Summer 2004

QoS Provisioning for Multi-Class Traffic in Wireless Networks

Mona El-Kadi Rizvi
Old Dominion University

QOS PROVISIONING FOR MULTI-CLASS TRAFFIC IN WIRELESS NETWORKS

by

Mona El-Kadi Rizvi
B.S. Computer Science, May 1985, Old Dominion University

A Dissertation Submitted to the Faculty of
Old Dominion University in Partial Fulfillment of the
Requirement for the Degree of

DOCTOR OF PHILOSOPHY

COMPUTER SCIENCE

OLD DOMINION UNIVERSITY
August 2004

Approved by:

Stephan Olariu (Co-Director)

Hussein Abdel-Wahab (Co-Director)

Ravi Mukkamala

Larry Wilson

Min Song

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

ABSTRACT

QOS PROVISIONING FOR MULTI-CLASS TRAFFIC IN WIRELESS NETWORKS

Mona El-Kadi Rizvi
Old Dominion University, 2004
Co-Director: Dr. Stephan Olariu
Co-Director: Dr. Hussein Abdel-Wahab

Physical constraints, bandwidth constraints and host mobility all contribute to the difficulty of providing Quality of Service (QoS) guarantees in wireless networks. There is a growing demand for wireless networks to support all the services that are available on wired networks. These diverse services, such as email, instant messaging, web browsing, video conferencing, telephony and paging all place different demands on the network, making QoS provisioning for wireless networks that carry multiple classes of traffic a complex problem. We have developed a set of admission control and resource reservation schemes for QoS provisioning in multi-class wireless networks.

We present three variations of a novel resource borrowing scheme for cellular networks that exploits the ability of some multimedia applications to adapt to transient fluctuations in the supplied resources. The first of the schemes is shown to be proportionally fair; the second scheme is max-min fair. The third scheme for cellular networks uses knowledge about the relationship between streams that together comprise a **multimedia session** in order to further improve performance. We also present a predictive resource reservation scheme for LEO satellite networks that exploits the regularity of the movement patterns of mobile hosts in LEO satellite networks. We have developed the cellular network simulator (CNS) for evaluating call-level QoS provisioning schemes. QoS at the call-level is concerned with call blocking probability (CBP), call dropping probability (CDP), and supplied bandwidth. We introduce two novel QoS parameters that relate to supplied bandwidth – the average percent of desired bandwidth supplied (DBS), and the percent of time spent operating at the desired bandwidth level (DBT).

ACKNOWLEDGMENTS

I would like to thank Dr. Olariu for introducing me to the area of my research, for motivating me to persevere through difficult times and for his help and guidance with my work. He always has time to discuss problems and ideas with me, and he has assisted me with every detail of this work. Dr. Abdel-Wahab has given me his unfailing support during all of my college education. Ever since I took my first course with him during my undergraduate studies more than 20 years ago, he has inspired me. I remember and use the lessons I have learned from him during all the years I have known him.

I am very thankful to my committee members, Dr. Mukkamala, Dr. Wilson and Dr. Song, for their constructive criticism and other suggestions and help. I am grateful to Dr. Todorova for having served on my committee and for her assistance to me in the area of LEO satellite networks. I want to thank Anjlica Malla for her excellent work on the max-min fair borrowing scheme.

I must thank the Computer Science Department at ODU, as I have spent many years here. Thanks to all the professors that have taught me and helped me with projects and also to the department staff – I have had to call on the both the office staff and the network staff for their assistance many times.

I thank my mother and father for their love and support during this long process. My mother was a role model for me in seeking a career in academia, just as she has always been a role model for me in the rest of my life.

My husband, Faraz, and daughter, Mariya, missed me a lot, especially during the last year, without complaining. And my husband put his own goals on hold while I met mine. I am very grateful to both of them.

TABLE OF CONTENTS

Page

List of Tables vii
List of Figures viii

CHAPTER

I Introduction 1
II Background and Related Work 5
II.1 Cellular Networks 5
II.1.1 Background and Early Research 5
II.1.2 Recent Research 9
II.2 LEO Satellite Networks 12
II.3 Multimedia Sessions 14
III The Rate-Based Borrowing Scheme 16
III.1 Oliviera's Schemes 17
III.2 The Rate-Based Bandwidth Borrowing Scheme 18
III.2.1 Cell and connection parameters 19
III.2.2 Fairness of RBBS 20
III.2.3 RBBS details 21
III.2.4 Performance of RBBS 25
IV The Max-Min Fair Borrowing Scheme 30
IV.1 Max-Min Fairness 30
IV.2 Max-Min Fair Bandwidth Allocation 33
IV.3 Performance of the max-min fair borrowing scheme 35
V The Multimedia Session-Aware Scheme 39
V.1 MSAS Details 40
V.2 MSAS in an End-to-End Architecture 42
V.3 Analysis 44
V.4 Performance of MSAS 45
V.4.1 Simulation Model 45
V.4.2 Experimental Results 47
VI The Predictive Reservation Scheme for LEO Satellite Networks 58
VI.1 Mobility and Traffic Models 58
VI.2 The predictive bandwidth reservation strategy 60
VI.2.1 New call admission strategy 61
VI.3 Experimental Results 63

VII Conclusions and Future Work 66

REFERENCES 68

APPENDIX

The Cellular Network Simulator 75

VITA 82

LIST OF TABLES

		Page
1	Traffic Characteristics for the Simulation	25
2	Traffic Characteristics for the Simulation, with Sessions	46

LIST OF FIGURES

		Page
1	Cells laid out in a honeycomb pattern	5
2	The footprint of a LEO satellite subdivided into spotbeams	13
3	RBBS connection parameters	19
4	The recovery of a Class I handoff	23
5	A comparison of bandwidth utilization	27
6	Call dropping probabilities for Class I traffic	27
7	Call dropping probabilities for Class I and Class II traffic combined	28
8	Call blocking probabilities for Class I traffic	29
9	Call blocking probabilities for Class I and Class II traffic combined	29
10	Insufficient bandwidth to satisfy new connection	31
11	Allocation of the equal share to all connections	31
12	Redistribution of the residual bandwidth	32
13	Final redistribution of the residual bandwidth	33
14	Call dropping probabilities for Class I traffic	35
15	Call dropping probabilities for Class I and Class II traffic combined	36
16	Call blocking probabilities for Class I traffic	37
17	Call blocking probabilities for Class I and Class II traffic combined	37
18	A comparison of bandwidth utilization	38
19	Call dropping probabilities for the combined traffic	48
20	Call blocking probabilities for the combined traffic	49
21	A comparison of bandwidth utilization	49
22	Call dropping probabilities for traffic type 1A	50
23	Call blocking probabilities for traffic type 1A	50
24	Call blocking probabilities for the session-only experiment	51
25	Call dropping probabilities for the session-only experiment	52
26	Percentage of time at desired bandwidth - traffic type 1A	53
27	Percentage of desired bandwidth supplied - traffic type 1A	53
28	Percentage of time at desired bandwidth - traffic type 2A	54
29	Percentage of desired bandwidth supplied - traffic type 2A	55
30	Borrow-related fluctuations per second - traffic type 1A	56
31	The trajectory of a connection	59
32	The sliding window concept for call admission	62
33	Call dropping probabilities for Class I traffic	64
34	Call dropping probabilities for Class I and Class II traffic combined	64
35	A comparison of the bandwidth utilization	65

CHAPTER I

INTRODUCTION

Quality of Service (QoS) provisioning for wireless networks is becoming more important by the day. Cellular phones are now ubiquitous throughout the world. Because the physical infrastructure costs of wireless networks are so much less expensive than those of wired networks, in many developing countries the wireless coverage, in area and population, is greater than the wired coverage. With more users, there is a growing demand for more services. People want to access services like email, instant messaging, web browsing, and even video streaming and video conferencing, as well as the more common services like telephony and paging, from anywhere or while on the move.

Most cell phone handsets and network providers in the U.S. provide users with access to telephony, paging, instant messaging, and trivial web browsing on the same device. Newer services and phones provide the ability for non-interactive multimedia such as sending and receiving still images and short video clips. Some new phones add the ability to do more complex web browsing by combining a small PDA with a phone handset. And users who have more powerful computers equipped with wireless network access want to be able to use the same, possibly high-bandwidth, applications they are accustomed to using with wired network access, such as video conferencing and other real-time networked applications.

QoS provisioning in single-class traffic wireless networks is a more complicated task than it is in wired networks due to physical constraints, bandwidth constraints, and user mobility. Clearly, providing QoS guarantees to users of wireless networks that handle different types or classes of service is an even more complex task, and it is currently a busy research area.

Some well-known QoS measures, such as delay, bandwidth, jitter and error rate, are important in both wired and wireless networks. The chance that a new connection request will be denied, the *call blocking probability* (CBP), is an important QoS measure in wireless networks due to limited bandwidth. And the chance that an in-progress connection will be forcibly terminated by the network, the *call dropping*

This dissertation follows the style of *IEEE Transactions on Parallel and Distributed Systems.*

probability (CDP), is important due to user mobility. Kalyanasundaram [1] proposes a hierarchical classification of QoS measures:

- **Packet level:** QoS at this level includes packet dropping probability, maximum packet delay, and maximum jitter.
- **Call level:** At this level, [1] mentions just CBP and CDP. We consider that supplied bandwidth also belongs at this level.
- **Class level:** QoS at this level describes the requirements for and relationships between the call-level QoS for different classes of traffic. Whether traffic is classified according to its call-level and packet-level QoS needs, or according to price paid for service, or by some other means, class-level QoS will describe the prioritization among the different classes of traffic.

Our work focuses on call-level QoS provisioning in cellular and LEO satellite networks. Our admission control and resource reservation schemes are local in the sense that they deal with the wireless link of a connection only. They deal with the specificities of the mobile environment, and they could serve as components in a larger end-to-end QoS provisioning framework. Our schemes consider the users' needs in relation to the service providers' needs; for example, minimizing the CDP and CBP while keeping the bandwidth utilization as high as possible. And we consider the users' needs with respect to one other, within and between traffic classes, i.e. making our algorithms *fair*. Three of the the schemes are designed for cellular networks. The key feature of these schemes is that they use resource reallocation - temporarily borrowing from existing users of network resources in order to accomodate new users. Many multimedia applications can adapt to fluctuations in their supplied resources, for example, by the use of an adjustable-rate codec. Non-real-time applications such as email or ftp can also tolerate fluctuations in their supplied bandwidth, although they cannot tolerate loss. At their startup, connections make an agreement with the network about how adaptive they can be to changes in their supplied resources.

The first scheme for cellular networks, the *rate-based borrowing scheme*, relies on connections providing a measure of their willingness or ability to adapt in the form of stated desired and minimum bandwidth levels. A portion of the difference between these two levels is divided up into a fixed number of portions, or *shares*. The network then borrows bandwidth, in times of need, in the amount of one share from

each connection. When bandwidth is freed, the network returns bandwidth to the connections one share at a time. The share concept ensures that connections experience gradual changes in their supplied resources. We show the rate-based scheme to be proportionally fair with respect to the connections' willingness to adapt. The bandwidth borrowing is combined with a fixed bandwidth reservation for handoffs. The performance of the rate-based scheme in terms of CDP and CBP is shown to be better than the best comparable bandwidth allocation and reservation schemes found in the literature at the time [2].

The second scheme for cellular networks is the *max-min fair borrowing scheme.* This scheme is similar to the rate-based borrowing scheme, but it substitutes max-min fairness for proportional fairness. It does not have the concept of shares, and all the bandwidth available for borrowing may be borrowed when necessary. Bandwidth is returned as soon as it is freed by a departing connection. This scheme does better than the rate-based scheme in terms of CBP and CDP; however, this is mainly due to the fact that most high bandwidth connections are blocked. The definition of fairness allows that high bandwidth calls may be rejected even when there is sufficient bandwidth available for them.

The third scheme for cellular networks that we introduce is an extension of the rate-based scheme. It adds an awareness of the fact that some real-time applications, for example, video conferencing, do not use a single connection, but rather a set of cooperating streams, known as a *multimedia session.* With awareness of the existence of these sessions, our *multimedia session-aware borrowing scheme* is able to eliminate the processing overhead suffered by QoS provisioning schemes that treat each stream of a session as a distinct entity. It also improves upon schemes, such as our own rated-based borrowing scheme, that treat a set of multimedia session streams as one single large connection. By allowing a simple prioritization between the streams in a session, significant improvements in CDP can be gained.

Schemes like ours and many other very recent call-level QoS provisioning schemes exploit the adaptiveness of some types of traffic and actually *remove resources from them or even end them prematurely* in order to meet overall goals like a lower CDP and CBP. Analysis of this class of schemes must take into account the negative aspects as well as the positive. Therefore we introduce some new metrics to use in analysis. Two new QoS metrics relate to a connection's supplied bandwidth:

- the percentage of the Desired Bandwidth that is actually Supplied (**DBS**), and

- the percentage of Time spent operating at the Desired Bandwidth (DBT).

For schemes which create fluctuations in the supplied bandwidth of ongoing connection, the fluctuation rate should also be measured.

In the following chapter, we review the state of the art in the areas of call-level QoS provisioning for cellular networks and LEO satellite networks, and supplying QoS to a multimedia session. In chapter 3, we present the rate-based bandwidth borrowing scheme and its performance measurements. Chapter 4 covers the max-min fair bandwidth borrowing scheme, and we compare it to the rate-based scheme. In chapter 5, we present a multimedia session-aware scheme that is based on the rate-based borrowing scheme and compare it to a similar, but session-unaware, scheme. We do a full analysis of the schemes that involves looking at the results for various traffic types in detail in terms of CDP, CBP, DBS, and DBT. We also consider the overhead caused by borrowing, and look at the rates of borrowing-produced fluctuation for different traffic types. Chapter 6 presents the predictive reservation scheme for LEO satellite networks. And the appendix contains the detailed design of the cellular network simulator (CNS).

CHAPTER II

BACKGROUND AND RELATED WORK

II.1 CELLULAR NETWORKS

II.1.1 Background and Early Research

Cellular networks usually consist of a set of base stations connected together through a wired network, and mobile hosts that communicate through their local base station to reach other mobile hosts or the wired internet. The area served by one base station is called a *cell*, and cells are commonly arrayed in a honeycomb pattern, with each cell having 6 neighbors (Figure 1). When a mobile host (MH) moves out of the range of one base station (BS) and into the area served by a neighboring base station, it is called a *handoff* or *handover*. A set of base stations that neighbor each other may be connected over a wired link to a local controller called a Mobile Switching Center (MSC).

Since our work does not focus on packet-level QoS, we will not review the work done in that area, but rather begin with call-level QoS and channel allocation. Because of co-channel interference, the entire frequency spectrum assigned to a network cannot be used in every cell. Much research has been done [3, 4, 5, 6] into how to allocate channels among the cells in order to maximize the available bandwidth for user requests. The simplest is the *fixed* channel allocation, in which all the channels

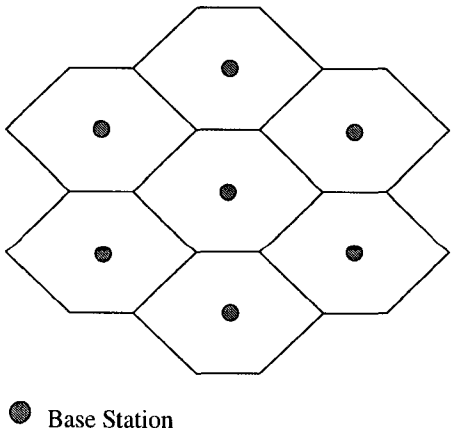

Fig. 1. Cells laid out in a honeycomb pattern.

of the spectrum are distributed statically to the base stations in a pattern that reduces or prevents cochannel interference. Each cell has a fixed number of channels to accommodate new and handoff calls. In *dynamic* channel assignment schemes, channels are assigned to cells as they are needed. *Flexible* channel assignment schemes are hybrids of fixed and dynamic assignment, where a set of channels is distributed statically and the rest of the channels are kept in a pool to be distributed as the need arises. Dynamic channel assignment will clearly outperform fixed assignment in times when a few cells are highly loaded, because highly loaded fixed assignment cells will exhaust all their available bandwidth with no method to obtain more. However, fixed assignment will perform better under uniformly high load conditions, unless the dynamic assignment scheme can produce the optimal configuration of fixed assignment as it progresses.

Many dynamic channel assignment schemes are complex and require a central controller. Prakash [7] proposes a distributed dynamic allocation scheme. A base station makes its channel assignment decisions itself, after conferring with neighboring base stations that are within cochannel interference distance of it. A distributed borrowing scheme is introduced by Das [8]. In this scheme, a fixed channel assignment is made initially, and then heavily loaded cells may borrow free channels from their neighbors when needed.

To provide greater frequency utilization, cells may be divided into concentric regions. In the region closest to the base station, channels can be used at low power without interfering with the same channels being used in the central regions of neighboring cells [9]. Although these types of schemes increase the available bandwidth, they are complex to manage and introduce an imbalance between resources available for calls, depending on the region in which they originate. Papavassiliou proposes an algorithm [10] to solve the problem that fewer channels are available in the region furthest away from the base station, and to create a uniform CBP across an entire cell. Dividing a cell into regions also introduces the possibility of handoff being necessary when a mobile host moves from the inner to the outer region of a cell.

