPRS. Based on the bandwidth requested in the requirement profile, a bandwidth segment is allocated to a real-time application at the upper end of the frequency spectrum. Similarly, bandwidth pages will be allocated to non-real-time applications successively from the rear end of the spectrum allocated to the cell (see Figure 3.4). The system will have the provision of
"removing" an active bandwidth page depending on the traffic load. This means that the amount of bandwidth allocated to non-real-time applications decreases because the scheduler schedules the non-real-time packets at a reduced rate. Bandwidth allocation in this fashion ensures that the bandwidth segments and pages form two disjoint sets for efficient utilization of the spectrum.

![Bandwidth Segments and Pages](image)

Figure 3.4: Bandwidth allocation pattern showing segments and pages

Since real-time users are assigned bandwidth segments, there can be unutilized bandwidth "holes" in the frequency spectrum caused by the call terminations. For example, in Figure 3.4(a), call terminations for users 2 and 4 lead to the creation of "holes". In a system requiring contiguous allocation of bandwidth for multimedia traffic, such a hole can only be filled again by a call request of equal or smaller bandwidth. We describe below a method called bandwidth compaction for efficient utilization of the available spectrum in such a system, which was originally suggested in [12, 35].

The idea behind bandwidth compaction is to shuffle the ongoing communications to place all free bandwidth together in one contiguous block. For example, the spec-
trum with holes as in Figure 3.4(a) becomes as in Figure 3.4(b) after compaction.

Moving a bandwidth segment to a different part in the frequency spectrum implies "re-tuning" the channels (constituting the segment) to a different set of channels. This might mean hopping to different carrier frequencies in FDMA systems or re-synchronization to different time slots in TDMA systems. The bandwidth compaction algorithm will sequentially shuffle the existing bandwidth segments towards the lower frequency side to absorb the holes, requiring the communicating users to "tune" to new channels almost instantaneously. This scheme can sometimes be very expensive and time consuming as a large number of channel re-assignments need to be done. Therefore, a complete bandwidth compaction is executed in the background when the traffic is stable.

A call request will generally not tolerate long delays to receive an acknowledgment. Usually, the user will hang-up and try again later. Hence, a partial bandwidth compaction algorithm will be executed during the call admission phase. The partial compaction scheme checks the spectrum allocation pattern to determine if by moving a few (in the range of 2 or 3) bandwidth segments to adjoining holes, sufficient contiguous spectrum can be released in order to satisfy the user request. This scheme only involves very few channel reassignments and sustains lower delays.

3.1.7. Call Admission Algorithm

The call admission algorithm is presented as a flowchart in Figure 3.5. As discussed earlier, the originator of the request specifies a requirement profile specifying its average and minimum bandwidth requirements, and whether it is a real-time or non-real-time application. The call admission criteria will be different for each class. For real-time users, admission is primarily based on the availability of bandwidth and compaction may have to be resorted to. The real-time user is classified as local or departing based on the location prediction scheme. If the user is departing, bandwidth reservation is initiated in the predicted destination cells. If these reservations are
successful, then only the call is admitted. The call of a local user is admitted based on the bandwidth availability in the current cell only.

For non-real-time users, the admission is primarily based on the availability of buffer space in the non-real-time packet queue (NRTQ). This criteria may be set against a queue length threshold, in order to prevent queue overflow once the new call is admitted. Once the non-real-time call is admitted, one or more bandwidth pages (if available) are assigned to transmit the packets depending on its requirement profile.

![Flowchart of the call admission algorithm](image)

Figure 3.5: Flowchart of the call admission algorithm

There is a QoS monitor function as part of the call admission controller, which will monitor QoS parameters affecting system-wide performance, e.g., interference level
in the cell, number of hand-off drops and bit error rate performance, and provide feedback to the call admission controller to help it make certain policy-based call admission decisions. These policies will be determined largely by network operators.

3.1.8. Bandwidth Stealing

In a fully loaded system (a sector/cell with all channels occupied), there is a provision of admitting non-real-time calls by "stealing" channel capacity from the real-time users. During the period of inactivity of the application source, the real-time user is not using the channels assigned to it, therefore, the channel can be used to transmit packets for non-real-time user. At high system loads, a channel can be owned by multiple users comprising of at most one real-time user and one or more non-real-time users. The maximum number of such users will depend on the policy enforced by the network operator. The real-time user will have true ownership of a channel, while the non-real-time users sharing the same channel will have a restricted ownership. This implies that a real-time user sharing a channel with another non-real-time user, will have preemptive priority over the latter and will continue packet transmission when the source is active again.

3.2. Performance Analysis

In this section, we will derive analytical models to evaluate the performance of our call admission scheme using predictive channel reservation in neighboring cells. Also, we evaluate expressions for the average carried traffic under call degradation, and the average size of the real-time and non-real-time packet queues.

3.2.1. Call Admission and Successful Hand-off for Real-time Calls

In this section we derive a QoS parameter for bandwidth reservation for real-time calls to be able to compare the performance of our scheme with that due to Oliveira.
et al. [30]. The parameter is the probability of admitting a call and its successful hand-off to one of the neighboring cells, assuming that the user changes cell at least once during the session. Since bandwidth reservation takes place only in case of real-time calls, the model is applicable to this class of calls only. We also assume for simplicity that all calls, on the average, require the same amount of bandwidth. Let $P_{av}$ be the stationary probability that this bandwidth is available in any cell. We define

$C \equiv$ Event denoting call admission,

$S \equiv$ Event denoting successful hand-off to a neighboring cell.

Our goal is to derive an expression for the probability, $P(CS)$.