Handoff is a problem for QoS provisioning because the new region or cell that a mobile host is entering may not have the resources available to grant the same level of QoS that the hosts had originally negotiated or may not be able to continue the host's call at all. And as cells become smaller (as in micro-/pico- cellular networks), handoffs will occur much more frequently. Handoff detection is in itself a complex

issue. As a mobile host moves away from one base station and towards a neighboring base station, the signal from one will get progressively weaker and from the other progressively stronger, until in some overlap area, they become equal for a time. But other things, such as terrain, can also affect signal strength. Tekinay [11] proposes some algorithms to assist in accurate handoff detection. Different methods of *handoff queueing* are presented as a means to lower the CDP in work by Tekinay [12] and by Agrawal [13]. If a MH is in the area where it can hear the signals of two BSs and its destination cell cannot support it, it is placed into a queue, with the idea that maybe the resources for it will become available before the time when it is completely outside the range of its current BS.

Reducing the CDP, or the handoff failure rate, is more important than reducing the CBP, because it is more disturbing to users to have an ongoing call cut off than to have a call attempt denied [12]. Dynamic and flexible channel allocation schemes aid in lowering the CDP because channels can be allocated or borrowed when handoffs arrive and find no available bandwidth. A common technique for reducing dropped calls is to reserve some portion of the bandwidth in a cell for use only by handoffs [14, 15, 16, 2]. The simplest method of bandwidth or channel reservation is static, where a fixed set of channels, or fixed percentage of a cell's bandwidth, is set aside for handoffs. Selecting the amount of bandwidth to reserve is a trade-off between the CBP and CDP: the larger the reservation, the less bandwidth available for new calls. Dynamic or adaptive reservation uses information about the network or the mobile hosts to vary the amount of reserved bandwidth with the current conditions in order to reach the optimal balance between CBP and CDP.

Yu presents a scheme [16] in which the traffic statistics of neighboring cells are polled to determine how much bandwidth should be reserved for handoffs. Knowledge of or prediction of the mobility characteristics (i.e. speed and direction) of the MHs can improve adaptive bandwidth reservation schemes. In Choi's scheme [14], traffic history is used to predict MH behavior, assuming that MHs entering a cell from the same direction that other MHs have come from in the past will probably go in the same direction as their predecessors and at a similar speed, because of roads. The *shadow cluster concept* [17] refers to the group of cells surrounding a MH's current cell which has the highest probability of being visited by the MH in the near future. This *region of influence* for each MH is calculated based on the probable call duration and reports sent by the MH to the BS containing position, speed and heading information.

Aljadhai proposes a scheme [18] which is based on the shadow cluster concept, but which uses historical information about a MH's movement to compute probabilities for its future movements, rather than relying on the MHs to report. This method is of course less precise, but probably more feasible. This scheme [18] also extends the original algorithms [17] to apply to multiple classes of traffic. Su [19, 20] has done extensive research into mobility prediction for bandwidth reservation schemes.

Call admission control schemes help a BS decide whether or not to grant service to connection requests. Admission control schemes are like bandwidth reservation, in that by using information about network conditions, they may deny a connection request even though the call may be serviceable in the short term, in order to leave resources available for demands expected later. Agrawal [13] proposes an admission control method that assumes single-class traffic requiring one channel per MH. The area at the perimeter of a cell is designated as the *handover zone*, and the area just inside that is the *pre-handover zone*. In this scheme, a new call is assigned a channel only if no MHs are in the cell's handover zone and if the number of MHs in the pre-handover zone is less than the number of available channels. Naghshineh [21] proposes a scheme in which neighboring cells periodically exchange information about their state. When a new call request arrives, a cell uses its knowledge about its neighborhood to decide whether or not to admit the new call, with the goal of limiting the CDP to a given value. This scheme has been criticized for its complexity [22]. A hybrid of the *weighted sum scheme* and the *probability index scheme* [23] has been shown to compare favorably to a static bandwidth reservation scheme and to Naghshineh's scheme [21]. The weighted sum scheme makes admission decisions based on the weighted sum of the number of ongoing calls in the neighboring cells, with the weight based on the distance between the cell and its neighbor. The probability index scheme estimates the dropping probability for the new call and compares it to a threshold value in order to make an admission decision. Sadeghi and Knightly [24] propose an admission control scheme called the *Virtual Bottleneck Cell*, which forms cells into clusters and avoids resource control at the user level in order to be highly scalable.

Many of the schemes we have discussed so far consider just single-class traffic, where each connection has the same requirements, using one channel or bandwidth unit per call. Some researchers have proposed generic algorithms and models for multi-class traffic of n types [25, 26, 27]. Others have modeled systems where each

call may request different QoS, but traffic is mainly classified based on whether it is real-time or non-real-time [2, 28, 29, 30, 31, 1]. Das [29] makes just this distinction in a framework he proposes for QoS at the packet and at the call level. The same classification is made by Naghshineh in [28], where an end-to-end QoS provisioning framework is proposed. These researchers suggest that some real-time applications have the ability to adapt themselves to fluctuations in the QoS. For example, video can be sent in *layers* of quality, so that if supplied bandwidth is reduced, layers can be dropped and the video quality lowered, without suffering a loss in the continuity of the video stream [32]. Non-real-time applications, such as email or ftp, can not tolerate loss, but are able to tolerate significant changes in supplied bandwidth.

Oliviera [2] has proposed a set of call admission control and bandwidth reservation algorithms for multi-class traffic. In these schemes, it is assumed that traffic falls into two classes, *Class I* designating real-time and *Class II* designating non-real-time, and that successful handoffs are more important to Class I users than Class II. The algorithms specify that in order for a Class I connection to be accepted in a cell, bandwidth must be available to be reserved for it in the neighboring cells. In one scheme, a fixed amount is reserved in a cell depending on the number of connections requesting a reservation. In another version, the amount reserved is equal to the maximum of all the current reservation requests for the cell. When a new connection is attempted in a cell, the service class, the desired bandwidth, and the minimum acceptable bandwidth are specified. If the desired bandwidth cannot be granted, the connection is blocked. Class I calls are blocked if reservations cannot be made for them. For handoff connection requests, only Class I connections have the cell's reserved bandwidth at their disposal. A Class I handoff must be granted at least its minimum bandwidth, and reservations must be made for it in neighboring cells, or it is dropped. A Class II handoff will succeed if there is at least one unit of bandwidth available for it.

II.1.2 Recent Research

In work related to our rate-based borrowing scheme [30], Lam [33] proposes a scheme that combines intra-cell borrowing with knowledge about the state of the client's video buffer in order to provide QoS to video streams. The QoS is measured mainly in terms of frame loss rate, and it is shown to outperform the rate-based scheme on that measure. However, Lam's scheme suffers from the fact that it does not consider

traffic other than video and assumes an intimate knowledge about the video buffers and their states on the clients, as well as the encoding method and playback rates of the individual streams. It differs from the rate-based scheme in both its assumptions and its goals. Tupelly [34] introduces a scheme also addressed solely to MPEG video traffic in cellular networks, and its performance is measured in terms of frame loss rate. It uses information about channel throughput, frame type and buffer state to schedule traffic.

Vakili and Aziminejad [35] propose a scheme in which the bandwidth in a cell starts with an initial reservation for audio connections and video connections. As calls arrive, bandwidth (of the same type - audio or video) can be borrowed from neighboring cells. Also, audio connections can temporarily borrow unused video bandwidth in the same cell. Non-real-time connections are given whatever bandwidth is left-over in the cell. This scheme makes fixed reservations based on traffic type, and then allows both inter- and intra-cell borrowing of unused bandwidth under certain conditions. It is easy to see that this scheme might lead to poor bandwidth utilization if the initial reservation were not chosen well. But with good choices based on knowledge of traffic patterns in the network, it could perform quite well.

At Monash University in Australia, a group of researchers are working on precisely the same problem set as we are. They have proposed a number of schemes [36, 37, 38, 39, 40, 41] that use knowledge about the mobility parameters of the MHs (speed and direction) in order to perform statistical bandwidth reservation, expanding on the ideas proposed in [17] and [18]. One scheme [38] combines a mobility-based reservation scheme with an inter-cell borrowing scheme similar to ours. In the earliest scheme [36], the MHs give information to the BS about their direction and the SNR and signal strength are monitored in order to enhance the mobility information. Bandwidth is allocated in a MH's current cell and reserved in the most likely cells to which a MH will travel. The new reservation scheme is compared to a simple fixed reservation technique in terms of CBP and bandwidth utilization. The second mobility-based scheme [37] adds a *speed-dependent directional probability* function that assumes that faster-moving MHs are less likely to change direction. This information makes the reservation even more precise. As in Oliviera [2], it is possible to drop a MH that can be accommodated in its current cell if reservations cannot be made on its behalf. The determination of when to drop is more forgiving in this new scheme than in Oliviera's.

The scheme that combines mobility-based reservation and a borrowing scheme

called *on-demand borrowing (ODB)* is compared favorably with our rate-based scheme. ODB borrows as much as needed from active calls in the same cell (according to their adaptiveness), and will even terminate ongoing Class II (using the same terminology as Oliviera) connections in order to free enough bandwidth to satisfy a Class I handoff. The borrowing scheme is said to be fair, but the type of fairness is not mentioned. It is clear that a scheme with movement prediction that supplies good information for statistical reservation should outperform a fixed reservation scheme (as is used in our schemes) in CDP, because it is more likely that the needed bandwidth will be waiting for calls when they arrive. A borrowing scheme that allows calls to be suspended (as many as necessary) also would deliver low CDP for class I calls at the expense of very poor QoS for the class II calls who got disconnected. Schemes which reserve bandwidth for calls in cells to which they are expected to travel in the future suffer from some communication overhead. This overhead is due to the message passing between cells to make and release reservations.

Zander [42] has proposed combining a bandwidth borrowing scheme with a statistical reservation technique that uses a database of historical movement patterns in order to estimate handoff arrival rates. The reservation technique is described in detail, but the borrowing technique is not; using our max-min fair technique is given as a possibility. The key feature of this scheme is that it allows a network operator to develop *penalty functions* which take into account relative priorities between traffic types, the adaptiveness of connections, and the desired and measured QoS. The scheme uses these functions to help make decisions about blocking and dropping, and borrowing and redistributing freed bandwidth. Its goal is to minimize annoyance as defined by the penalty functions. For example, it may be more acceptable to drop a connection than to borrow from a connection that has already suffered from previous borrows.

Curescu [43] has developed an interesting scheme that focuses on utility, rather than CBP and CDP. Traffic types are assigned a utility value, and the QoS they receive affects their utility. A scheme is developed that tries to maximize the overall system utility. Resource reallocation is used, and on-going connections may even be terminated in favor of new calls in order to maximize utility. When simulated against our rate-based scheme the utility-based scheme gets good results, but the goals and assumptions of the two schemes differ.

Ibrahim [44] has developed a novel charging and resource allocation framework

for wireless networks. Connections supply a QoS profile upon seeking service from the network which specifies a set of levels of desired QoS and what the user is willing to pay for each. A group of schemes that feature resource reallocation with the goal of maximizing revenue for the network service provider is proposed. These schemes have also compared favorably to our rate-based scheme, but again the assumptions and goals are different.

II.2 LEO SATELLITE NETWORKS

There are some places, such as low population areas, harsh terrain or the middle of the ocean, where it is impossible or economically infeasible to deploy land-based wireless systems. Because of their ability to provide service in such areas, satellite networks hold the promise of true global communications coverage. *Low Earth Orbit (LEO)* satellite networks, deployed at relatively low altitudes (500km-2000km) as compared to geosynchronous earth orbit (GEO) satellites, have a small signal propagation delay. In addition, the power required to transmit that signal is comparatively low. Therefore, the ground-based systems for LEO satellite networks can be hand-held devices, allowing for world-wide personal communications services (PCS).

The area on the earth that is covered by a certain satellite is called its *footprint.* A LEO satellite is composed of a number of *spotbeams* that illuminate the earth's surface inside its footprint in a pattern of slightly overlapping circular cells. A spotbeam controller can be considered similar to a ground-based cellular system's base station. But unlike a cellular base station, a LEO satellite is moving at a very high velocity. This implies that an end-user, even a stationary one, will suffer frequent handoffs as he is switched from one spotbeam to another and from one satellite to another. A handoff of a connection from spotbeam to spotbeam is an *intra-satellite handoff*, and one from satellite to satellite is an *inter-satellite handoff.* Handoff management is a key part of QoS provisioning in LEO satellite networks.

In order to analyze the problem of intra-satellite handoffs, it is common practice to model a satellite network like a cellular network in which the mobile hosts are all moving in straight lines in the same direction from spotbeam to spotbeam at a constant speed that is equal to the satellite's orbital velocity [45, 46]. Some researchers are applying techniques designed for cellular systems directly, albeit with a loss in efficiency, because the specific characteristics of the LEO networks are not considered. Del Re [45] proposes a dynamic channel allocation scheme and applies the *handoff*

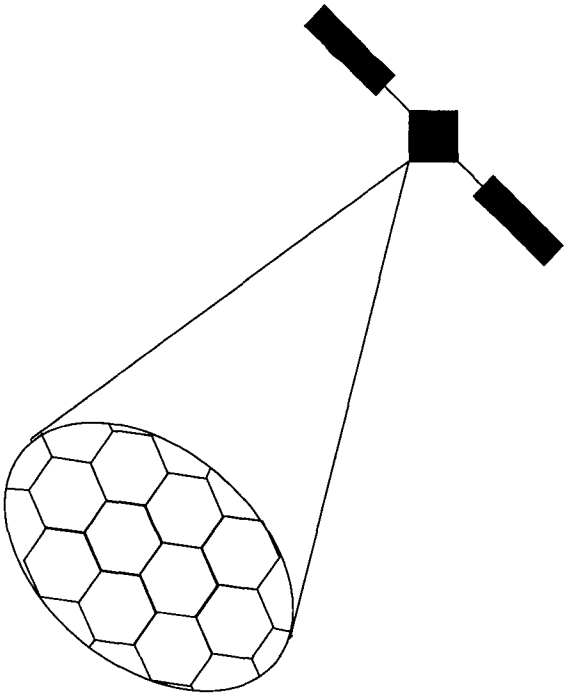

Fig. 2. The footprint of a LEO satellite subdivided into spotbeams.

queuing technique.

The need for mobility prediction does not exist in the satellite model because the motion of MHs is defined by the motion of the satellites. This simplifies some of the cellular schemes, or provides them with better performance because the knowledge of the speed and direction of the MHs can be exploited. For example, Kalyanasundaram [1] applies a *channel-sharing* scheme to satellite networks. In this scheme, each adjacent pair of cells, $\{n, n+1\}$, forms a *meta-cell* which has a set of channels assigned to it, thus allowing some users to go from cell n to $n+1$ carrying the same channels with them. Initially, a fixed channel allocation scheme partitions the channels among the spotbeams according to the correct reuse distances. A call admission scheme determines which channels to assign to a connection request in order to improve the connection's chances of carrying the same channels through its next spotbeam handoff. Applied to a land-based cellular network, Kalyanasundaram's scheme [1] is much more complex.

Because the movement pattern of MHs in a LEO satellite network is known, many researchers have developed resource reservation techniques that use the knowledge of the MH mobility to improve their efficiency [47, 48, 49, 50, 51, 52].

II.3 MULTIMEDIA SESSIONS

Applications like video phone, video conferencing and distance education use a set of streams operating together in what is known as a *multimedia session.* For example, an interactive distance learning application might include a connection for the teacher's audio, one for the teacher's video, one for slides, one for a whiteboard, an audio stream for students' questions and a connection for control messages. The set of individual streams can be considered a single entity, a session, because the streams will start and end together, have the same endpoints and be interrelated in other ways as well, i.e., audio/video synchronization. And if the user of such an application is mobile, of course, the streams handoff together. Therefore, it makes sense to consider the session, rather than the individual streams, in terms of call-level QoS, particularly CDP.

Some protocols recently defined by the Internet Engineering Task Force (IETF) that treat a set of streams as a single entity are the Session Description Protocol (SDP), the Session Initiation Protocol (SIP), and the Session Announcement Protocol (SAP). SDP (RFC 2327) [53] allows a session and its component streams to be described. It is used by SIP (RFC 3261) [54], which is the control protocol for initiating a session, and SAP (RFC 2974) [55], which is used to announce sessions that will occur in the future. SAP and SDP were originally designed with MBONE multimedia conferences in mind, as a way to announce upcoming conferences – SAP uses SDP for the session description. Now, with SDP being used in many more environments than it was originally designed for, such as providing media descriptions to media gateways [56, 57], a new version is being developed – SDPng [58].

SDP is a text-based protocol that describes the structure of a multimedia session. It consists of a session information block and zero or more media information blocks. The session block may contain data items like name, valid times and encryption key and other attributes. A media block contains information such as media type (e.g., video), the transport protocol (e.g., UDP), the media format (e.g., MPEG), bandwidth specifications, the transport address, and other attributes.

There has been a lot of research and standards work looking at providing QoS to sessions. RFC 3524 [59], *Mapping of Media Streams to Resource Reservation Flows*, discusses an extension to SDP that allows one or more media streams to be grouped into a single reservation flow (SRF). This information can then be used by a resource

reservation protocol or scheme, such as RSVP (Resource Reservation Setup Protocol) [60] to map the media streams to resource reservations. In RFC 3312 [61], the integration of resource management with SIP is discussed with the introduction of a framework for *preconditions.* Preconditions for QoS in a SIP session ensure that QoS reservations are made before the session begins. The preconditions are specified in the session description using SDP and can indicate the importance, or *strength* of the QoS requirement, the *direction* of the requirement (send, recv, or sendrecv), and the *type* of requirement, which refers to whether it is end-to-end or *segmented.* Segmented refers to QoS in the access networks only, with the possible values of *local* and *remote.*

So the ability to specify QoS requirements to the local network for the streams within a session exists, but most work has focused on the provision of end-to-end QoS. We will not review end-to-end frameworks for session QoS here because our focus is on the local wireless link, but we offer a few references. Siddiqui [62] looks at end-to-end QoS for SIP sessions in CDMA networks. Prior to the definition of SIP and SDP, Youssef [63] introduced the term *QoSess*, Quality-of-Session, QoS for multimedia sessions. The ITU-T H.323 is a very popular standard for multimedia conferencing that predates SIP, but SIP is becoming the protocol of choice as it is more open and general.