Assuming that unavailability of bandwidth is the only reason behind a hand-off failure, we proceed to determine the conditional probability of a successful hand-off of a call to a given neighboring cell, given that the call is admitted. The algorithm in [30] admits a real-time call only if there is enough bandwidth to be reserved in the originating cell and its six neighbors. Thus, the conditional probability $P(S|C) = 1$. The call admission probability is $P(C) = P_{av}$, and the call blocking probability is given by $P(B) = 1 - P(C)$. Hence, for the scheme in [30],

$$P(CS) = P(C)P(S|C) = (P_{av})^7. \quad (3.2)$$

The admission policies in our scheme are different for a local and a departing user initiating the call request. For a local user, a call is admitted if there is enough bandwidth available in the cell, implying that $P(C)_{local} = P_{av}$. After admission, the probability of successful hand-off of the call to a given neighboring cell is $P(S|C)_{local} = P(BW_D)$, where $BW_D$ denotes the event that sufficient bandwidth (through borrowing, if required) is available in the set $D$ of destination cells for the user. For a departing user, the call is admitted if enough bandwidth is available in the cell and the set $D$ of its destination cells. Thus, $P(C)_{departing} = P_{av}P(BW_D)$. Once the call is admitted, the probability of successful hand-off is $P(S|C)_{departing} = 1.$
Thus,

\[ P(CS)_{\text{departing}} = P(C)_{\text{departing}} P(S|C)_{\text{departing}} = P_{av} P(BW_D) = P(C)_{\text{local}} P(S|C)_{\text{local}} = P(CS)_{\text{local}} \]  

(3.3)

implying that \( P(CS) \) is the same for both the local and departing users.

To derive \( P(BW_D) \), we consider separately the cases where bandwidth is available in all three, two, one and none of the three destination cells in the set D. Therefore,

\[ P(BW_D) = P_{av}^3 + 3P_{av}^2 (1 - P_{av})(3P_{av}) + 3P_{av}(1 - P_{av})^2 (3P_{av})^2 + (1 - P_{av})^3 (3P_{av})^3 \]

\[ = [P_{av} \{1 + 3(1 - P_{av})\}]^3 \]

The probability of call admission and a successful hand-off in our scheme is thus

\[ P(CS) = P_{av}^4 [1 + 3(1 - P_{av})]^3 \]  

(3.4)

To compare our scheme with that in \([30]\), we compute the ratio \( \eta \) of the two probabilities in Equations (3.4) and (3.2). It is given as

\[ \eta = \frac{P_{av}^4 [1 + 3(1 - P_{av})]^3}{P_{av}^4} = \left( \frac{4}{P_{av} - 3} \right)^3 \]  

(3.5)

Since \( P_{av} \leq 1 \), we set \( \eta \geq 1 \) which implies that the probability of call admission and successful hand-off in our proposed scheme is always greater than or equal to that obtained from the approach in \([30]\). The values of \( \eta \) as a function of \( P_{av} \) are plotted in Figure 3.6.

It is observed that with low probabilities of resource availability, our scheme shows a huge improvement over \([30]\). As the probability \( P_{av} \) increases, the performances of both the schemes tend to be similar, and finally, the ratio \( \eta \) converges to unity when \( P_{av} = 1 \).

3.2.2. Carried Traffic under Bandwidth Degradation

Recall that, the real-time traffic has preemptive priority over non-real-time traffic. Bandwidth degradation is applicable only for a fully loaded system, where all the
channels are occupied. Let us consider the case where all the channels are occupied by real-time users. A new real-time call can be admitted either if bandwidth degradation leads to the availability of sufficient channels or another user terminates a call. Assuming sufficient buffer space is available, the admitted non-real-time users “steal” bandwidth from real-time users who are inactive. Hence, the average number of non-real-time users in the system will depend on the number of such “idle” channels of real-time users. We will continue our analysis assuming a TDMA system where a channel is equivalent to a time slot in a frame.

Let $N$ be the total number of slots in a frame. Let the call arrival process for real-time users be Poisson with average rate $\lambda_r$, and the service time be exponentially distributed with rate $\frac{1}{\mu_r}$. Also, assume that the real-time users are using $N_r^{avg}$ slots on the average. We denote

$$p_i = \text{probability that user } i \text{ is degradable}$$

$$n_i = \text{number of slots by which user } i \text{ can degrade}$$

The average number of slots obtained by degrading the existing real-time users is
given as
\[ S_{\text{degrade}} = \sum_{i \notin S_i} p_i n_i \]  
(3.8)

where \( S_i \) is the set of real-time users using a single slot who are non-degradable. We also assume that a new user is never admitted into the system in a degraded mode [11]. The average carried real-time traffic per carrier is given as
\[ U_r = \frac{N + S_{\text{degrade}}}{N_{\text{avg}}} \]  
(3.9)

Note that the non-real-time users are allocated the "idle" slots of the real-time users. Let \( a_i \) denote the average activity of the \( i \)th user who is occupying \( N_i \) slots. Hence, the probability that a slot is idle is given as
\[ p_{\text{idle}} = \sum_{i=0}^{m} (1 - a_i) \frac{N_i}{N} \]  
(3.10)

where \( m \) is a random variable denoting the number of real-time users in the system. The average value of \( m \) in this case can be equated to \( U_r \) given by Equation (3.9).

The probability of \( k \) slots being idle is given by the Binomial distribution,
\[ \binom{N}{k} p_{\text{idle}}^k (1 - p_{\text{idle}})^{N-k} \]  
Hence, the average number of idle slots is \( N p_{\text{idle}} \). Let the average bandwidth requirement for non-real-time users be \( N_{\text{avg}}^{nr} \). Then, the total carried non-real-time traffic is given by
\[ U_{\text{nr}} = \frac{N p_{\text{idle}}}{N_{\text{avg}}^{nr}} \]  
(3.11)

Thus the total carried traffic in the system is given by \( U = U_r + U_{\text{nr}} \).

3.2.3. Average Queue Lengths for RTQ and NRTQ

For this analysis, we assume that the real-time users are admitted depending only upon the available bandwidth. Hence, the maximum number of real-time users who can be served is
\[ C_r = \frac{N}{N_r^{\text{avg}}} \]  
(3.12)
Let the average activity of real-time users be denoted by \( a_r \), while the arrival and service rates of real-time requests (both Poisson distributed) be denoted by \( \lambda_r \) and \( \mu_r \) respectively. Since the non-real-time users are accommodated by “stealing” bandwidth from inactive real-time users, the average number of non-real-time requests served is given by

\[
C_{nr} = \frac{N - a_r (\frac{\lambda_r}{\mu_r}) N^{avg}}{N^{avg}}
\]  

(3.13)

We model both classes with the help of \( M/M/C \) queuing model, where \( C \) is approximated by \( C_{nr} \) for non-real-time calls and as \( C_r \) for real-time calls.