CHAPTER III

THE RATE-BASED BORROWING SCHEME

As noted in the previous chapter, much of the work in QoS for cellular networks has considered only single-class traffic, ignoring the more complex case of multi-class traffic. But in light of the current and expected future demands for wireless networks, it is essential to create QoS provisioning schemes that address networks carrying a mixture of traffic classes. Multimedia applications are known to be able to tolerate and adapt to transient fluctuations in QoS [64, 65, 66]. This adaptation is typically achieved by the use of an adjustable-rate codec or by employing hierarchical encoding of voice and/or video streams [64, 65, 67, 68]. The codec, along with appropriate buffering before play-out, can allow applications to adapt gracefully to temporary bandwidth fluctuations with little or no perceived degradation in overall quality. The additional flexibility afforded by this ability to adapt can be exploited by protocol designers to significantly improve the overall performance of cellular systems. We propose an admission control and bandwidth reservation scheme that takes advantage of the adaptability of some the classes of traffic in order to improve the QoS provided by the network in terms of CBP and CDP while maintaining good resource utilization.

As mentioned in the previous section, it is more acceptable to deny a new connection request than to terminate an in-progress connection [8, 9, 11, 12]. The numerous strategies proposed to lower the CDP include channel rearrangement, handoff queuing, and channel reservation. Channel rearrangement and handoff queuing techniques are effective; however, our scheme uses reservation. There are, essentially, three approaches to resource reservation:

- *fixed reservation*, where a certain percentage of the available resources in a cell are permanently reserved for handoff connections, and
- *statistical reservation*, where resources are reserved using a heuristic approach. These approaches range from allocating the maximum of the resource requirements of all connections in neighboring cells, to reserving only a fraction of this amount [17, 2].
- *hybrid reservation*, where a small amount of fixed reservation is maintained,

while a statistical method is used to determine the amount of resources that need to be reserved beyond the fixed amount.

Our rate-base borrowing scheme (RBBS) combines a resource reservation scheme with a companion fair borrowing scheme.

III.1 OLIVIERA'S SCHEMES

In order to set the stage for the description of our algorithms, we will now describe the bandwidth allocation and reservation schemes proposed by Oliviera [2] in more detail. We chose these schemes as a benchmark since they are arguably better than other comparable bandwidth allocation and reservation schemes found in the literature [2].

When a MH requests a new connection in a given cell, it provides the following parameters:

- the class of traffic (either I or II);
- the desired amount of bandwidth for the connection; and
- the minimum acceptable amount of bandwidth, that is, the smallest amount of bandwidth that the source requires in order to maintain acceptable quality, e.g. the smallest encoding rate of its codec.

One of the significant features of the call admission control and bandwidth reservation schemes in [2] is that in order to admit the connection, bandwidth must be allocated in the originating cell and, at the same time, bandwidth must be reserved for the connection in all the neighboring cells. Specifically, for a new connection to be admitted in a cell, the cell must be able to allocate the desired bandwidth to the connection. For Class I connections, the call will be blocked unless the desired bandwidth can be allocated to it in the original cell and some bandwidth can be reserved for it in each of its six neighboring cells.

During a handoff, an established Class I connection is dropped if its minimum bandwidth requirement cannot be met in the new cell or if appropriate reservations cannot be made on its behalf in the new set of neighboring cells. However, Class II traffic has no minimum bandwidth requirement in the case of a handoff, and a call will be continued if there is any free bandwidth available in the new cell.

The schemes presented in [2] use statistical reservation techniques based on the number of connections in neighboring cells, the size of the connections in neighboring

cells, the predicted movement of mobile hosts, and combinations of these factors. It is worth noting that the reservation schemes in [2] keep the dropping probability for Class I connections very low, since the mobile host should find bandwidth reserved for it, regardless of the cell to which it moves. But bandwidth may be wasted in the neighboring cells (the host can only move to one neighbor), and the blocking probability in those cells may increase because unused bandwidth is being kept in reserve. In general, the schemes described in [2] favor minimizing the CDP at the expense of the CBP and give Class I traffic precedence over Class II traffic.

III.2 THE RATE-BASED BANDWIDTH BORROWING SCHEME

It is clear that keeping a small pool of bandwidth always reserved for handoffs, as in [2], yields low CDP. However, in our schemes, the size of the reserved pool is not determined by requests from neighboring cells, but is *fixed* at a certain percentage of the total amount of bandwidth available in the cell. We found that this produced results similar to the best results reported in [2], without the overhead of communication between neighboring base stations to request and release reservations. It was also reported in [23] that simple fixed reservation can outperform more complex statistical techniques. To further reduce the CDP in our scheme, we treat the reserved pool very carefully. We do not allow bandwidth from the reserved pool to be allocated to incoming handoffs unless the bandwidth is needed to meet the minimum bandwidth requirements of the connection. Just as in [2], our scheme gives precedence to Class I connections; Class II traffic does not make use of the reserved bandwidth.

RBBS attempts to allocate the desired bandwidth to every multimedia connection originating in a cell or being handed off to that cell. The novelty of our scheme is that in case of insufficient bandwidth, in order not to deny service to a requesting connection (new or handoff), bandwidth will be borrowed, on a temporary basis, from existing connections. Our borrowing scheme guarantees that no connection will give up more than its *fair share* of bandwidth, in the sense that the amount of bandwidth borrowed from a connection is proportional to its tolerance to bandwidth loss.

Our borrowing strategy has the following interesting features:

1. It guarantees that the bandwidth allocated to a real-time connection never drops below the minimum bandwidth requirement specified by the connection at call setup time. This is very critical to ensuring that the corresponding application

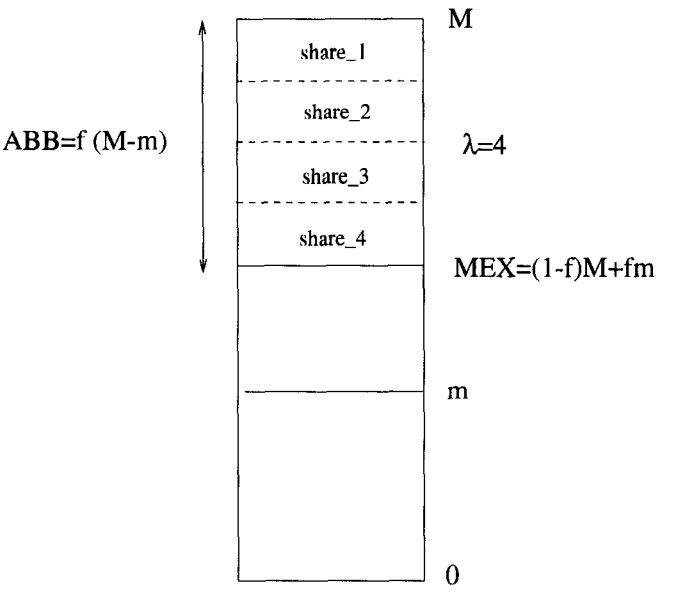

Fig. 3. RBBS connection parameters.

can still function at an acceptable level.

2. It guarantees that if bandwidth is borrowed from a connection, it is borrowed in small increments, allowing time for application-level adaptation.

3. It is *fair* in the sense that if bandwidth is borrowed from one connection, it is also borrowed from the other existing connections. Specifically, if borrowing is necessary in order to accommodate a requesting connection (new or handoff), every existing connection will give up bandwidth in proportion to its tolerance to bandwidth loss. This motivated us to refer to our scheme as *rate-based* fair.

4. Finally, the borrowed bandwidth is returned to the connections as soon as possible. Thus, the degradation in the QoS is transient and limited to a minimum.

III.2.1 Cell and connection parameters

Each cell maintains a pool of bandwidth reserved for Class I handoffs which, initially, represents r percent of the total bandwidth. At setup time, each connection specifies to the cell in which it originates a *maximum bandwidth* M (termed the *desired bandwidth*) and a *minimum bandwidth* m as illustrated in Figure 3. The difference between

these two values is the *bandwidth_loss tolerance* (BLT) of the connection. Thus,

$$BLT = M - m.$$

We note that for constant bit rate (CBR) connections $M = m$, indicating no bandwidth_loss tolerance and, thus, $BLT = 0$.

Each cell maintains a local parameter, f, $(0 \leq f \leq 1)$, which represents the fraction of the BLT that a connection may have to give up, in the worst case. This fraction is the *actual borrowable bandwidth* (ABB) of the connection. Thus,

$$ABB = f \times BLT = f(M - m).$$

By accepting a new call, the base station agrees that the supplied bandwidth will not fall below a certain level that we call the *minimum expected* (MEX) bandwidth that the connection is guaranteed to receive during its stay in its starting cell. By definition, $MEX = M - ABB$. It is worth noting that $MEX \geq m$. Simple computation shows that MEX is a weighted average of M and m in the sense that

$$MEX = (1 - f) \cdot M + f \cdot m.$$

To prevent borrowing from producing noticeable changes in a connection's QoS, we introduce another cell parameter, λ. The ABB is divided into λ *shares*, each share being equal to $\frac{M-MEX}{\lambda}$. This provides the basis for a method of borrowing bandwidth gradually from a set of connections whose allocated resources may be quite different. A cell is said to be operating at level L, $(0 \leq L \leq \lambda)$, when all its ongoing connections have had L (or more) shares borrowed from them.

It is important to note, however, that it is possible for a connection to be missing more than L shares after a handoff, due to the sacrifices made to prevent call dropping. However, our scheme attempts to restore bandwidth to handoff connections as soon as it becomes available.

III.2.2 Fairness of RBBS

We now introduce a further connection parameter that we call *adaptivity* (AD), which underlies our borrowing scheme. Specifically, for a given connection, AD is the ratio between the connection's bandwidth loss tolerance and the maximum bandwidth that the connection can use.

$$AD = \frac{\text{bandwidth\_loss\_tolerance}}{\text{desired\_bandwidth}} = \frac{M - m}{M}. \quad (1)$$

It is worth noting that the higher the AD the more adaptive the connection, and the lower the probability of a forced termination in case of a handoff. Notice, again, that for CBR connections the adaptivity is 0.

Consider an arbitrary cell operating at level L. Recall that this implies that every connection in the cell has given up L of its shares. Consider an arbitrary connection C with desired and minimum bandwidth M and m, respectively. Since the cell operates at level L, connection C must have lost L of its shares operating at an effective bandwidth of

$$M - L \times \frac{ABB}{\lambda}.$$

The *loss ratio* (LR) of connection C is the ratio between the amount of bandwidth borrowed from C and the maximum bandwidth M specified by C at setup time. In other words

$$LR = \frac{L \times \frac{ABB}{\lambda}}{M}. \tag{2}$$

Direct manipulations of (2) reveal that

$$LR = \frac{Lf}{\lambda} \times \frac{M - m}{M} = \frac{Lf}{\lambda} \times AD. \tag{3}$$

Since for a given cell, and a given point in time, $\frac{Lf}{\lambda}$ is a constant, (3) shows that the connection will give up an amount of bandwidth proportional to its adaptivity.

Let C' be an arbitrary connection in the same cell as C, and let $LR(C)$ and $LR(C')$ be the corresponding loss ratios. Then (3) allows us to write

$$\frac{LR(C)}{LR(C')} = \frac{\frac{Lf}{\lambda} \times AD(C)}{\frac{Lf}{\lambda} \times AD(C')} = \frac{AD(C)}{AD(C')}.$$

Thus, the ratio of the loss ratios of two connections is invariant to L and is only a function of the adaptivity of the connections. This is the sense in which we consider our borrowing scheme to be fair.

III.2.3 RBBS details

New call admission protocol

When a new call requests admission into the network in a cell operating at level L, the cell first attempts to provide the connection with an amount of bandwidth equal to its desired bandwidth minus L shares of its ABB, that is

$$M - L \cdot \frac{ABB}{\lambda} = \left(1 - \frac{Lf}{\lambda}\right) \cdot M + \frac{Lf}{\lambda} \cdot m. \tag{4}$$

If the amount of bandwidth specified in (4) exceeds the amount of bandwidth available, the cell tests to see if the call could be admitted if the cell progressed to level $L+1$. If transition to level $L+1$ will provide enough bandwidth to admit the call, the bandwidth is borrowed, the level is incremented, and the call is admitted; otherwise, the call is blocked. When the cell is operating at level $L = \lambda$, no more borrowing is allowed. It is important to note that our scheme never borrows from CBR connections or from connections that have already lost more than L shares.

Every time bandwidth becomes available in a cell due to a connection releasing its bandwidth allocation, the cell will attempt to make a transition to the next level. As a result, the available bandwidth is returned to the connections that have lost bandwidth due to borrowing. All fluctuations in a connection's allocated bandwidth are gradual as only one share can be borrowed or returned at a time.

Handoff management

The handoff admission policies differentiate between Class I and Class II connections. The reserved bandwidth is used only for Class I connections, which are admitted only if their minimum bandwidth needs can be met. When a Class I connection requests admission into a cell as a handoff, the cell checks to see if the minimum bandwidth requirement can be met with the sum of the available free and reserved bandwidth in the cell. If such is the case, the call is admitted into the cell and given bandwidth from the pool of free bandwidth up to its desired level minus L shares. The connection is given bandwidth from the reserved bandwidth pool only if its minimum requirement cannot be met using free bandwidth. If the minimum cannot be met using the free and reserved bandwidth together, the cell tests to see if scaling to level $L + 1$ would free up enough bandwidth to admit the call. If so, the cell scales the other calls in the cell and provides the handoff call with bandwidth according to the guidelines described above.

On the other hand, Class II traffic will be dropped only if there is no free bandwidth left in the cell at all. The reserved pool is not available to these connections, because, as in [2], we assume that Class II traffic is able and willing to incur a possibly substantial fluctuation in service rather than be disconnected. Calls that have suffered a lowering of bandwidth due to a handoff will eventually be brought back to a reasonable level as their new cell has free bandwidth to give them. This is in sharp contrast to the schemes presented in [2], which have no facility to improve connections

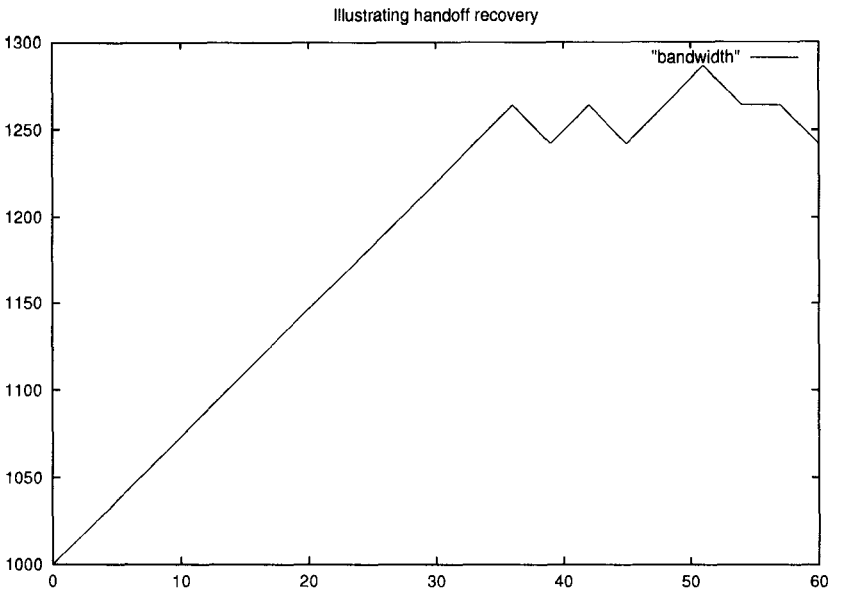

Fig. 4. The recovery of a Class I handoff.

which have been degraded due to a handoff.

How well does a handoff do?

Recall that a handoff Class I connection may be cut down to its minimum in order to avoid dropping the call. In addition, our scheme specifically disallows borrowing from connections that are below the cell level L. When bandwidth becomes available, our scheme attempts to bring all Class I connections to the cell level L. In particular, this means that handoff connections are expected to *recover* from a bandwidth loss incurred at handoff time.

Figure 4 illustrates this recovery process by plotting the bandwidth allocated to a Class I handoff connection over time. At time 0 the connection is admitted into the cell at its minimum acceptable level. In roughly 35 time units (seconds, in our simulation) the bandwidth has been replenished to the cell level. It is evident that the connection has reached the cell level L at the point when its bandwidth begins to fluctuate, indicating that borrowing has resumed [1].

[1] In Figure 4 this shows as a small decrease in bandwidth.

Complexity and Overhead of RBBS

Unlike Oliviera's scheme, all decisions in RBBS are made locally, within a base station, and do not require any communication with other base stations. In Oliviera's scheme, messages must be passed between a cell and all its neighbors each time a mobile host arrives and leaves a cell. When a MH arrives, at least one message must be sent to all neighbors, and at least one message must be received from all neighbors in order to continue. When a MH leaves a cell, at least one message must be sent to the cell's neighbors. All schemes that require information from other cells in order to make decisions in one cell, which are many of the statistical reservation schemes as well as all inter-cell borrowing schemes, suffer from this communication overhead.