**Average RTQ Length**

As per the \( M/M/C \) model, the probability that the queue length of the real-time packet queue is \( n \) is given as [26]:

\[
P_r^n = \frac{\lambda_r^n}{n! \mu_r^n} P_0, \text{ for } 1 \leq n \leq C_r
\]

\[
= \frac{\lambda_r^n}{C_r! \mu_r^n} P_0, \text{ for } n \geq C_r
\]

where

\[
P_r^n = \left[\sum_{i=0}^{C_r-1} \frac{1}{i!} \left(\frac{\lambda_r}{\mu_r}\right)^i + \frac{1}{C_r! \left(\frac{\lambda_r}{\mu_r}\right)^C_r} \left(\frac{C_r \mu_r}{C_r \mu_r - \lambda_r}\right)^{-1}\right]^{-1}.
\]

The length \( L_r^q \) of the queue of requests waiting to be served is then

\[
L_r^q = \sum_{n=C_r+1}^{\infty} (n - C_r) \ P_r^n
\]

\[
= \left[ \frac{(\lambda_r)^C_r \lambda_r \mu_r}{(C_r - 1)!(C_r \mu_r - \lambda_r)^2} \right] P_r^0.
\]

The waiting time in the queue is given by Little's formula [26], \( W_r^q = \frac{L_r^q}{\lambda_r} \).

Assuming that the traffic generated by an admitted call has a Poisson-distributed
packet burst size with an average rate \( \lambda_r \), the number of packets \( (B_r) \) that need to be buffered has a simple upper bound given by

\[
B_r \leq L_r^q \cdot W_r^q \cdot \lambda_r^p
\]  

(3.14)

Here, \( L_r^q \cdot \lambda_r^p \) gives the maximum packet arrival rate when all the admitted real-time users are simultaneously active.

**Average NRTQ Length**

For non-real-time traffic, let \( \lambda_{nr} \) be the request rate and \( \mu_{nr} \) be the service rate for each request, both Poisson distributed. Using the \( M/M/C \) queuing model with \( C \) approximated by \( C_{nr} \) given by Equation 3.13, the number of non-real-time packets, \( B_{nr} \), that need to be buffered can be bounded as

\[
B_{nr} \leq L_{nr}^q \cdot W_{nr}^q \cdot \lambda_{nr}^p
\]

(3.15)

where the parameters \( L_{nr}^q \) and \( W_{nr}^q \) are derived similar to the corresponding parameters for the real-time case.

3.3. Simulation Experiments

In this section, we provide details of our simulation model and the experimental results obtained.

3.3.1. Simulation Model

Our simulation is comprised of applications that generate requests and data packets, the admission controller (AC) that admits these requests, the packet sorter (PS) that sorts the data packets, the queue manager (QM) that keeps track of the length of the real-time and non-real-time queues, the compaction manager (CM) that performs compaction on a frame, and the scheduler that schedules packets onto a frame. Hand-off control and reservation messaging for hand-off calls are taken care by the call control manager (CCM).
3.3.2. Admission Control

User requests are modeled as requirement profile, which is a 3-tuple: average bandwidth, minimum bandwidth, and delay tolerance. Figure 3.7 shows the sequence of events that are initiated by the AC when a new request has to be admitted. Once an applications request has been admitted and data packets are received from the application, the PS appends a single degradation information bit (DIB) to convey degradation information to the scheduler according to a simple algorithm that checks if the minimum bandwidth is less than the average bandwidth. If so, we set DIB = 1, otherwise DIB = 0.

3.3.3. Frame Structure and Scheduling

Our scheduling mechanism is adapted for the FRAMES radio interface proposal [28]. In FRAMES mode-1 (called FMA1), a unit TDMA frame length is 4.615 msec and can consist of 1/64 (72 μs), and 1/8 (577 μs) slots fitting together in a TDMA frame.
We denote a 1/64 slot by $S_A$ and a 1/8 slot by $S_B$, where $S_A$ will map to an indivisible unit time slot and $S_B = 8S_A$ per frame. For a bit rate of 64 Kbps, either a single $S_B$ slot or eight $S_A$ slots can be used in a single radio link frame. For the purpose of link layer scheduling, we abstract that the slot $S_A$ can carry $K$ bytes of data and the slot $S_B$ can carry $8K$ bytes of data.

Our scheduling algorithm considers the time slots of a frame indexed by increasing time. A unit frame is represented as $S_{t_0}, S_{t_1}, \ldots, S_{t_n}$. Two indices are maintained while populating a frame. These indices indicate the last time slot index ($S_{t_{\text{min}}}$) occupied before the “hole”, and the first slot index ($S_{t_{\text{max}}}$) occupied at the end of the “hole”, respectively.

The link layer scheduling algorithm is implemented as follows.

**Algorithm** Scheduler

WHILE (Real-time queue (RTQ) not empty) DO

Dequeue a packet from RTQ

\[ p = \lceil (\text{packet size in bytes} / K) \rceil \]

\[ q = \lceil (\text{packet size in bytes} / nK) \rceil \]

CASE (packet's DIB)

0: /* not degradable */

IF \((S_{t_{\text{min}}} + (q \times S_B)) < S_{t_{\text{max}}})\) THEN

Assign $q$ $S_B$ slots at $S_{t_{\text{min}}}$

\[ S_{t_{\text{min}}} = S_{t_{\text{min}}} + (q \times S_B) \]

ELSE /* cannot assign RTQ packet */

BREAK; /* exit RTQ loop and schedule NRTQ packets */

ENDIF

1: /* degradable */

IF \((S_{t_{\text{max}}} - (p \times S_A)) > S_{t_{\text{min}}})\) THEN

Assign $p$ $S_A$ slots at $S_{t_{\text{max}}}$

\[ S_{t_{\text{max}}} = S_{t_{\text{max}}} - (p \times S_A) \]
ELSE /* frame is full */
EXIT;
ENDIF
ENDCASE
ENDWHILE

WHILE (Non-real-time queue not empty) DO
IF 
((S_{t_{max}} - (p \times S_A)) > S_{t_{min}}) THEN
Assign p S_A slots at S_{t_{max}}
S_{t_{max}} = S_{t_{max}} - (p \times S_A)
ELSE /* frame is full */
EXIT;
ENDIF
ENDWHILE
END ALGORITHM Scheduler

Table 3.1 shows a random mix of traffic packets and Figure 3.8 shows how the packets are mapped to a FMA1 (mode-1) frame. We have assumed that K = 8 bytes.

<table>
<thead>
<tr>
<th>Packet Number</th>
<th>Packet size (bytes)</th>
<th>Packet source</th>
<th>DIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100</td>
<td>RTQ</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>170</td>
<td>RTQ</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>RTQ</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>NRTQ</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>NRTQ</td>
<td>-</td>
</tr>
</tbody>
</table>
3.3.4. Compaction Implementation

The compaction algorithm checks for idle slots (from terminated calls) and reassigns idle slots to ongoing calls. We show below the implementation for compacting the lower end of a frame. Compaction on the upper end of the frame is carried out in a similar loop.

**Algorithm Compactor**

FOR $(i = 0$ TO $St_{min}$) DO

IF (CALLSTATUS of $St_i$ is TERMINATED) THEN

FOR $(j = (St_i + q$ TO $St_{min})$ DO

Assign $\text{CONTENTS}(St_i) = \text{CONTENTS}(St_j)$

$St_i = St_i + 1$

ENDFOR

$St_{min} = St_{min} - q$

ENDIF

ENDFOR

END ALGORITHM Compactor
3.3.5. Simulation Parameters

1. **Data generation**: Admitted requests act as bursty sources of data, with inter-arrival time of the bursts being exponentially distributed with mean \( \frac{1}{\gamma} \).

2. **Request origination and termination**: Call arrival is programmed as a Poisson process with inter-arrival time exponentially distributed with mean \( \frac{1}{\lambda} \). The call holding time is also programmed as a exponentially distributed random variable with mean \( \frac{1}{\mu} \).

3. **Modeling of mobility**: Number of hand-off requests is programmed as an exponentially distributed random variable with mean \( \frac{1}{\lambda^h} \). No differentiation is made between real-time and non-real-time users for hand-off calls.

4. **User classification**: In order to classify the users correctly, we need to model the signal strength received by the base station from the user. Since actual signal strengths cannot be generated, we overlay on each cell a mesh. A user position within a cell is given by a pair of co-ordinates \((x, y)\) on this mesh. A fixed set of co-ordinates defines the peripheral shaded region of a cell shown in Figure 3.2. User mobility is modeled as a random walk on the mesh.

3.3.6. Simulation Results

The simulation results are shown in Figures 3.9-3.13. From Figure 3.9, we observe that the system capacity is fully utilized only when the compaction technique is used. With arrival rate less than the service rate, the call admission probability is approximately unity when compaction is used. Without compaction, some incoming requests are blocked due to insufficient bandwidth in the left-over "holes". The observed improvement varies from 21% for an arrival rate of .03 to about 4% for an arrival rate of 0.07.

Figure 3.10 shows an improvement of call admission probability of non-real-time calls with compaction over that without compaction. The percentage improvement
Figure 3.9: Call admission probability for real-time users versus call arrival rate for queue size = 100 and service rate $\mu = 0.03$

varies from 4% at an arrival rate of 0.05 to about 17% at an arrival rate of 0.09. The improvement is greater at higher call arrival rates due to the fact that at high arrival rates, most of the incoming real-time calls are blocked providing scope of scheduling the buffered non-real-time calls.

Figure 3.11 demonstrates the call admission probability with varying queue sizes with and without compaction. The admission probability for non-real-time users is observed to be higher with compaction at small queue lengths over that without compaction. Note that, at some threshold queue size, the benefit of compaction is lost. This implies that, given an average call arrival and service rates, there exists a queue size beyond which the compaction algorithm should not be executed, as compaction is generally very computation intensive.

Figure 3.12 shows that as the service rate is increasing, the real-time calls are terminated quickly leading to an increase of higher priority real-time calls in the system. This reduces the admission probability of the non-real-time calls in the
Figure 3.10: Call admission probability for non-real-time users versus call arrival rate for queue size = 100 and service rate $\mu = 0.03$

system.

Figure 3.13 shows that the hand-off dropping probability also decreases with compaction. The observed improvement is around 17% for a hand-off rate of .001 and 11% for a rate of .009.

3.3.7. Comparison of Reservation Scheme

Figure 3.14 shows the comparison in terms of the probability of admission and successful hand-off of a call with that given by the scheme proposed in [30]. As expected from the analytical results in previous sections, the $P(CS)$'s for both the schemes tend to merge for high values of $\lambda$. An improvement of about 12% is observed for $\lambda = 0.9$. 
Figure 3.11: Call admission probability for non-real-time users versus queue size for arrival rate $\lambda = 0.08$ and service rate $\mu = 0.03$

Figure 3.12: Call admission probability for non-real-time users versus call arrival rate for queue size $= 100$
Figure 3.13: Hand-off drop probability versus hand-off rate for queue size = 100, service rate $\mu = 0.01$ and call arrival rate $\lambda = 0.1$

Figure 3.14: Call admission and successful hand-off probability versus traffic load
CHAPTER 4

A Reservation Mechanism for Mobile Nodes in the Internet

In this chapter, we address the problem of resource reservation at the Network layer for mobile nodes. We propose a simple and scalable resource reservation mechanism that allocates resources to a currently ongoing data flow of the mobile node in all neighboring subnets. The resource reservation setup and maintenance is achieved through two simple protocols, The Neighbor Mobility Agent Discovery Protocol (NMADP), and the Mobile Reservation Update Protocol (MRUP). Our reservation concepts can be incorporated with ease into current Internet routing such as the Open Shortest Path First protocol (OSPF) [24] and Mobile IP [32, 20]. The reservation protocols are adaptable with both intserv and diffserv architectures since the scope of our reservation is restricted to a leaf subnet. For intserv, our protocols can work with MRSVP, by providing the list of foreign agents to MRSVP so that passive reservations can be setup. In case of a diffserv architecture, the reservation is made by NMADP and MRUP.

4.1. Internet Architecture

An important assumption for the effective working of NMADP and MRUP is that during the lifetime of a connection, the mobile node will move from a subnet to a geographically adjacent subnet. This is a very reasonable assumption since a user cannot hop from one subnet to another subnet a large distance away with an open connection. Accordingly, the Internet architecture we consider is as shown in Figure 4.1. A geographic area has to be covered by several wireless local area networks (WLANs) or some other similar wireless networking technology, for a mobile node to truly communicate on the move. This architecture is very similar to the current archi-
tecture of the *Public Land Mobile Telephony System* (PLMTS), where a geographical area is partitioned into *cells*. Analogous to cells are subnets in our architecture.