The concept of shares in RBBS makes the scheme simple to implement. The size of a share for a given connection does not depend on what cell the mobile host is in, nor how busy the cell is (as we have shown in section III.2.2). Therefore, the amount of bandwidth available from a round of borrowing also changes only with the number of connections in a cell, not with the cell level. So it has to be computed only at the time a MH enters or leaves the cell, not with each round of borrowing, and that computation is, of course, trivial; the entering/departing MH's share size is added/subtracted from the current value of a round of borrowing.

With respect to the complexity of the adaptation needed by the application experiencing the changes in bandwidth, we have already mentioned some ways in which a multimedia application can be adaptive. With the gradual degradation and improvement afforded by the share concept, we make that task easier. And, later, in section V.2, we cover a few ways that the network could assume some of the work of adaptation from the mobile application.

Finally, we must consider the complexity of actually reapportioning the bandwidth and signaling the mobile host about changes in supplied bandwidth. Consider that the cellular network uses a TDMA (time division multiple access) channel access scheme. In TDMA, a channel is divided up into time slots and users are periodically assigned slots in which they can transmit or receive data. A set of slots forms a TDMA frame. These slots are assigned to users dynamically, with the assignment information being sent on slots permanently reserved for signalling. With dynamic assignment of slots on the frame, a better channel utilization can be achieved by allowing busy users to transmit during times when other users are idle [69]. So it is nothing new to vary supplied bandwidth dynamically. The reapportionment of bandwidth by RBBS

TABLE 1
Traffic Characteristics for the Simulation

CLASS	BANDWIDTH (Kbps)			DURATION (sec)		
	AVG	MIN	MAX	AVG	MIN	MAX
I	30	30	30	180	60	600
I	256	256	256	300	60	1800
I	3000	1000	6000	600	300	18000
II	10	5	20	30	10	120
II	256	64	512	180	30	36000
II	5000	1000	10000	120	30	1200

translates into a simple calculation of a new number of slots on a frame for each user, who is already prepared to discover when and how much to transmit dynamically.

III.2.4 Performance of RBBS

Simulation Model

To fairly contrast our scheme with other schemes from the literature, we have used the traffic types and characteristics given in [2] as input to our simulations, and modeled traffic behavior just as described there, with the exception of the handoffs. In [2], a handoff occurs during a connection with some given probability, and that probability decreases exponentially with each successive handoff during the connection. We have chosen a different approach that seems more realistic. We give each MH a speed characteristic specifying the amount of time that will be spent in each cell during a call. Thus, longer calls are likely to experience more handoffs than shorter ones.

As in [2], the traffic offered to the cellular system is assumed to belong to two classes:

1. Class I traffic – real-time multimedia traffic, such as interactive voice and video applications;

2. Class II traffic – non real-time data traffic, such as email or ftp.

Table 1 shows the exact characteristics of the traffic used in our simulations. Each of the six types occurs with equal probability. This set represents an estimation of what the traffic mix will look like in a future multimedia wireless network [2].

Experimental Results

In order to evaluate the performance of our rate-based borrowing scheme, we implemented and simulated two other schemes for comparison. First, we implemented a request-based statistical reservation scheme from [2], termed the uniform and bandwidth-based model. According to this scheme, when reservations are made on behalf of a connection in neighboring cells, an equal amount of bandwidth is reserved in each neighboring cell, with no consideration of the most likely cell to which the host might travel. A cell does not reserve the sum of all the bandwidth it is asked to reserve, but just the largest of all the current requests.

We also simulated a simple scheme that reserves 5 percent of the total bandwidth in each cell for handoffs. New calls are admitted into the network if their desired bandwidth can be met; otherwise, they are blocked. Class I handoffs are admitted if at least their minimum bandwidth requirements can be met. They are given only enough bandwidth from the reserved pool to meet their minimum, if there is too little free bandwidth available. Class II handoffs are admitted if there is any free bandwidth in the cell.

To simulate our rate-based borrowing scheme, we used a fixed reservation pool representing 5 percent of the total bandwidth. We set f to 0.5, thus permitting borrowing up to half of the bandwidth_loss tolerance. And we set λ to 10, so that each call had 10 shares to give.

For the results shown in Figures 5-9, the speed of the MHs was set to that of a host spending from 1 to 15 minutes in a cell, with an average of 5 minutes per cell. Each cell had 30Mbps of bandwidth. The network was a hexagonal grid of size 6×6 consisting of 36 cells. Traffic was provided to each cell at the level being measured.

Figure 5 compares the values of bandwidth utilization for the request-based reservation scheme from [2], for a fixed reservation scheme with $r = 5\%$, and for our rate-based borrowing scheme with $r = 5\%$, $\lambda = 10$ and $f = 0.5$, so that, at most, half of a connection's bandwidth_loss tolerance can be borrowed. For the fixed reservation scheme and the rate-based borrowing scheme, at the maximum connection rate, the bandwidth utilization comes close to equaling the bandwidth outside of the reserved pool. The results for the request-based reservation scheme are worse than for the other two, because we did not implement a cap on the size of the reserved pool.

Figures 6 and 7 show, respectively, the CDP for Class I traffic alone and for Class I and II traffic combined. The borrowing scheme outperforms the other two

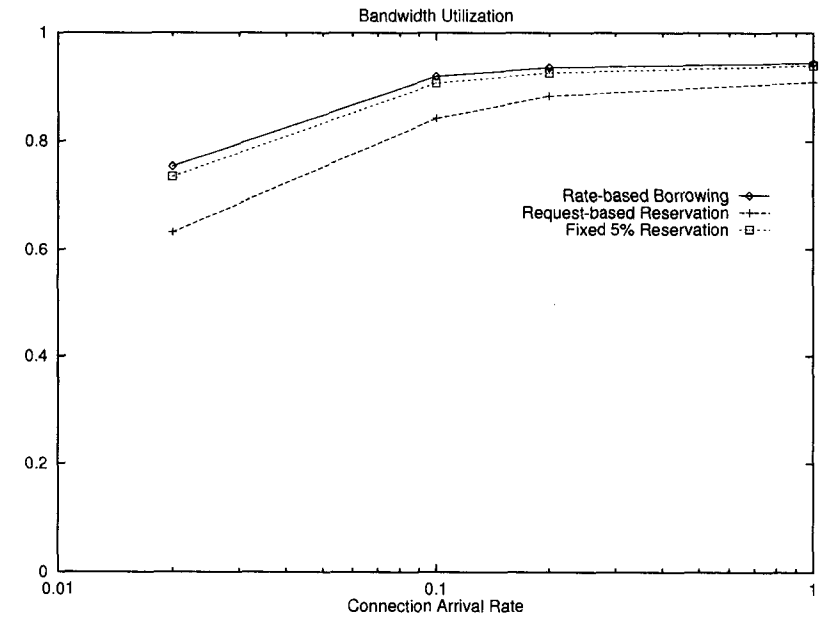

Fig. 5. A comparison of bandwidth utilization.

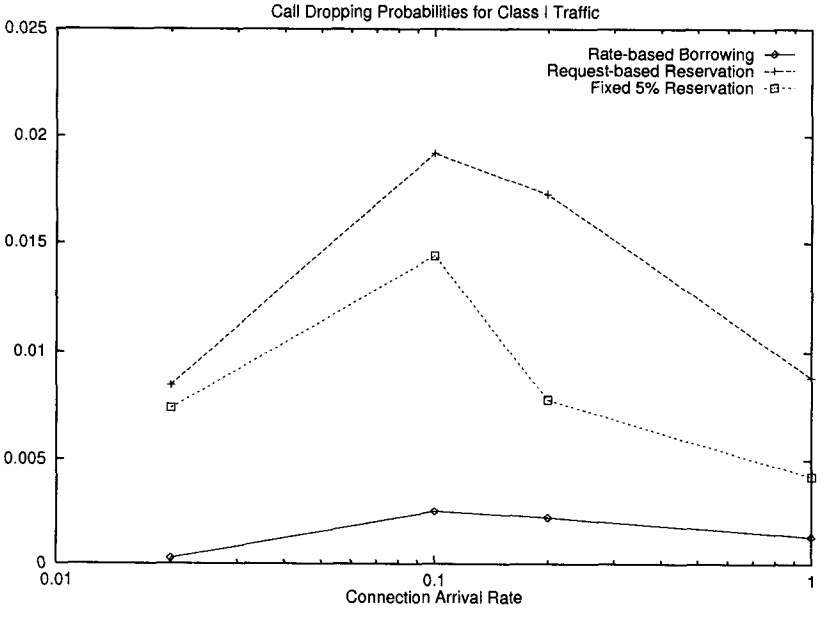

Fig. 6. Call dropping probabilities for Class I traffic.

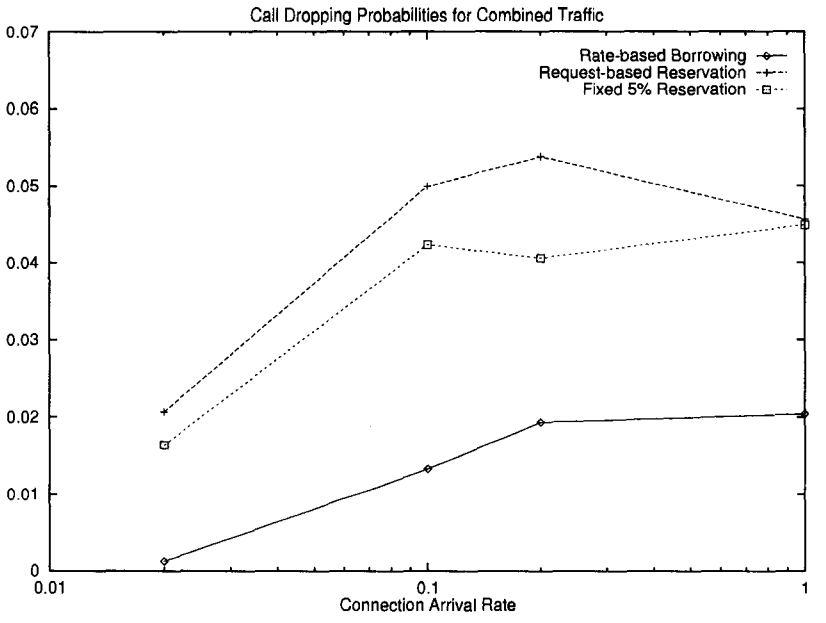

Fig. 7. Call dropping probabilities for Class I and Class II traffic combined.

schemes in both cases. In fact, the dropping probability for Class I connections is very close to zero. The motivation, of course, for favoring Class I connections by giving them exclusive use of the handoff reserves is that real-time connections would suffer an actual loss by being dropped. We assume that a Class II application, although inconvenienced by being dropped, would be able to resume its transmission at a later time, without any significant loss. Despite this, Class II traffic fares significantly better under our rate-based borrowing scheme than under the others; it is especially important that our scheme returns bandwidth to connections which have suffered cuts during a handoff.

Next, Figures 8 and 9 illustrate, respectively, the call blocking probabilities for Class I traffic alone and for Class I and II traffic combined. They demonstrate how borrowing allows a significant reduction in the CBP while also lowering the CDP. As it did in terms of CDP, the combined traffic also fares worse than Class I traffic alone in terms of CBP. However, this is not due to any bias in the algorithms, but rather to the characteristics of the traffic being simulated. The Class II traffic requires more bandwidth on average.

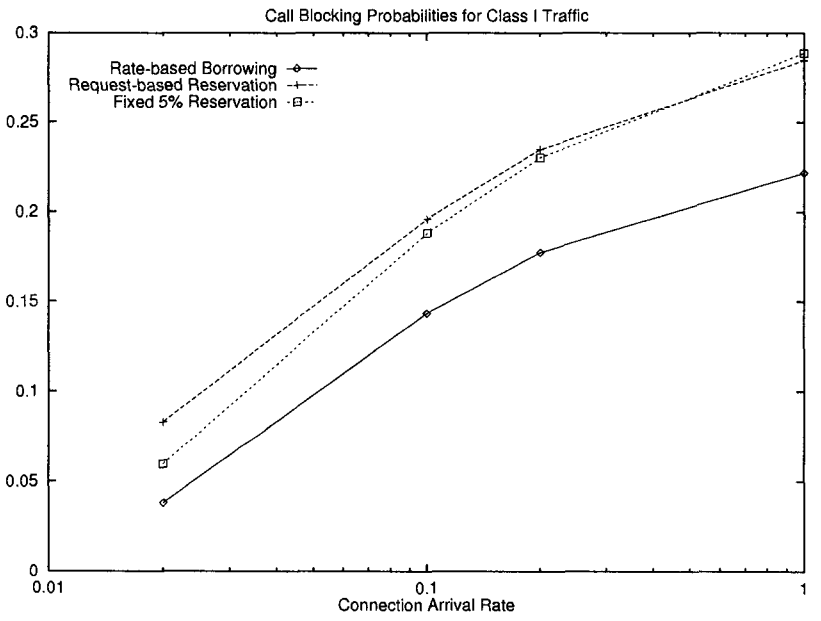

Fig. 8. Call blocking probabilities for Class I traffic.

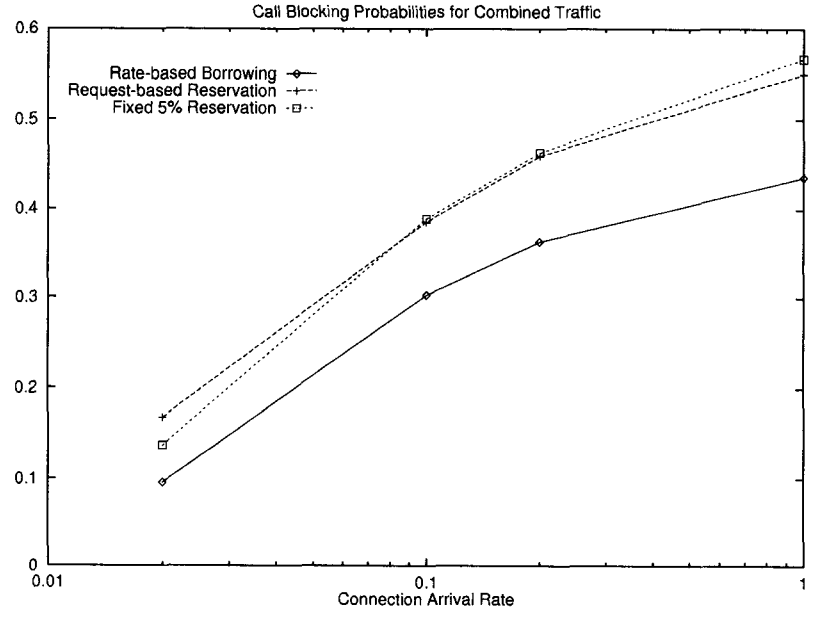

Fig. 9. Call blocking probabilities for Class I and Class II traffic combined.

CHAPTER IV

THE MAX-MIN FAIR BORROWING SCHEME

The second scheme we have developed is also a reservation and borrowing scheme, but it is based on max-min fairness, rather than the proportional, rate-based fairness described in the previous chapter. Max-min fairness is reviewed in the following section.

IV.1 MAX-MIN FAIRNESS

When the amount of bandwidth requested by the connections in a cell exceeds the total bandwidth available in the cell, it is unavoidable that some of the connections will receive less than their desired amount of bandwidth. It is, however, important that the bandwidth allocation be fair in some sense. One of the best-known fair allocation schemes is the classic max-min fairness. An allocation is max-min fair if there is no way to give more bandwidth to a connection without decreasing the allocation of a connection of lesser or equal bandwidth [70, 71].

As an illustration, consider a cell that has a total of 20 units of bandwidth. At some point there are four active connections A, B, C, and D using, respectively, 5, 4, 3, and 2 units of bandwidth. Referring to Figure 10, assume that a new connection E, requesting 7 units, has been accepted into the cell. Clearly, the total amount of bandwidth requested by these connections is 21 units, exceeding the capacity of the cell. How should the bandwidth be partitioned between the connections in a fair manner?

The key idea in max-min fairness is that if n connections need to partition b units of bandwidth, then each is guaranteed its equal share of $\frac{b}{n}$ units. Connections that require at most their equal share are granted their desired bandwidth and are referred to as *satisfied.* For definiteness, assume that of the n connections, m are satisfied and the total bandwidth requested by the m satisfied connections is S units. The residual bandwidth, $R = \frac{mb}{n} - S$, is partitioned among the remaining connections, each receiving $\frac{R}{(n-m)}$ units.

Notice that, in the current example, the addition of the new connection E brings the total number of connections to 5, and consequently, each is guaranteed 4 units

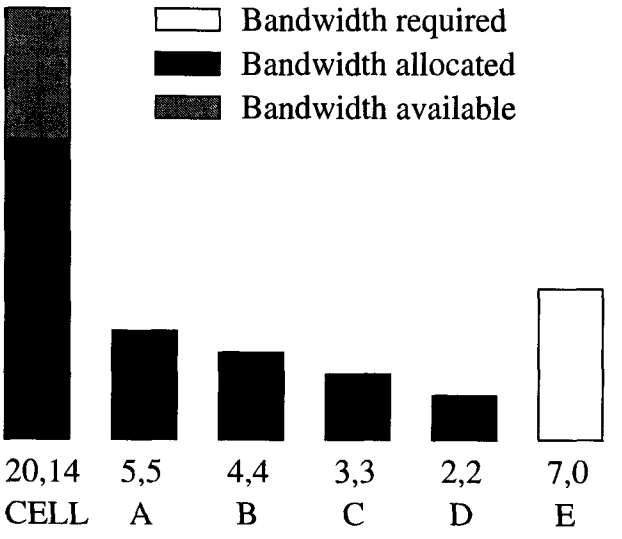

Fig. 10. Insufficient bandwidth to satisfy new connection.