![Subnet Architecture Diagram](image)

**Figure 4.1: Subnet Architecture**

4.1.1. Scope of Resource Reservation

The physical medium of communication between a mobile node and the network is via a wireless interface such as radio or infrared. There are several issues to be addressed for successful communication at the physical layer. For example, framing, coding, synchronization, and scheduling or access issues. In this chapter, by resource reservation, we mean the following:

1. There is enough buffer (memory) space and processor capacity reserved at the
gateway router to a subnet.

2. There are enough wireless resources reserved. For example, if the physical medium is radio and the multiple-access technique is TDMA, the packets of a reserved flow are guaranteed to receive their required number of time slots. Mechanisms for propagation of these reservations from the network layer to the lower layers are out of the scope of this chapter.

4.2. Neighbor Mobility Agent Discovery Protocol (NMADP)

Mobile IP defines two functional entities to handle mobility management for mobile nodes in the Internet [32]. A Home Agent (HA) is a router on a mobile node's home network that maintains current location information for the mobile node, and tunnels datagrams for delivery to the mobile node when it is away from home. A Foreign Agent (FA) is a router on a mobile node's visited network which provides routing services to the mobile node while registered. We use the term Mobility Agent (MA) to refer to an HA or an FA.

The purpose of NMADP is for mobility agents of neighboring subnets to discover each other. Currently, there is no mechanism in the literature that allows two neighboring routers to discover that they are both mobility agents for their respective subnets. In the case of normal IP routing, the most common method of neighboring routers to “know of” each other is by manual configuration at the time the routers are brought up. Assuming that two neighboring routers “know of” each other's presence, NMADP allows them both to communicate their mobility-support capabilities.

4.2.1. Direct Method

1. Each MA (a router that is a home or a foreign agent) periodically sends a Neighbor Mobility Agent (NMA) message to its neighboring routers with the IP address of the interface on which it is an MA.
2. If the receiver router of a *Neighbor Mobility Agent* (NMA) message is also an MA, the receiver-MA records the sender-MA as being mobility handling router in its routing table, and returns an acknowledgment back to the sender-MA indicating its own mobility handling capability. Otherwise, the sender-MA’s *Neighbor Mobility Agent* message is discarded.

The NMA message structure is shown in Figure 4.2(a). As an example, consider that router R4 in Figure 4.1 is an MA. The NMA message generated by R4 and sent to router R3 is shown in Figure 4.2(b). For simplicity, we have refrained from showing all the interface addresses of R4 and R3. A single bit flag $M$ could be added to the routing table to maintain a routers mobility handling capability information. Continuing the example, Figure 4.2(c) shows the difference in the routing table of R4 before and after the *Neighbor Mobility Agent* discovery message has been sent from R4 to R3 and an acknowledgment from R3 has been received. The other flags $U$ indicates if the router is up, and $G$ indicates if the router is a gateway router.

<table>
<thead>
<tr>
<th>Destination Router IP address</th>
<th>e.f.g.h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Router Address [1]</td>
<td>MA flag 1</td>
</tr>
<tr>
<td>a.b.c.d</td>
<td>1</td>
</tr>
<tr>
<td>p.q.r.s</td>
<td>0</td>
</tr>
</tbody>
</table>

(a)

<table>
<thead>
<tr>
<th>Dest.</th>
<th>Gateway</th>
<th>Flags</th>
</tr>
</thead>
<tbody>
<tr>
<td>e.f.g.h</td>
<td>a.b.c.d</td>
<td>UG</td>
</tr>
<tr>
<td>e.f.g.h</td>
<td>a.b.c.d</td>
<td>UGM</td>
</tr>
</tbody>
</table>

(b)

Figure 4.2: NFA Discovery Message and Routing Table
4.2.2. Indirect Method

In many situations two neighboring routers may not "know" of each other's presence, if they do not have a direct communication link between them. We call this problem the Hidden-Router problem. The Indirect Method or Mobile Assisted method is proposed to handle the mutual discovery of two neighboring MA's, both of which are hidden from each other. In Figure 4.1, assume routers R3 and R5 are both MA's. Although R3 and R5 are both gateways to wireless LANs, and their wireless LANs are adjacent to each other, R3 and R5 do not know of each other's mobility handling capabilities since they are not connected to each other directly and hence the direct method will fail. The scenario could be worse if a large internet is the only connection between R3 and R5.

We make use of Mobile Registration for NMA discovery, when two neighboring routers are hidden from each other. Mobile Registration is a mechanism for mobile nodes to communicate their current reachability information to their HA [32]. Registration has to be performed by every mobile node as soon as it enters a foreign network. It is only after a mobile node has registered, the HA can tunnel packets to the mobile node. Registration is performed by the exchange of two messages registration request and registration reply. When a mobile node enters a foreign network, it sends a registration request to the FA. The FA forwards the request to the mobile node's HA (the FA obtains the HA's address from the mobile node). The HA authenticates the mobile node, and sends back a registration reply to the FA. Thereupon, the FA provides services to the mobile node. The NMA can be performed in either distributed or centralized manner.

Distributed Discovery

In distributed NMA discovery, every MA discovers a hidden neighbor-MA when the MA is playing the role of an HA. Assume a mobile node is in its home network and has an open connection with a sender node through the HA. When the mobile node
wanders into a foreign subnet and sends a registration request through the foreign subnet’s FA, the registration request is received at the HA. The HA processes the registration request, and infers that since the mobile node had an open connection, it has moved into a neighboring subnet. Had the HA known that the neighboring subnet and its FA existed, it would have initiated a reservation for the mobile node at the neighboring FA, and the mobile node’s connection would not have been dropped. After processing the registration request and returning it to the FA, the HA sends an NMA message to the FA through an intermediate router, and both the HA and the FA discover that they are neighbors. The drawback of this approach is that the mobile node loses its connection the first time it roams into an “undiscovered” neighboring subnet. The neighbor mobility agent information matures over a certain period of time.
Centralized Discovery

Centralized NMA discovery is a variant of the distributed NMA discovery, in which the HA of a mobile node can inform two FA's that they are neighbors.

We illustrate the operation of both the distributed and centralized schemes with the help of an example. Consider a network as shown in Figure 4.3. The routers HA and FA1, and FA1 and FA2, are MAs for geographically adjacent subnetworks, but they do not have a direct communication link between them. HA and FA1 are physically wired through intermediate routers IR1 and IR2, and FA1 and FA2 are connected via IR3.

Suppose that a mobile node is registered with the HA, and has an open connection to a sender host somewhere in the Internet, through the HA. Next, if the mobile node roams and registers with FA1 then, through distributed NMA discovery, HA and FA1 will discover each other. As mentioned earlier, a mobile node's connection may be dropped if enough resources were not available at FA1.

Suppose that at a later time, a mobile node again started with an open connection at the HA. This time however, the HA would have reserved resources at FA1, and when the mobile moves into FA1's subnet, its connection will not be dropped. Now, if the mobile node continued on the roaming path and registered with FA2, its connection would be dropped again, since FA1 and FA2 have not discovered each other. The HA now infers that the previous location of the mobile node was FA1 and the new location is FA2, which means that FA1 and FA2 are hidden-neighbors. The HA can direct either FA1 or FA2, to initiate an NMA discovery message towards FA2 or FA1 respectively.

4.3. Mobile Reservation Update Protocol (MRUP)

The purpose of MRUP is to guarantee resource reservation for a mobile node throughout the lifetime of an open data flow, when the mobile node is roaming from one subnetwork to another subnetwork.
We define resources (bandwidth, processor capacity, or memory space) to transition among three distinct states—free, reserved, and in-use. Resource reservation means that the state of a resource is changed from free to reserved. When the resources are actually used, the state of the resource changes from reserved to in-use. As an example, consider a TDMA frame comprised of a number of time slots. When no data is being scheduled in a time slot of the frame, the slot is free. When the slot is reserved, the slot is not available for scheduling, and when the slot is in-use, data is being sent in the slot. Note that the state of a resource can change from free to in-use directly. We can also view the state reserved as a subset of free, with the constraint that when the application that set the resource to reserved requires the resource, the application's request should not be denied. This implies that other applications can view a reserved resource as free until the reserving application changes the state of the resource from reserved to in-use.

MRUP operates in the following steps:

1. When a mobile node requests a guaranteed data flow connection with an MA, the MA checks if enough free resources are available in its subnetwork. If so, the requested resources are set to in-use and the request is accepted. Otherwise, the mobile node's request is denied.

2. If a request is accepted, the MA sends a Reserve Resource message to all its neighboring MA's (as discovered by NMADP). The Reserve Resource contains the amount of resources requested by the mobile.

3. Once an MA has initiated a reservation for a mobile node, it periodically sends Reservation Refresh messages to its neighboring MA's, until it realizes that the mobile node is no longer present in its subnetwork through the link or lower layers.

4. An MA, upon reception of a Reserve Resource message, changes the state of a certain amount of resources from free to reserved. The amount of resources
for which the state change is made, is equal to or less than the amount of resources requested by the mobile node.

5. When an MA receives a registration request from a mobile node, it checks if there is a resource reservation for the mobile node. If so, the MA changes the state of the resources from reserved to in-use. The MA also initiates a reservation as in step 2.

<table>
<thead>
<tr>
<th>Destination Router IP address</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reservation time-to-live</td>
</tr>
<tr>
<td>Mobile Node IP address</td>
</tr>
<tr>
<td>Wireless Link Bandwidth</td>
</tr>
<tr>
<td>Delay Requirement</td>
</tr>
<tr>
<td>Buffer Requirement</td>
</tr>
</tbody>
</table>

Figure 4.4: Format of Reserve Resource message

The format of reserve resource message is shown in Figure 4.4. The destination router IP address is the IP address of the receiving MA. The reservation time-to-live field denotes the time limit until which this Reserve Resource message is valid. This can be implemented as a counter that is decremented by the receiver-MA every minute. The value of the field is an engineering parameter. The IP address of the mobile node for which the reservation is being made is included. The requested bandwidth, delay constraints, and buffer requirements are also included. These may be obtained from the flow profile when the mobile node first starts the data flow. Other wireless link related parameters may be added as needed. The Reservation Refresh message is comprised of only the first three fields.
4.4. Analytical Model for Resource Reservation

In the following, we use the terms mobility agent and subnet interchangeably. We assume a 1:1 correspondence, i.e. each subnet has a single mobility agent.

According to our resource reservation protocol, if there is a reservation request in a subnet, we allocate the requested resources at the current MA, as well as the MAs of all neighboring subnets that are one hop away. For analysis purposes, we consider the resource to be bandwidth on the air interface between the MA and the mobile node. For a user, the bandwidth allocated in the current MA is called in-use bandwidth, and the bandwidth allocated in the adjacent subnets is called reserved bandwidth. In a subnet, users using bandwidth are called in-use users, and users for whom there is a reservation are called reserved users.

For analysis, we assume that each subnet has a fixed number of channels denoted by $N$, and each call request demands a fixed number of channels denoted by $k$. Hence the number of users who can be accommodated in each subnet is denoted as $N' = \frac{N}{k}$. We define the state of each MA in a subnet as a tuple $(i, j)$. Here $i$ denotes the number of in-use users and $j$ denotes the number reserved users. Since there is a fixed number of users who can be served simultaneously, an MA can be in state $(i, j)$ under the following three conditions:

- $C1 : 0 \leq i \leq N'$
- $C2 : 0 \leq j \leq N'$
- $C3 : 0 \leq i + j \leq N'$

Consider a MA $B$ in state $(i, j)$. Then the following cases arise:

1. If there is a call request in the MA $B$ from a user in the subnet, the new state of the MA is $(i + 1, j)$. Let this event occur at rate $\lambda_{use}$.

2. If there is a call request at a MA one hop away from $B$, or if a user moves into a subnet which is one hop away from $B$, the new state of the MA $B$ is $(i, j + 1)$. Let this event occur at rate $\lambda_{resv}$. 
3. If there is a call termination by a mobile user at MA B, the new state of B is \((i - 1, j)\). Let this event occur at rate \(\mu_{\text{use}}\).

4. If there is a call termination by a mobile user at an MA that is one hop away from B, or if the user moves from B to another non-adjacent MA, then the state of B is \((i, j - 1)\). Let this event occur at rate \(\mu_{\text{resv}}\).

5. If a mobile user moves from B to an adjacent MA, the new state of B is \((i - 1, j + 1)\). Let this event occur at rate \(\alpha_{\text{resv}}\).

6. If a user moves from an MA, which is one hop away from B, to B the state of the MA B is \((i + 1, j - 1)\). Let this event occur at rate \(\alpha_{\text{use}}\).

7. If there is a call request simultaneously from a user at B, and a new user moves to an MA adjacent to B, the state of B is \((i + 1, j + 1)\). Let this event occur at rate \(\lambda_{\text{use}+\text{resv}}\).