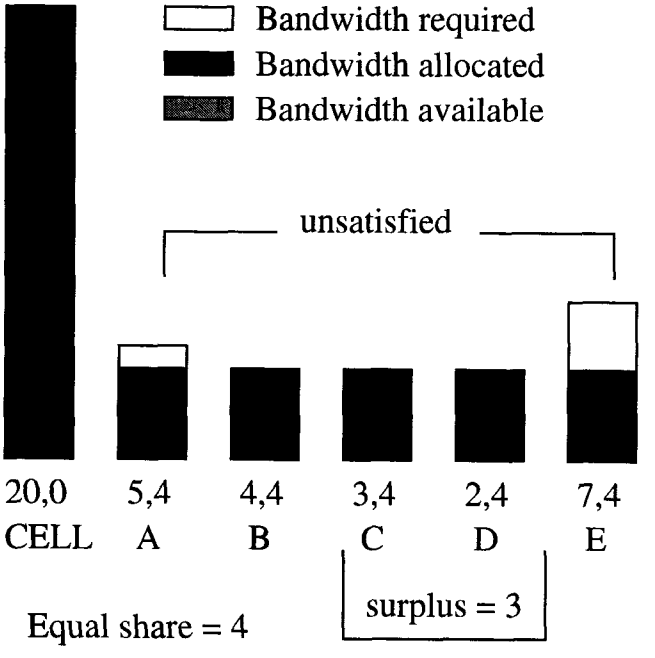

Fig. 11. Allocation of the equal share to all connections.

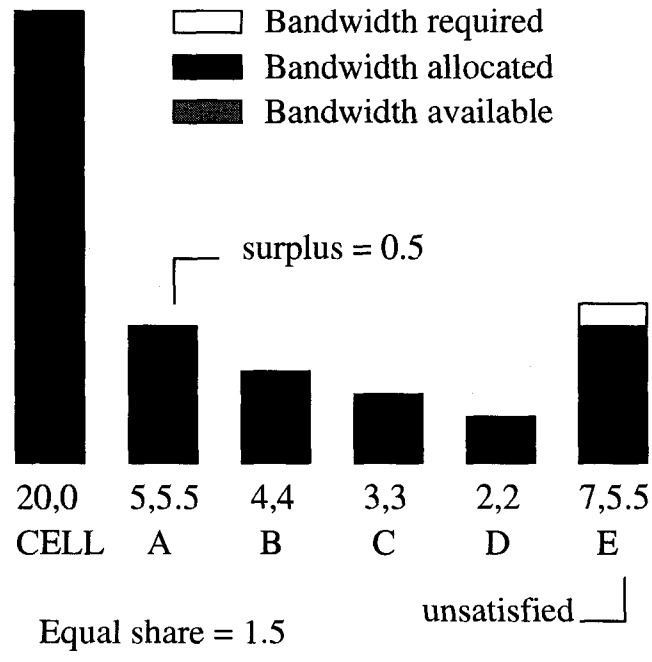

Fig. 12. Redistribution of the residual bandwidth.

of bandwidth. As illustrated in Figure 11, connections B, C and D are satisfied, while connections A and E are not. Since connections C and D are requesting less than their equal share, there are 3 units of residual bandwidth. As shown in Figure 12, the residual bandwidth is now partitioned between the unsatisfied connections A and E, each receiving 1.5 units of additional bandwidth. With this new allocation, connection A becomes satisfied, leaving E as the only unsatisfied connection.

Finally, the residual 0.5 units of bandwidth are now given to connection E. The final max-min fair bandwidth allocation is shown in Figure 13. Notice that connections A, B, C and D are satisfied and that the only unsatisfied connection is the relatively bandwidth-intensive connection E.

As this example shows, max-min fairness attempts to maximize the bandwidth allocation to the connections requesting the least amount of bandwidth. The max-min algorithm considers that each connection is entitled to an equal share of the limited bandwidth. Some connections request less bandwidth than others. In the end, connections with a low bandwidth demand receive their desired amount of bandwidth, thus becoming satisfied, while the remainder of the bandwidth is equally apportioned

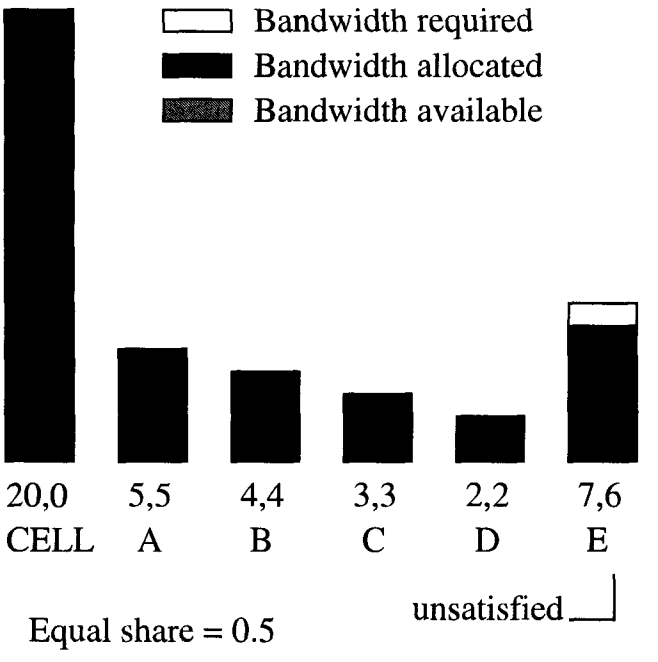

Fig. 13. Final redistribution of the residual bandwidth.

among the unsatisfied connections.

IV.2 MAX-MIN FAIR BANDWIDTH ALLOCATION

The max-min fair scheme keeps a pool of bandwidth reserved for handoffs, just as RBBS does. This reserved pool is used for Class I handoffs only, with the assumption that real-time connections have more stringent QoS requirements than Class II connections. It has been shown already that this reservation technique significantly reduces the CDP without adversely affecting the CBP. In order to also improve the CBP, our scheme allows for the temporary borrowing of bandwidth from existing connections in order to accommodate new and handoff connections.

The parameters that must be specified in a connection request are the connection class and the desired, minimum, and *expected* levels of bandwidth. The expected amount falls between the desired and the minimum and represents a comfortable working level for the application. In fact, in our scheme, the expected bandwidth is guaranteed to a connection while it remains in its initial cell. The difference between

the desired and expected amounts of a connection is the *actual borrowable bandwidth (ABB)*, and the cell may borrow some of this bandwidth from an existing connection in order to accommodate other incoming connections.

A new connection, whether Class I or Class II, is accepted into a cell only if its expected bandwidth is less than or equal to the total bandwidth of the cell (not including the reserved handoff pool) divided by the number of connections in the cell. This requirement is actually a key feature of our scheme, because it can cause a new connection to be blocked even if there is enough bandwidth for it in the cell, in order not to give it an unfairly large amount of bandwidth. If all the connections in a cell are functioning at their desired levels, and a new connection can also be accommodated at its desired level, it is simply admitted. If a connection cannot be given its expected amount using all the borrowable and free bandwidth in the cell, it is rejected. However, if it can be accommodated at its expected level, then the sum of the borrowable amounts of each connection plus any free bandwidth in the cell is divided equally among all the connections. Specifically, the equal share in a cell is determined by the formula below:

$$equalShare = \frac{totalBandwidth}{activeConnections + 1}$$

If a new mobile user requests a connection, the cell will accept or deny the connection based on the following algorithm:

```
if (expected bandwidth of new user <= equalShare)
    if (expected bandwidth of new user
            <= (borrowable bandwidth of all users
                + free bandwidth in cell))
        accept connection
    else
        block call
else
    block call
```

Handoff management differs for Class I and Class II connections. In the case of Class II connections, we assume that they want to be continued, even if the bandwidth allocated is very small, since they do not have stringent QoS requirements. Therefore,

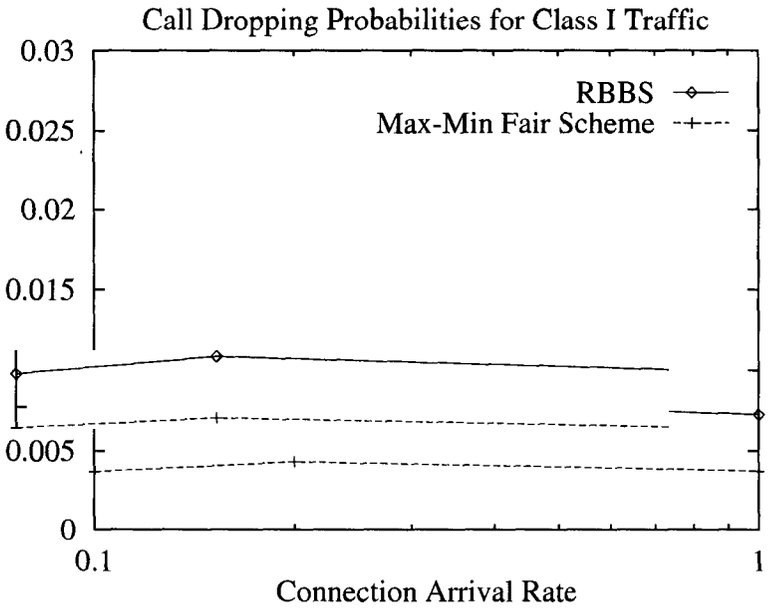

Fig. 14. Call dropping probabilities for Class I traffic.

Class II handoffs are not dropped as long as there is some free bandwidth in the new cell; clearly, their stated minimum may not be honored. The reserved bandwidth pool is for the exclusive use of Class I handoffs, and not available to Class II connections. However, a Class I handoff is given bandwidth from the reserved pool only if its minimum cannot be met using the borrowable and free bandwidth in the cell.

When a connection terminates, the freed bandwidth is used first to replenish the reserved pool, with the leftover allocated to connections that are functioning below their expected level. Finally, any residual bandwidth is distributed in a max-min fair manner among the connections that are functioning below their desired level.

IV.3 PERFORMANCE OF THE MAX-MIN FAIR BORROWING SCHEME

The simulation environment used for our max-min experiments is exactly the same one used for the rate-based experiments. In fact, the two algorithms were compared against each other directly in a simulator subjecting both to exactly the same traffic.

Figures 14 and 15 show, respectively, the CDP for Class I traffic alone, and for Class I and Class II traffic, combined. The results show that the max-min scheme

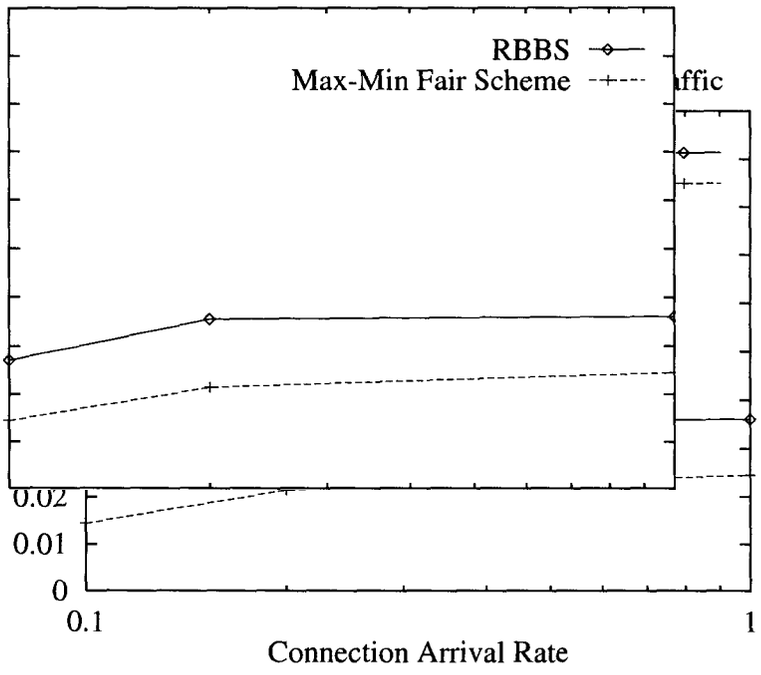

Fig. 15. Call dropping probabilities for Class I and Class II traffic combined.

outperforms the other scheme in both cases. By blocking slightly more bandwidth-intensive connections, our scheme makes more bandwidth available to both new and handoff connections. Somewhat surprisingly, the more strict call admission regimen of our scheme does not adversely impact the overall CBP, as illustrated in Figures 16 and 17. We note here that the significant difference in performance between our scheme and the proportionally fair scheme can be partially attributed to the traffic mix assumed in the simulation model. This mix features connections whose bandwidth requirements differ dramatically. However, this is anticipated to be a realistic scenario in future wireless multimedia networks [2].

It is important to note that all the desirable characteristics of the max-min scheme do not adversely impact the network bandwidth utilization. This is shown in Figure 18. The bandwidth utilization of the max-min fair scheme is just slightly worse than that of RBBS. We note here that the better performance of the max-min scheme in terms of CDP and CBP more than compensates for the slight degradation in bandwidth utilization. But as mentioned before, this comes at the price of denying admission into the system to those connections that are requesting an inordinate amount of resources. Since bandwidth-intensive connections have a high probability

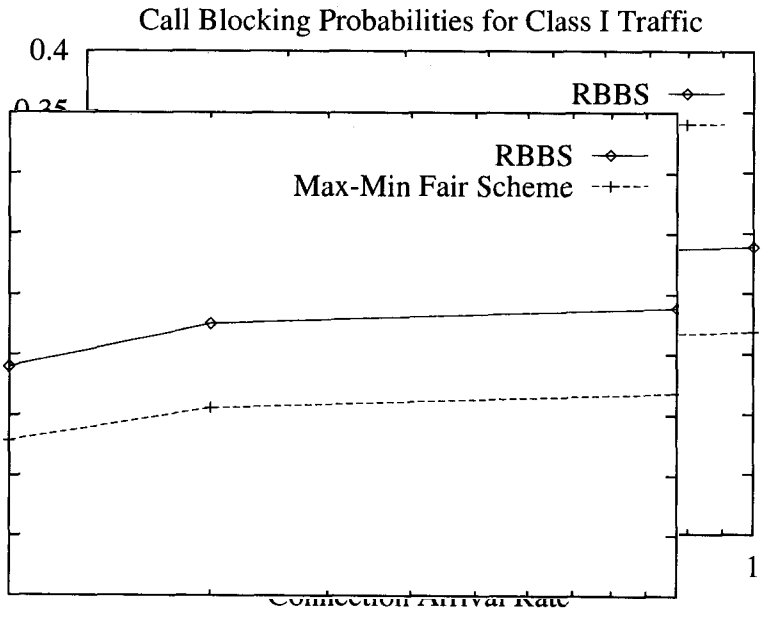

Fig. 16. Call blocking probabilities for Class I traffic.

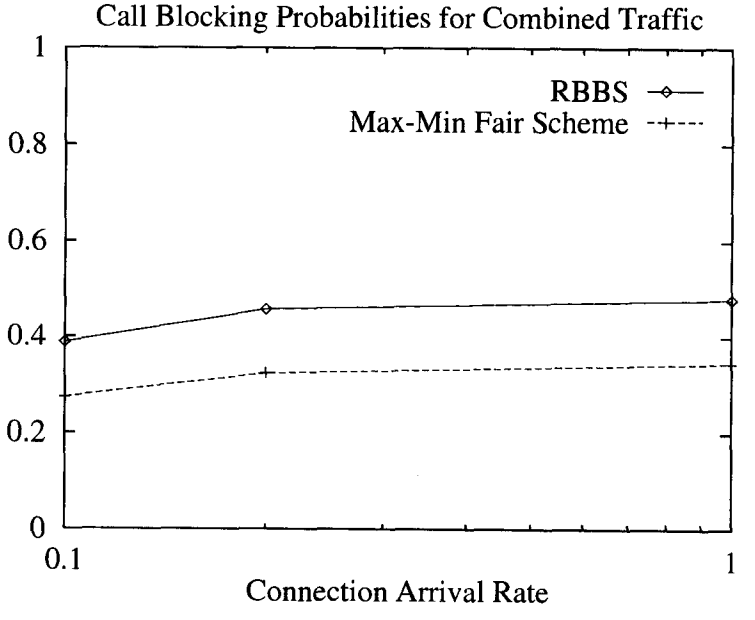

Fig. 17. Call blocking probabilities for Class I and Class II traffic combined.

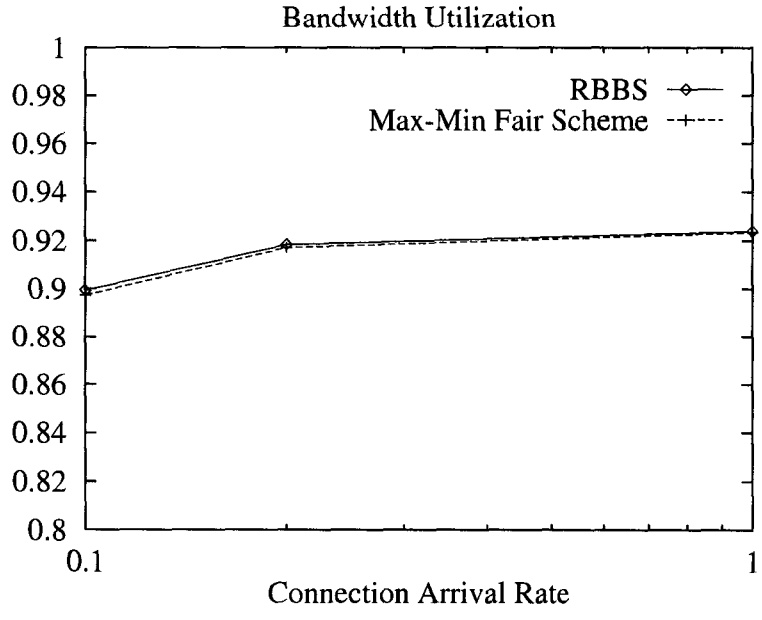

Fig. 18. A comparison of bandwidth utilization.

of being dropped in handoff situations, denying them admission is likely to improve both CBP and CDP.