8. If there is a call termination at B, and if a user at an adjacent MA B moves to a new non-adjacent (to B) MA, the state of the MA B is \((i - 1, j - 1)\). Let this event occur at rate \(\mu_{\text{use}+\text{resv}}\).

Note that the above state transitions are valid as long as the new state is valid, otherwise the mobility agent B remains in the same state.

Figure 4.5 shows a markov model for the transition between the states of a mobility agent in terms of the number of in-use and reserved users. In this figure,

\[ a = \lambda_{\text{resv}} \]
\[ b = \mu_{\text{resv}} \]
\[ c = \lambda_{\text{use}} \]
\[ d = \mu_{\text{use}} \]
\[ e = \lambda_{\text{use}+\text{resv}} \]
\[ f = \mu_{\text{use}+\text{resv}} \]
\[ g = \alpha_{\text{resv}} \]
Figure 4.5: Markov Model for State of a Mobility Agent
\( h = \alpha_{use} \)

In the following analysis, we use the summations below:

\[
sum = \lambda_{resv} + \mu_{resv} + \lambda_{use} + \mu_{use} + \lambda_{use+resv} + \mu_{use+resv} + \alpha_{resv} + \alpha_{use}
\]

\[
sum_1 = \lambda_{resv} + \mu_{resv} + \lambda_{use} + \mu_{use} + \alpha_{use} + \alpha_{resv} + \mu_{use+resv}
\]

\[
sum_2 = \mu_{resv} + \mu_{use} + \alpha_{use} + \alpha_{resv} + \mu_{use+resv}
\]

\[
sum_3 = \lambda_{use} + \mu_{resv} + \alpha_{use} + \lambda_{use+resv}
\]

\[
sum_4 = \mu_{use} + \lambda_{resv} + \alpha_{resv} + \lambda_{use+resv}
\]

\[
sum_5 = \mu_{use} + \lambda_{use} + \lambda_{resv} + \alpha_{resv}
\]

\[
sum_6 = \mu_{use} + \alpha_{resv}
\]

\[
sum_7 = \lambda_{use} + \lambda_{resv} + \lambda_{use+resv}
\]

\[
sum_8 = \lambda_{use} + \mu_{resv} + \lambda_{resv} + \alpha_{use}
\]

\[
sum_9 = \mu_{resv} + \mu_{use+resv}
\]

The Markov model in Figure 4.5 is formulated below, where the probability of being in state \((i,j)\) is denoted by \(P_{(i,j)}\).

FOR \(i \leftarrow 1\) TO \(N-2\)

FOR \(j \leftarrow 1\) TO \(N - i - 2\)

\[
P_{(i,j)} \times \text{sum} = P_{(i-1,j)} \times \lambda_{use} + P_{(i+1,j)} \times \mu_{use} + P_{(i,j-1)} \times \lambda_{resv} +
\]

\[
P_{(i,j+1)} \times \mu_{resv} + P_{(i+1,j-1)} \times \alpha_{resv} + P_{(i,j-1,j+1)} \times \alpha_{use} +
\]

\[
P_{(i-1,j-1)} \times \lambda_{use+resv} + P_{(i,j+1)} \times \mu_{use+resv}
\]

FOR \(i \leftarrow 1\) TO \(N-1\)

FOR \(j \leftarrow 1\) TO \(N - i - 1\)

\[
P_{(i,j)} \times \text{sum}_1 = P_{(i-1,j)} \times \lambda_{use} + P_{(i+1,j)} \times \mu_{use} + P_{(i,j-1)} \times \lambda_{resv} +
\]

\[
P_{(i,j+1)} \times \mu_{resv} + P_{(i-1,j-1)} \times \alpha_{resv} + P_{(i,i,j+1)} \times \alpha_{use} +
\]

\[
P_{(i-1,j-1,j+1)} \times \lambda_{use+resv}
\]

FOR \(i \leftarrow 1\) TO \(N-1\)

FOR \(j \leftarrow 1\) TO \(N - i\)
\[ P_{(i,j)} \times \text{sum2} = P_{(i-1,j)} \times \lambda_{\text{use}} + P_{(i,j-1)} \times \lambda_{\text{resv}} + P_{(i+1,j-1)} \times \alpha_{\text{resv}} + \\
\quad P_{(i-1,j+1)} \times \alpha_{\text{use}} + P_{(i-1,j-1)} \times \lambda_{\text{use}+\text{resv}} \]

FOR \( j \leftarrow 1 \) TO \( \mathcal{N} - 2 \)

\[ i \leftarrow 1 \]

\[ P_{(i,j)} \times \text{sum3} = P_{(i+1,j)} \times \mu_{\text{use}} + P_{(i,j-1)} \times \lambda_{\text{resv}} + P_{(i,j+1)} \times \mu_{\text{resv}} + \\
\quad P_{(i+1,j-1)} \times \alpha_{\text{resv}} + P_{(i+1,j+1)} \times \mu_{\text{use}+\text{resv}} \]

FOR \( i \leftarrow 1 \) TO \( \mathcal{N} - 2 \)

\[ j \leftarrow 0 \]

\[ P_{(i,j)} \times \text{sum4} = P_{(i-1,j)} \times \lambda_{\text{use}} + P_{(i+1,j)} \times \mu_{\text{use}} + P_{(i,j+1)} \times \mu_{\text{resv}} + P_{(i-1,j+1)} \times \alpha_{\text{use}} \]

\[ i \leftarrow \mathcal{N} - 1 \]

\[ j \leftarrow 0 \]

\[ P_{(i,j)} \times \text{sum5} = P_{(i-1,j)} \times \lambda_{\text{use}} + P_{(i,j+1)} \times \alpha_{\text{use}} \]

\[ i \leftarrow 0 \]

\[ j \leftarrow 0 \]

\[ P_{(i,j)} \times \text{sum6} = P_{(i+1,j)} \times \mu_{\text{use}} + P_{(i,j-1)} \times \mu_{\text{resv}} + P_{(i+1,j+1)} \times \mu_{\text{use}+\text{resv}} \]

\[ i \leftarrow 0 \]

\[ j \leftarrow \mathcal{N} - 1 \]

\[ P_{(i,j)} \times \text{sum7} = P_{(i+1,j)} \times \mu_{\text{use}} + P_{(i,j-1)} \times \lambda_{\text{resv}} + P_{(i,j+1)} \times \mu_{\text{resv}} + P_{(i+1,j-1)} \times \alpha_{\text{resv}} \]

\[ i \leftarrow 0 \]

\[ j \leftarrow \mathcal{N} \]

\[ P_{(i,j)} \times \text{sum8} = P_{(i,j)} \times \lambda_{\text{resv}} + P_{(i-1,j-1)} \times \lambda_{\text{use}+\text{resv}} \]