One of the shortcomings of the max-min scheme is that users can be subjected to significant bandwidth fluctuations. While high-quality codecs can dampen the effects of these changes, it would be desirable to smooth out, to the largest extent possible, these bandwidth fluctuations. In this respect, the RBBS is superior to the max-min fair one, as it borrows and returns bandwidth in a graded manner. Unlike RBBS, in the max-min fair scheme, the amount of bandwidth available to be borrowed must be recomputed after each borrow as well as when a mobile host enters or leaves the system. And the amount of bandwidth to be borrowed from each mobile host must be computed at the time of borrowing. Thus, the max-min fair scheme is more computationally complex than RBBS.

CHAPTER V

THE MULTIMEDIA SESSION-AWARE SCHEME

Now that email, web browsing and SMS or instant messaging are very common applications available to cellular network users from their handheld devices, the most attractive new applications are surely video telephony, streaming multimedia and interactive games. These types of applications, along with more complex ones such as video conferencing with multiple video streams, slides and a whiteboard, usually require a set of multiple connections working together (e.g., video and audio) in a multimedia session. As noted earlier, there has been significant work done creating protocols that specify and control multimedia sessions, as well as that supply end-to-end QoS to them. Since this type of traffic already exists on cellular networks and will become more common, we propose that admission control and resource allocation schemes for the wireless link must have knowledge of the existence of these sessions and their attributes, and that this knowledge can be used both to improve call-level QoS and reduce processing overhead in the wireless link.

Putting multimedia session awareness into the admission control and resource allocation strategies is important because the relationship between connections has a direct bearing on the data that these strategies use to make their decisions. For example, suppose there is an application which uses four cooperating streams. An admission control scheme without session awareness might accept three of the component streams of the session into the network, but then deny the fourth stream. The three connections would get set up, only to have to be torn down when the fourth connection is denied and the application realizes it cannot continue. In a wireless environment, where resources are scarce, such a scenario could cause other users to be denied service unnecessarily. It would also cause unnecessary processing overhead.

These are examples of two of the three major benefits of our multimedia session-aware scheme (MSAS):

- Eliminate the overhead of provisioning a partial set of individual streams in the case where the whole session must be blocked or dropped.
- Eliminate the possibility that the condition of having resources reserved for streams that are destined to be terminated by their application (as in the example above) will cause other users' calls to be denied service.

- Create a higher probability for a session to survive a handoff by prioritizing among the streams in a session.

MSAS is built upon an updated version of the rate-based borrowing scheme. In order to provide QoS to sessions, rather than individual connections, the information about the session must be supplied to the network at the initial service negotiation. The information necessary for our schemes can be provided by the SDP session description provided in the SIP protocol.

In section V.1, we will describe MSAS in detail. In section V.2, we will discuss how we could use the session description information in combination with a media gateway at the MSC in order to improve our scheme's performance. Then in section V.3, we introduce the new QoS metrics we will be using to analyze the scheme's performance, and finally, in section V.4, we will discuss the simulation model and the simulation results.

V.1 MSAS DETAILS

The multimedia session-aware scheme treats each service request similarly, regardless of whether the request is for one or for multiple streams. At a minimum, MSAS needs the following information about each individual stream that is part of a request for service from the network:

- the stream type - class I or class II,
- the desired bandwidth,
- the acceptable bandwidth,
- the minimum bandwidth, and
- whether or not the stream (which belongs to a session) can be temporarily suspended in order to prevent the session from being dropped.

We do not consider the possibility that streams that are not part of a session, for example a class II connection for email, might also elect to be temporarily suspended in order to avoid being dropped, which would become something akin to queuing handoffs. For now, we consider that handoff decisions are made immediately; therefore only streams that belong to a session can agree to suspension, because the session

itself is continued. An example of using suspension would be if in a video phone call, the user was willing to temporarily suspend the video and continue with only the audio rather than get dropped during a handoff.

We update the rate-based borrowing scheme slightly from the version described in chapter III. In the MSAS, the difference between each connection's desired and acceptable bandwidth is divided up into the shares used for borrowing, rather than using the parameter f. Thus, bandwidth cannot be borrowed from a class I connection that is operating at or below its acceptable level. In the event of a handoff of a single class I connection or multimedia session, two shares, rather than one, are borrowed from each class II connection that is available for borrowing, as this improves the overall CDP of the cell. A Class II connection operating at or below its minimum level will not be borrowed from. In the case of multimedia sessions, the individual streams are borrowed from, according to their individual attributes.

Recall from chapter III that a constant bit rate (CBR) connection has an *adaptivity* value of 0. Note that an individual stream may have no adaptivity, but, in terms of the session it belongs to, it may be of low priority and thus be suspendable. A session's adaptivity is the ratio between the sum of the bandwidth loss tolerances of its non-suspendable connections and the sum of the desired bandwidth levels of all its connections. The higher the adaptivity level of a session, the lower the probability that it will be dropped during a handoff. A component connection will only be suspended in the event that there is no other way for the session to avoid being dropped.

Admission Control

For traffic types that consist of just a single connection, MSAS works exactly the same as the RBBS. In the case of sessions, in order to admit a new session to the network, bandwidth must be available to meet the desired level to each of the session's streams. If the cell has already undergone some borrowing, the desired bandwidth of each stream is scaled down to the appropriate cell level. If there is not enough bandwidth to accommodate the new connection, a check to see if a round of borrowing – removing one share from each individual connection in the cell that can be borrowed from – will free up enough resources. If so, the borrowing is done, the new session's streams are scaled down to the new cell level, and the session is admitted. If not, the new session is blocked.

Handoffs

In the event of a handoff, class I connections may be scaled down to their negotiated minimum level, whereas class II connections may be scaled down to the minimum unit of bandwidth. Suspendable connections may be allocated no resources at all, temporarily. If one or more of a session's component streams are real-time (class I), then bandwidth from the reserved bandwidth pool may be used for this session, if needed. As noted earlier, if borrowing is done to help continue a session that contains a real-time stream, then two shares may be borrowed from class II connections (unless it would bring them below their negotiated minimum level). Bandwidth is allocated proportionally among the streams in a session at a level between their minimum and their scaled desired levels, depending on what is available. The reserved bandwidth is used only if it is essential to providing the sum of the minimum values of a session. A suspendable connection cannot be allocated less bandwidth than its negotiated minimum; if there is not enough to meet its minimum, it is allocated no bandwidth. Any extra bandwidth is divided among the other streams, up to their scaled desired levels.

Bandwidth Recovery

When bandwidth becomes free because a session or connection terminates or moves to another cell, it is first used to bring any suspended connections back up to their minimum levels. New calls are not blocked for the sake of suspended connections; we found that this made little to no difference to the average recovery time for a suspended stream. As long as there are no suspended connections in the cell, connections which are operating at a level below their acceptable level are replenished, one share at a time. Finally, once there are no connections operating below the cell level, freed bandwidth is used to return borrowed bandwidth and raise the cell's operating level.

V.2 MSAS IN AN END-TO-END ARCHITECTURE

In this section we will explain how our scheme could fit into a larger framework of end-to-end QoS using SIP, SDP, media gateways. The SIP and SDP protocols come from the IP world but are not constrained to running on IP networks. A session description in SDP consists of one session section and one or more media sections within that session section. This description language, which promises to be more

robust in SDPng, may be used to specify all the information that our local call-level QoS provisioning scheme needs. There may be one or more bandwidth specifications in the session section or in each media section. A bandwidth specification takes the form:

`b=<bwtype>:<bandwidth>`

The bandwidth type field is meant to explain what the bandwidth value field represents. Although only two values for *bwtype* are currently defined, which both represent a kind of maximum value, it is with this specification that the bandwidth values for our scheme (desired, acceptable and minimum) could be defined. Attribute lines may be contained in both the session section and the media section as well. An attribute specification could be used to set our scheme's suspendable flag for a stream. An attribute line may take the following forms:

`a=<flag>`

`a=<attribute>:<value>`

In section II.3, we mentioned the QoS precondition model for SIP/SDP, which can be used to specify both local and end-to-end QoS requirements for a session. Preconditions are negotiated using the *offer/answer* model of SIP, in which SDP is used to carry the offers and the answers. SDP is also used to carry the information for negotiating the encoding algorithms and quality levels for the various media streams [72]. Consider a media stream that can be encoded/decoded at multiple quality levels. If that information were made available to our QoS scheme, we could scale such a stream by its predefined levels, instead of by the arbitrary *shares*.

Now, imagine that a media gateway is installed at the MSC. The media gateway would be informed by SIP/SDP about the formats of the various streams passing through it from and into the network backbone. A base station could signal the MSC if connections belonging to mobile hosts in its cell have their supplied bandwidth changed, whether due to handoff or to borrowing. For received streams, the MSC might be able to adjust the stream to fit the new bandwidth level, maybe by changing the encoding (if the receiver's codec can make the same adjustment), or by intelligently discarding packets. For received streams that get suspended, the media gateway might just discard the data until the stream's bandwidth is restored. For send or send/receive streams, the mobile host might also have to be signaled in order for it to adjust to changes in its supplied bandwidth.

SDP can supply the information necessary for our session-aware scheme. The

offer-answer model of SIP with SDP provides a way to negotiate QoS with a client and to send updates to the status between the network and client. A media gateway also fits nicely into this framework, and provides a way to move some of the effort of adaptation away from the client application. Finally SIP with SDP has the ability to negotiate both local and end-to-end QoS, so MSAS could be seen as being the local component of, or cooperating with, an end-to-end architecture.

V.3 ANALYSIS

Section II.1.2 illustrated that an increasing number of QoS provisioning schemes, including our own, exploit the adaptability of certain types of traffic in order to gain performance in parameters like CBP and CDP. Some schemes, like ours, borrow from all types of traffic that will tolerate it, but treat best-effort traffic a little worse; other schemes remove bandwidth only from best-effort type traffic, sometimes even terminating on-going connections. Decreased bandwidth, fluctuations in bandwidth, and forced termination are all negative effects caused by schemes attempting to improve QoS. In order to correctly analyze this class of schemes, both the positive and negative effects of a scheme on the traffic should be measured. We have defined two new QoS parameters that both reflect the supplied bandwidth over the life of a call, but from different viewpoints.

The first is the percentage of the Desired Bandwidth actually Supplied over time (DBS). For every bandwidth level at which a connection operates during its lifetime, let b_i be the bandwidth supplied to the connection between time t_i and t_{i+1}. The weighted bandwidth B_i is $b_i(t_{i+1} - t_i)$. The average weighted bandwidth is the sum of all the weighted bandwidth values divided by the total amount of time. Finally, the DBS is the average weighted bandwidth divided by the connection's desired bandwidth value. The second metric is the percentage of Time spent operating at the Desired Bandwidth (DBT). In this case, for each connection, we sum the time intervals $(t_{i+1} - t_i)$ for every b_i, where b_i is equal to the connection's desired bandwidth, which yields the total time spent operating at the desired bandwidth level. We divide that value by the actual call length to get the DBT.

Together these two metrics provide a picture of the QoS a connection receives with respect to supplied bandwidth. For CBR traffic, which has no adaptivity, both the DBS and DBT will, of course, always be 100 percent. The amount of adaptation to which a certain traffic type agrees will clearly affect the DBS and DBT. But the two

metrics provide a means to compare different schemes that deal with the same traffic. Schemes with more aggressive bandwidth borrowing or redistribution techniques may do well on CDP, but poorly on these metrics. Analysis that takes all these factors into account will assist the designer in coming to the right balance for her network requirements.

We also analyze MSAS by measuring the *fluctuation rate*, which we consider to be the rate of changes, both increases and decreases, caused by the scheme itself, not by mobility. Therefore, a bandwidth loss due to a handoff and the corresponding increases back up to the desired level are not counted as fluctuations.

V.4 PERFORMANCE OF MSAS

V.4.1 Simulation Model

In our simulation, we use as input a set of traffic types and characteristics similar to the one given in [2] and used for the simulating the rate-based and max-min schemes. Two of the traffic types were intended to represent video phone and video conference type traffic, but were treated as single connections by Oliviera, our own previous schemes, and by other researchers who have adopted the same traffic model [43]. Realistically, video phone and video conference traffic would be carried in multistream sessions. In order to simulate our new scheme, we have modelled the video phone type traffic as a session containing one audio and one video stream. The video conference type traffic is a session containing one large or high quality video stream, one small or low quality video stream, two audio streams, and two data streams. We have tried to keep the total bandwidth for the sessions close to the total bandwidth originally defined for those traffic types. Table 2 shows the exact characteristics of the traffic used in the simulations. The parameters have the same meaning they did in the experiments with the earlier schemes, and the traffic is introduced to the system in the same way - each traffic type occurs with equal probability. We have given each traffic type a designation for ease of reference. Types 1A-1C are real-time continuous (class I) traffic, and types 2A-2C are class II traffic. The table shows each traffic type on one line, with its name, duration values in seconds, and example application, and the stream(s) that compose it on the following line(s) with the stream class, and the bandwidth values. As with the rate-based scheme simulation, the desired bandwidth chosen for each connection is distributed around the average value, given

TABLE 2
Traffic Characteristics for the Simulation, with Sessions

NAME	**AVG DUR**	**MIN DUR**	**MAX DUR**	**EXAMPLE**
CLASS	AVG KBPS	MIN KBPS	MAX KBPS	
1A	**600s**	**300s**	**18000s**	**video conference**
I	2048	1024	4096	high quality video
I	224	192	256	low quality video
I	30	30	30	audio
I	30	30	30	audio
II	512	256	1024	data
II	512	256	1024	data
1B	**300s**	**60s**	**1800s**	**video phone**
I	224	192	256	low quality video
I	30	30	30	audio
1C	**180s**	**60s**	**600s**	**audio phone**
I	30	30	30	
2A	**120s**	**30s**	**1200s**	**ftp**
II	5000	1000	10000	
2B	**180s**	**30s**	**36000s**	**rlogin, http**
II	256	64	512	
2C	**30s**	**10s**	**120s**	**paging, fax**
II	10	5	20	

the minimum and maximum. The minimum bandwidth value is taken directly from the table. The acceptable bandwidth value is chosen to be the midpoint between the desired and the minimum.

The general attributes of the network are the same in this set of simulations. A fixed reservation of 5 percent is kept for handoffs. The simulated network consists of a six-by-six honeycomb of 36 cells. Traffic is generated in each cell at a constant rate in a geometric distribution around the connection rate being measured. Each cell has 30Mbps of bandwidth. Mobile hosts move at a speed varying between 1 and 15 minutes per cell, with an average of 5 minutes per cell. The length of stay of a MH in a cell changes as it moves from cell to cell. Mobile hosts move in a slightly directed manner, where there is a higher probability of moving forward than to the sides or remaining in the same cell, and an even lower probability of moving backward. If

a mobile host reaches the edge of the grid and has to continue ahead, its session is ended, as though it has left the network. The number of shares, λ, is 10.

When simulating the rate-based scheme and its comparison schemes, because we were comparing with an existing scheme, we measured CDP as the number of drops over the total connections arriving in a cell, and we measured CBP as the total blocks over the total connection requests. In simulating the session-aware scheme, we are defining CDP to be the total drops over the total handoff requests and the CBP as the total blocks over the total new call requests.

V.4.2 Experimental Results

In order to evaluate the performance of MSAS, we have implemented a scheme that takes the same input and uses the same borrowing and reservation techniques, but treats each stream as though it has been introduced to the network as an individual connection, with no knowledge that the connections are related. It measures its performance with respect to the sessions, in order to be comparable. In the case of a new call that consists of separate streams, it is possible that the negotiation of the individual streams will be interleaved with the negotiation of handoffs in that same cell. They are not interleaved with other new calls. When a session attempts a handoff, the streams are considered individually, but are not interleaved with any other requests to the cell. We noticed that this session-unaware scheme (SUS) suffers from excessive borrowing because a round of borrowing can be performed for each component stream of a session. So we also implemented a session-unaware scheme that allows only one round of borrowing per mobile host in a cell, essentially one per session, like MSAS. This scheme is called the session-unaware scheme with limited borrowing (SUS-LB). Finally, we ran a set of input to MSAS in which the video streams of the traffic types 1A and 1B are set to be suspendable, so that those sessions may temporarily cut their video streams in order to survive a handoff. These results are referred to as MSAS-S.