By solving the set of linear equations for the given set of traffic conditions we derive the value of \( P_{(i,j)} \). Once the probability of being in state \((i,j)\) is known, the
expected number of in-use users and reserved users are estimated as

\[ E_{use} = \sum_{i=1}^{M} \sum_{j=0}^{M-i} P(i,j) \]  
(4.1)

\[ E_{res} = \sum_{i=0}^{M} \sum_{j=1}^{M-i} j \cdot P(i,j) \]  
(4.2)

4.5. Simulation Model for Reservation Mechanism

4.5.1. Assumptions and Model

To simulate the behavior of the reservation mechanism and obtain performance parameters, we model the arbitrarily shaped subnets as hexagonal cells. We assume each subnet is serviced by a single MA, and this correspond to a base station serving a cell. In a single cell, we assume a new call request event and a hand-off into the cell event are mutually exclusive. This assumption forces the parameter in the analytical model \( \lambda_{use+reser} \) to be set to zero. Along similar lines, we assume there can be no in-use call termination and reserved call termination simultaneously. This forces the parameter \( \mu_{use+reser} \) to be set to zero. In simulation experiments we fixed the bandwidth available in each cell to 50 channels, and no call is allowed to request more than 8 channels. Simulation was carried out for one cell. For the simulated cell and the six adjacent cells, there can be at most one call request and one call hand-off in each simulation cycle. The simulation model assumes that the type of a new call can be real-time or non-real-time with equal probability. Since we assumed an hexagonal cell, the ratio of the probability of call arrival, service, and hand-off rate in the current cell to the corresponding probabilities in the neighboring cells is 1:6.

4.5.2. Experiments and Results

Our primary parameter of interest is the Call dropping probability, with and without reservation. Since we simulate a single cell, hand-off calls are treated as a subset of
new call requests. Unsuccessful hand-off of non-real-time calls from the current cell to an adjacent cell does not contribute to the total number of dropped calls. The call arrival rate, service rate and hand-off rate have been scaled by a factor of 10 in Figure 4.6.

![Figure 4.6: Arrival rate vs. Call dropping probability for service rate = 0.1](image)

As expected the call dropping probability increases with an increase in the call arrival rate in both the reservation and non-reservation cases. Observing the trend of call dropping probability with the hand-off rate for the reservation mechanism, we see that as the hand-off rate increases, the call dropping rate decreases. This is because we are pre-empting the non-real-time calls when a reserved user becomes in-use user, and we admit more non-real-time calls users once the in-use user once again becomes reserved user. Observing the same statistics for the non-reservation scheme, the call dropping probability increases as we increase the hand-off rate. This is because we are not preemitting non-real-time calls.
The increase in the call dropping probability between the reservation and non-reservation cases is because we are allowing non-real-time call requests to use the bandwidth reserved for reserved users. This means more number of real-time call requests will be rejected. Although a marginal loss in the number of new call requests is seen, a significant gain in the number of hand-off calls dropped is observed.
CHAPTER 5

Conclusions

In this chapter, we review the contributions made by this thesis and provide directions for future work.

5.1. Research Contributions

Next generation wireless applications are widely believed to be multimedia in nature. From a delay in information delivery from source to destination point of view, multimedia traffic is classified as real-time (delay-sensitive) and non-real-time (delay-tolerant). Support of these two classes requires different levels of response from the wireless network. This differentiation in traffic type formed the foundation for the quality-of-service framework developed in this thesis. The dynamic, error-prone behavior of the wireless physical link necessitates low level QoS control, and hence the framework focused on low layer (e.g., radio link layer) QoS provisioning.

The framework integrated our approach to known aspects of QoS provisioning such as bandwidth reservation, bandwidth degradation, and introduced a technique of improving spectrum utilization called bandwidth compaction. A call admission algorithm that admits requests based on the request class and existing traffic conditions at the link layer was proposed. Algorithms for accurate bandwidth reservation by predicting the user location within a cell were also proposed. Sorting and Scheduling algorithms to aid in the differential treatment were developed and implemented.

The performance of the QoS provisioning framework was captured through analytical models and simulation experiments. Analytical models compared our reservation and call admission schemes with an existing scheme in the literature. Expressions
for the carried traffic under bandwidth degradation, and estimates for the average lengths of the real-time and non-real-time queues were derived.

Simulation experiments showed an improvement of about 12% using our channel reservation approach. Simulation experiments showed upto 21% improvement in call admission probability of real-time calls and upto 17% improvement in admission probability of non-real-time calls, when the various call control techniques proposed within this framework, were employed. Threshold queue sizes that maximize the benefit of bandwidth compaction and scheduling were determined.

The Internet is increasingly being used for real-time communication of voice and data. Due to the "best-effort" delivery mechanism employed in the Internet, support of real-time applications requires resource reservation. As computers become lighter, and wireless connectivity grows, mobile computers will connect to the Internet.

The second major contribution of this thesis has been the development of a resource reservation protocol for mobile computers in the Internet. The fundamental idea behind the reservation mechanism is to reserve resources for a mobile computer, not only in the subnet it currently resides in, but also in all adjacent subnets.

The Neighbor Mobile Agent Discovery Protocol was developed for a router to discover neighboring routers that are capable of supporting mobile computers. The Mobile Reservation Update Protocol creates and maintains reservations for a mobile computer. The merit of the protocols is in their simplicity of implementation, effectiveness in achieving the reservation with reasonable overhead, and scalability. The protocols can be easily incorporated into the current Internet infrastructure as extensions to OSPF and Mobile IP. The reservation mechanism can be adapted to both intserv and diffserv architectures since we restricted the scope of our reservation mechanism to a leaf subnet.
5.2. Future Work

It would be interesting to implement some of the components of the QoS framework in a real world system such as IS-136 TDMA, or IS-99 CDMA, and observe the performance improvements.

The reservation protocols can be integrated into Mobile IP and OSPF, either on a network simulator [29], or on a prototype testbed.
REFERENCES


http://www-mash.cs.berkeley.edu/ns/.