CBP and CDP

In Figure 19, we see call dropping probabilities for the combined traffic. MSAS-S has the best results by a slight margin, as expected, because the session types of traffic are able to be more flexible at a handoff. Session awareness lowers the CDP because the streams of a session do not compete against one another for bandwidth. SUS performs

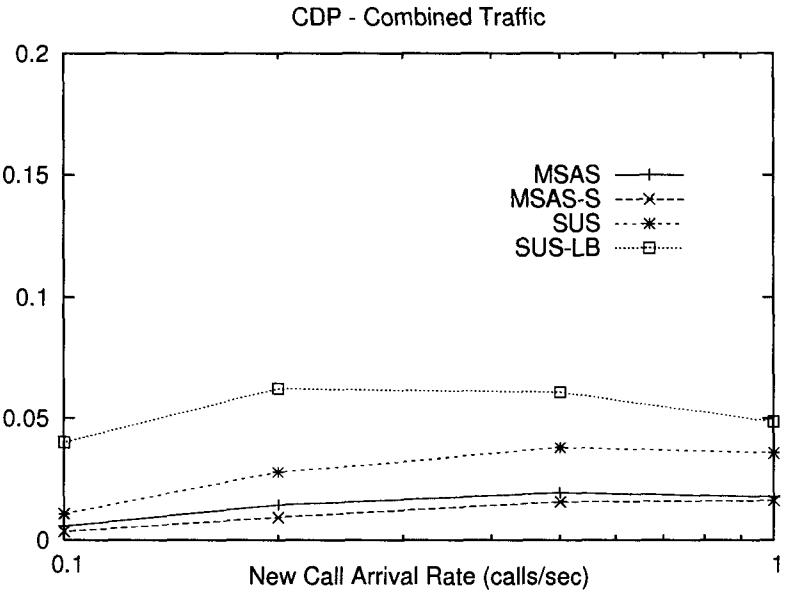

Fig. 19. Call dropping probabilities for the combined traffic.

better than SUS-B because borrowing can be done for each stream in a session. Due to the better CDP, the session-aware schemes experience a slightly higher CBP, as is shown in Figure 20. In fact, the ranking of each scheme with respect to CDP is reversed with respect to CBP, and this is expected.

Figure 21 shows the average bandwidth utilization achieved by the cells in the network. In a busy network, the bandwidth utilization should approach 95 percent, because of the 5 percent reservation for handoffs. SUS does just slightly better in this metrics, and SUS-LB just slightly worse, than the session-aware scheme.

When we examine the statistics for each traffic type separately, we find that the session-aware scheme produces very good results, not surprisingly, for session traffic. For traffic type 1A, which has 6 streams, Figure 22 shows the CDP and Figure 23 shows the CBP. The CBP for MSAS and MSAS-S is slightly worse, but the CDP is significantly better. In fact, MSAS-S produces a CDP of less than 0.2 percent in the highest connection rate case, whereas SUS-LB drops nearly 50 percent of the handoffs.

For traffic types 1B and 1C the CDP for all schemes is negligible. However, for types 2A, 2B and 2C, which are all non-real-time class II connections, the session-unaware schemes produce CDPs near zero at all rates, but the session-aware schemes

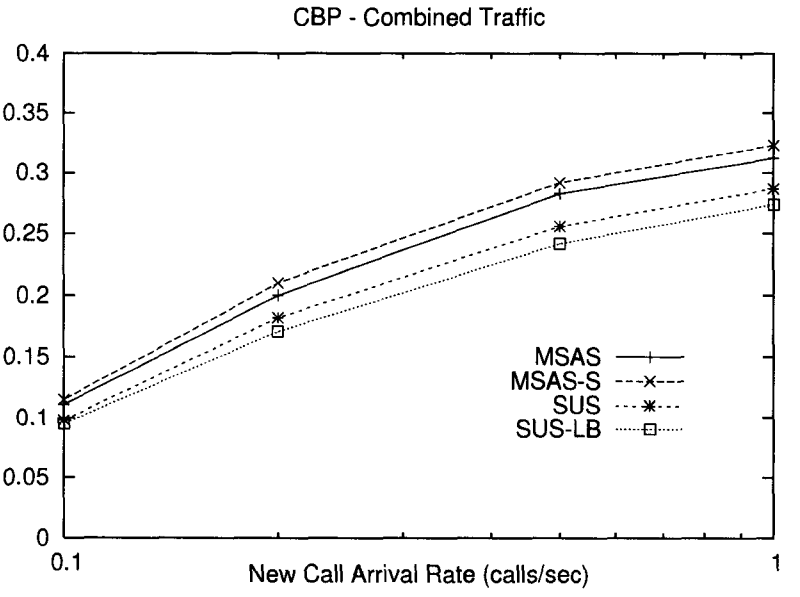

Fig. 20. Call blocking probabilities for the combined traffic.

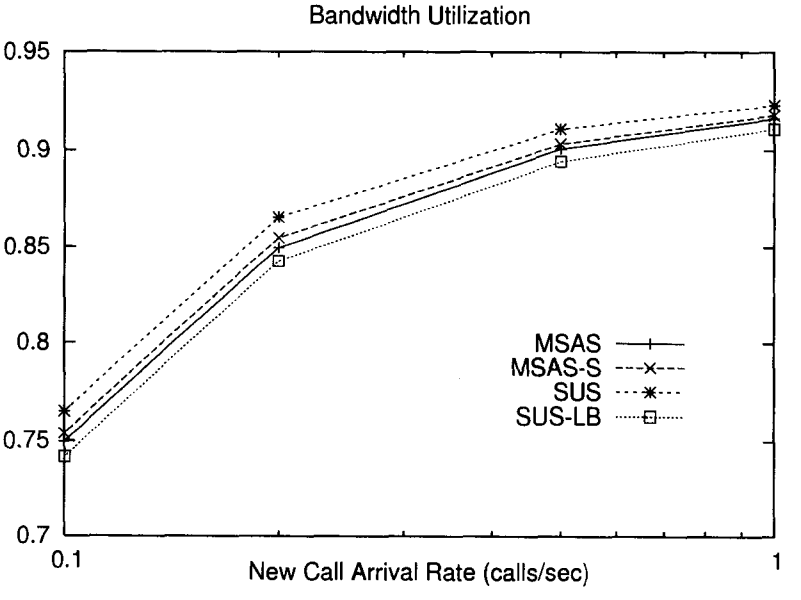

Fig. 21. A comparison of bandwidth utilization.

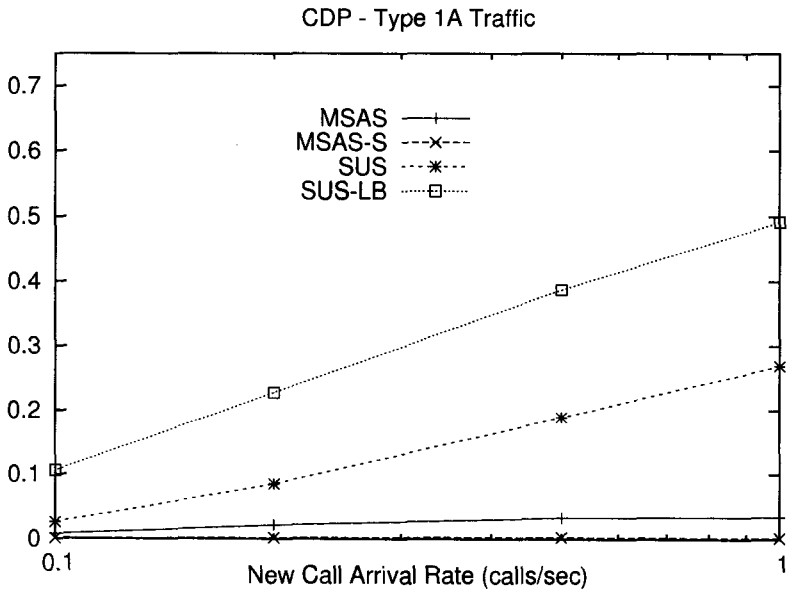

Fig. 22. Call dropping probabilities for traffic type 1A.

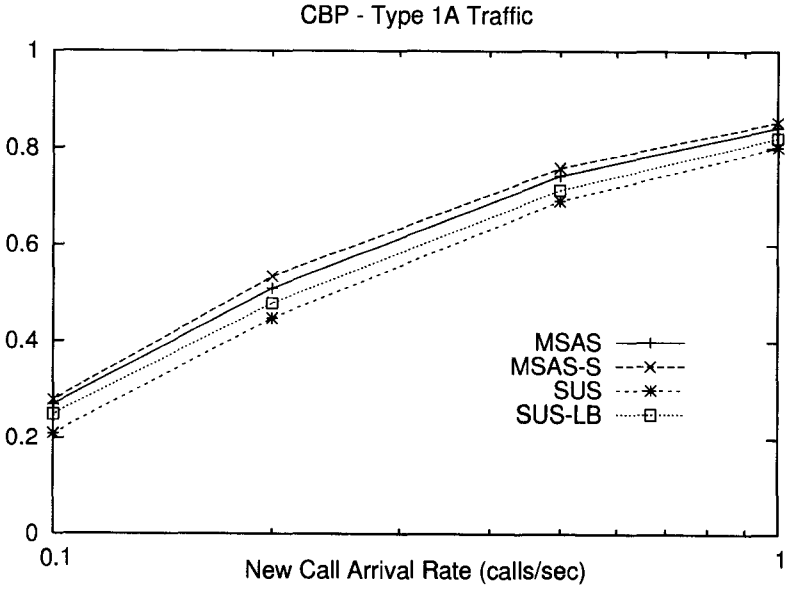

Fig. 23. Call blocking probabilities for traffic type 1A.

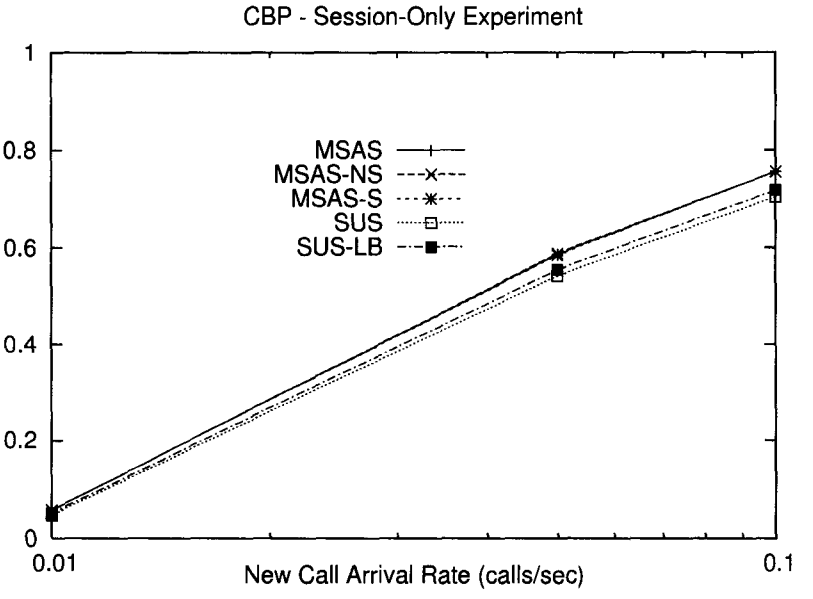

Fig. 24. Call blocking probabilities for the session-only experiment.

go up to around 5 percent at the highest connection rate.

The Session-Only Experiment

We also studied the performance of these schemes in an experiment where only session traffic (type 1A) was introduced to the network in order to isolate the effects of session awareness. We had to use lower connection arrival rates because the needs of this traffic type are so high. For this experiment, we ran three trials, one each for SUS, SUS-LB and MSAS, but in the MSAS trial, half of the sessions agreed that their video streams could be suspended and half did not. In this case, the label MSAS represents the results of the combined traffic in the session-aware experiment, and then, individually, the non-suspending half is MSAS-NS, and MSAS-S is the suspending half. Figure 24 shows the CBP, and Figure 25 shows the CDP. In terms of CBP, suspending makes no difference. The session-aware scheme does slightly worse than the unaware schemes. With CDP, the suspending sessions naturally do the best, making the combined MSAS results good too. MSAS-NS and SUS are almost the same in less busy networks, but in busy networks, MSAS-NS does better.

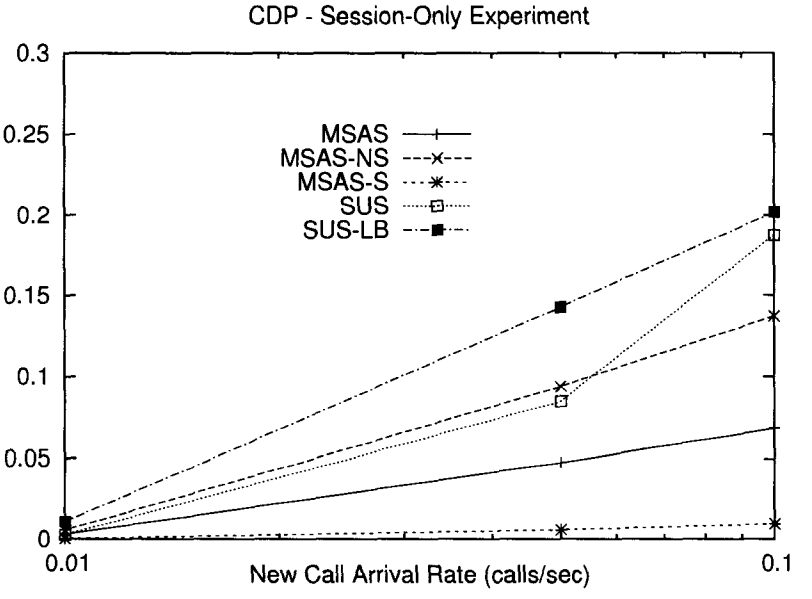

Fig. 25. Call dropping probabilities for the session-only experiment.

DBS and DBT

Now we will look at the results for DBT and DBS for some of the most adaptive traffic types. Figures 26 and 27 show the DBT and DBS values for traffic type 1A in the mixed traffic experiment, and Figures 28 and 29 show the same measurements for type 2A. These are the traffic types that require the most resources, one real-time and one non-real-time, and that have the highest given adaptivity.

With respect to DBT for traffic type 1A, the session with 6 streams, SUS-LB does the best - more time is spent at the desired bandwidth, and SUS does the worst, because of the excessive borrowing. SUS-LB does a little better than the other schemes on DBT for type 1A traffic because it does so badly on CDP for the same traffic type. Dropping a lot of handoffs means that it can provide slightly better service to the calls that it does keep in the system. SUS-LB also does the best on DBS, for the same reason. MSAS-S does the worst on DBS because, although it is not the worst at DBT, suspending streams and cutting their bandwidth to zero contributes to bringing down the average of the supplied bandwidth.

For type 2A traffic, the two session-aware schemes get the best results in both DBT and DBS. Suspending does not make a difference because this traffic type has

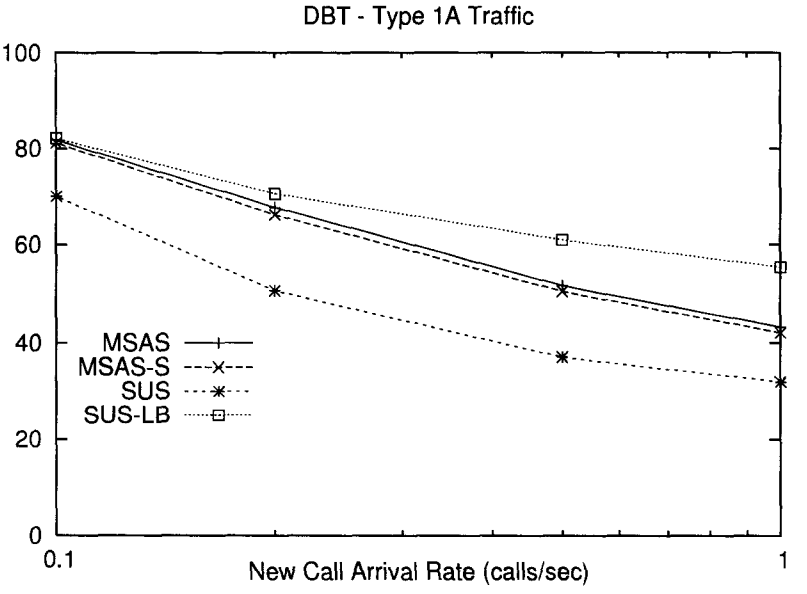

Fig. 26. Percentage of time at desired bandwidth - traffic type 1A.

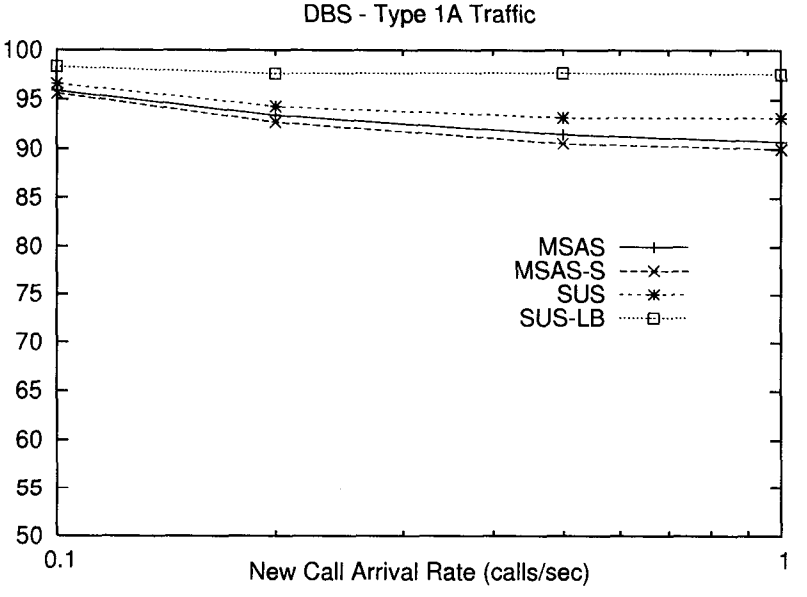

Fig. 27. Percentage of desired bandwidth supplied - traffic type 1A.

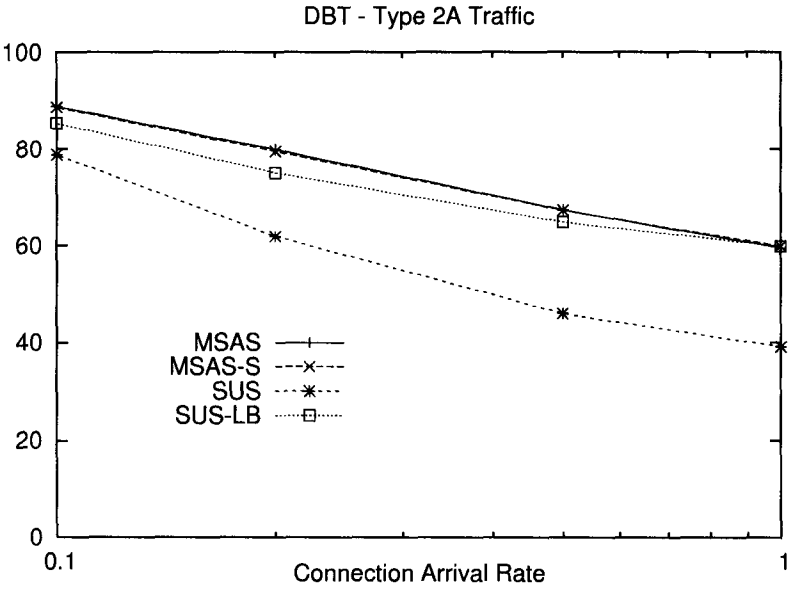

Fig. 28. Percentage of time at desired bandwidth - traffic type 2A.

just a single class II stream. SUS, of course, does the worst, because of the excessive borrowing, while SUS-LB gets almost the same results as MSAS. It is worth noticing how well this high-bandwidth, high-adaptivity class II traffic type does with respect to DBS despite the new practice of borrowing two shares from class II connections for the sake of a class I handoff.

Fluctuations in Bandwidth

Like any scheme which allows resource reallocation, rate-based borrowing has the side-effect of causing fluctuations in the supplied bandwidth of connections in the network. We have measured these fluctuations by traffic type since some traffic types do not allow borrowing. Figure 30 shows the number of fluctuations per second for type 1A, where a fluctuation is a change in bandwidth caused by borrowing or returning bandwidth that has been borrowed. It does not include a loss in bandwidth caused by a handoff or the replenishment of bandwidth lost in a handoff. The session-aware scheme has the lowest fluctuation rate, and the results for all traffic types that allow borrowing are similar. SUS has the highest fluctuation rate because it is capable of performing a round of borrowing for each individual stream in an arriving

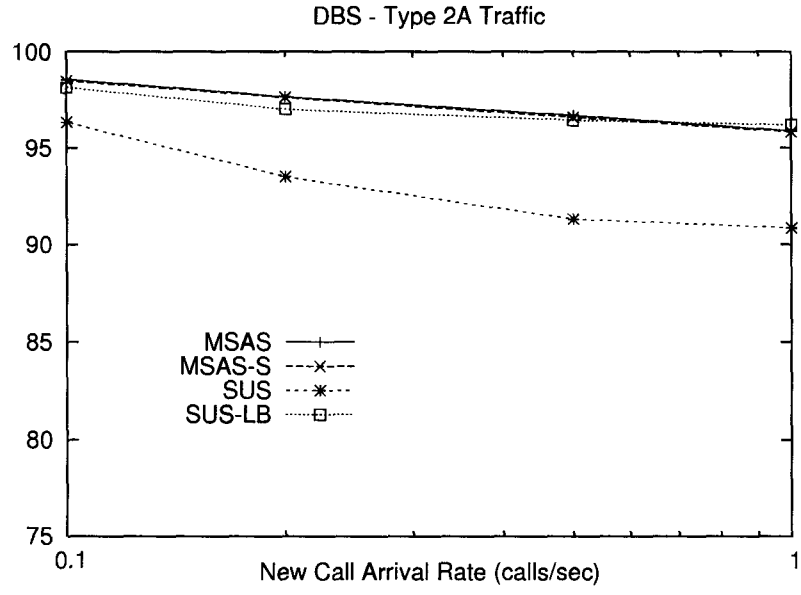

Fig. 29. Percentage of desired bandwidth supplied - traffic type 2A.

session, while the other schemes allow only one round of borrowing for the sake of an arriving session. SUS-LB has a higher rate than MSAS because sometimes a round of borrowing is performed for a session that will never completely make it into a cell.

Recovery of Suspended Streams

Simulation has shown that the practice of temporarily suspending a stream is very effective in lowering the dropping probability for multimedia session traffic. We measured the rate of recovery for suspended streams. In the full traffic mix experiment at the highest connection rate (one new connection per second per cell), suspended streams were brought back to their minimum level in an average of 17 seconds. The maximum time to recovery was 129 seconds and the minimum was 0, meaning that bandwidth was freed by another session and reassigned to this stream during the same simulation time step. As in all the other experiments, the video streams from traffic types 1A and 1B were set to suspendable - two of those streams have a minimum bandwidth of 192 Kbps, and the other's minimum level is 1 Mbps. The recovery from the minimum rate back up to the cell level or the desired level would be similar to what is depicted in section III.2.3.

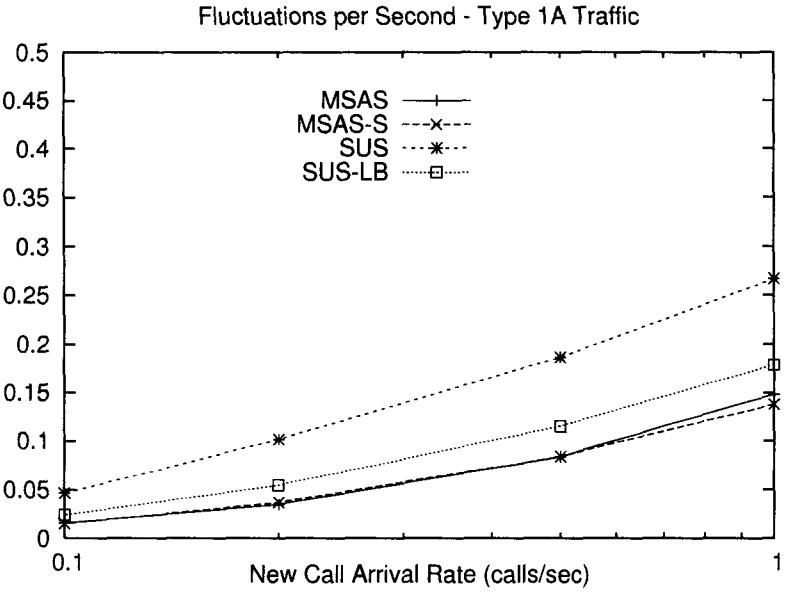

Fig. 30. Borrow-related fluctuations per second - traffic type 1A.

Overhead of Session-Unawareness

We claimed in the beginning of this chapter that a scheme which is not aware of the existence of multimedia sessions will waste network resources. It will incur a processing overhead by incorrectly admitting streams that belong to a session that cannot be supported in its entirety. If an admission control scheme is unaware of the relationship between streams in a session, then when one of the streams of a session gets blocked or dropped, the streams that may have already been admitted will have to be terminated. If n is the number of streams in the session, it is easy to see that up to n-1 wasted allocations/deallocations may occur for every dropped session. The actual number depends on the order the streams are introduced to the network and the resources they are seeking. In the case of a handoff, the streams in a session will essentially be competing with one another for network resources, as each will get as close to its desired bandwidth as possible.

Not only is there a processing and a communications overhead associated with the unnecessary setting up and tearing down of connections caused by session-unawareness, but there is also the problem of reserving bandwidth improperly. When bandwidth is assigned to a stream that is not destined to survive, other connections

may be blocked or dropped. We have measured the extent of this problem, and we refer to it as the unnecessary dropping rate (UDR). The UDR is computed by finding the total number of drops that occur due to bandwidth being held by streams that are going to be terminated when another stream in their session is blocked/dropped, and dividing that number by the total number of drops. In our simulation, we did not model the possibility of unnecessary blocks.

In the mixed traffic experiment, the UDR for SUS was less than 1 percent at all connection rates. For SUS-LB, the UDR was 6 percent regardless of the connection rate. The problem would naturally be more severe in the case where all the traffic in the network is multimedia sessions. In that experiment, the worst case for SUS, at the highest rate this time, was 14 percent. For SUS-LB, the UDR was 29 percent at the highest connection rate.

CHAPTER VI

THE PREDICTIVE RESERVATION SCHEME FOR LEO SATELLITE NETWORKS

Due to the high velocity of LEO satellites, in LEO systems, a connection established in one of the cells of a satellite's footprint is likely to experience a large number of handoffs during its lifetime as it passes through several spotbeams. Consequently, LEO satellite systems require sophisticated handoff management and call admission control protocols. We have proposed a predictive handoff management and admission control strategy for multimedia LEO satellite networks. A key ingredient of our handoff management scheme is an adaptive resource reservation protocol. Simulation results have confirmed that our scheme offers very low CDP, while at the same time it keeps resource utilization high.

Our predictive handoff management and call admission control strategy involves some processing overhead. However, as it turns out, these overheads do not affect the mobile hosts as all the processing is handled by the satellite. Consequently, the scheme scales to a large number of users.

VI.1 MOBILITY AND TRAFFIC MODELS

The LEO system model we have adopted for simulating our scheme is based on the well-known Iridium satellite system, where each satellite rotates around the Earth in a polar orbit [46]. We follow common practice and assume that the speed of individual mobile hosts is negligible with respect to the orbital velocity of the satellite [47, 49, 46]. Consequently, mobile host trajectories are straight lines. The footprint of a satellite is partitioned into spotbeams, each approximated by a regular hexagon.

In order to model mobile host behavior we make the following assumptions:

- Mobile hosts move at constant speed, essentially equal to the orbital speed of the satellite, and they cross each cell along its maximum diameter.
- The time t_s it takes a mobile host to cross a cell is $t_s = \frac{\text{cell\_width}}{\text{speed}}$
- A mobile host remains in the cell where the connection was initiated for t_f time, where t_f is uniformly distributed between 0 and t_s; thus, t_f is the time until the first handoff request, assuming that the call does not end in the original cell.

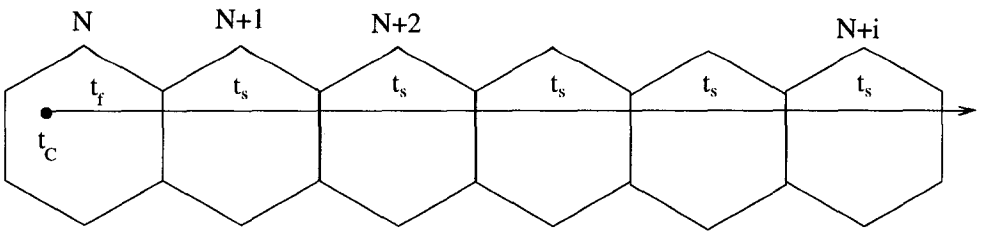

Fig. 31. The trajectory of a connection.

- After the first handoff, a fixed time t_s is assumed between subsequent handoff requests until call termination.

Referring to Figure 31, when a new connection is requested in cell N, it is associated with a *trajectory*, consisting of a list $N, N+1, N+2, \ldots, N+i, \ldots$ of cells that the connection may visit during its lifetime. For a generic call C, we let H be the random variable denoting the holding time of C. We assume that H is exponentially distributed with mean $\frac{1}{\mu}$.

Assume that C was accepted in cell N. After t_f time units, C is about to cross into cell $N+1$. Let p_f be the probability of this first handoff request. Clearly,

$$p_f = \Pr[H > t_f] = \frac{1}{t_s}\int_0^{t_s} e^{-\mu t}dt = \frac{1-e^{-\mu t_s}}{\mu t_s}. \tag{5}$$

Due to the memoryless property of the exponential distribution, the probability of the $(k+1)$-th, $(k \geq 1)$ handoff request is

$$\Pr[H > t_f + k\cdot t_s | H > t_f + (k-1)\cdot t_s] =$$

$$\frac{\Pr[H > t_f + k\cdot t_s]}{\Pr[H > t_f + (k-1)\cdot t_s]} = \frac{e^{-\mu(t_f + k\cdot t_s)}}{e^{-\mu(t_f+(k-1)\cdot t_s)}} = e^{-\mu t_s}$$

which, as expected, is independent of k. Consequently, we will let

$$p_s = e^{-\mu t_s} \tag{6}$$

denote the probability of a *subsequent* handoff request. It is important to note that t_f, t_s, p_f and p_s are mobility parameters that can be easily evaluated by the satellite using its on-board processing capabilities.

When a mobile host requests a new connection C in a given cell, it provides the following parameters:

- The desired class of traffic for C (either I or II).

- M_C, the desired amount of bandwidth for the connection.

If the request is for a Class I connection, the following parameters are also specified:

- m_C, the minimum acceptable amount of bandwidth, that is the smallest amount of bandwidth that the source requires in order to maintain acceptable quality, e.g. the smallest encoding rate of its codec.
- θ_C, the largest acceptable call dropping probability that the connection can tolerate.
- $\frac{1}{\mu_C}$, the mean holding time of C.

VI.2 THE PREDICTIVE BANDWIDTH RESERVATION STRATEGY

Our handoff admission policies distinguish between Class I and Class II. As in our previously explained schemes and in [2], Class I handoffs are admitted only if their minimum bandwidth requirements can be met. However, Class II handoff requests will be accepted as long as there is some bandwidth left in the cell. Thus, bandwidth reservation pertains only to Class I handoffs. The basic idea is to reserve for each accepted Class I connection a certain amount of bandwidth in each cell along its trajectory.

Let p_h denote the handoff failure probability of a Class I connection, that is, the probability that a handoff request is denied for lack of resources. Let S_i denote the event that a Class I connection C admitted in cell N goes successfully through i handoffs and will, therefore, show up in cell $N+i$. It is easy to confirm that the probability of S_i is

$$\Pr[S_i] = p_f \cdot (1 - p_h) \cdot [p_s \cdot (1 - p_h)]^{i-1}. \tag{7}$$

Equation (7) suggests the following natural reservation strategy: in preparation for the arrival of connection C, an amount of bandwidth equal to

$$B_{N+i} = M_C \cdot \Pr[S_i] \tag{8}$$

will be reserved in cell $N+i$, $(i \geq 1)$, during the time interval $I_{N+i} = [t_C + t_f + (i-1)t_s, t_C + t_f + it_s]$, where t_C is the time C was admitted into the system.

It is worth noting that our reservation scheme is lightweight: since the mobility parameters t_f and t_s are readily available, and since the trajectory of connection C is

a straight line, it is a straightforward task, for every $i > 1$, to compute the amount of bandwidth B_{N+i} to reserve as well as the time interval I_{N+i} during which B_{N+i} must be available.

The bandwidth reservation strategy discussed above is meant to ensure that the parameter p_h is kept as low as possible. We emphasize here that p_h and CDP are *not* the same: while p_h quantifies the likelihood that an arbitrary Class I handoff request is denied, CDP has a long-term flavor denoting the probability that a Class I connection will be dropped at some point during its lifetime. It is straightforward to show that

$$\begin{aligned} \text{CDP} &= p_f \cdot p_h + p_f \cdot (1 - p_h) \cdot p_s \cdot p_h + \cdots = \\ &= p_f \cdot p_h \sum_{i=0}^{\infty} [p_s(1 - p_h)]^i = \frac{p_f \cdot p_h}{1 - p_s(1 - p_h)} \end{aligned} \tag{9}$$

where the first term in (9) is the probability that the call will be dropped on the first handoff attempt, the second term denotes the probability that the call will be dropped on the second attempt, and so on.

We now point out a strategy for ensuring that for an accepted Class I connection C the negotiated CDP is maintained below the specified threshold θ_C. Thus, the goal is to ensure that CDP $< \theta_C$. By (9) this amounts to insisting that

$$\frac{p_f \cdot p_h}{1 - p_s(1 - p_h)} < \theta_C.$$

Solving for p_h we get

$$p_h < \frac{\theta_C \cdot (1 - p_s)}{p_f - \theta_C \cdot p_s}. \tag{10}$$

All the quantities in the right-hand side of (10) are either specified by the connection or can be determined by the satellite from the mean holding time of C by using equations (5) and (6). Thus, in order to enforce the CDP commitment, the satellite keeps track of the handoff failure probability p_h for Class I connections. If p_h is close to the value of the right-hand side of (10), new calls are temporarily blocked.

VI.2.1 New call admission strategy

Our new call admission strategy involves two criteria. The first call admission criterion, which is local in scope, applies to both Class I and Class II connections, and attempts to ensure that the originating cell has sufficient resources to provide the connection with its desired amount of bandwidth.

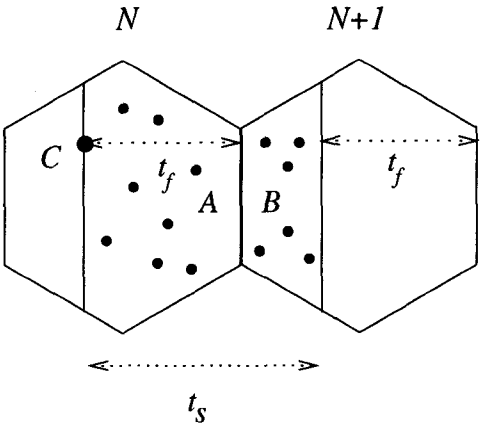

Fig. 32. The sliding window concept for call admission.

The second admission control criterion, which is global in scope, applies to Class I connections only, and attempts to minimize the chances that, once accepted, the connection will be dropped later due to a lack of bandwidth in some cell into which it may handoff.

Consider a request for a new Class I connection C in cell N at time t_C and let t_f be the estimated residence time of C in N. Referring to Figure 32, the key observation that inspired our second criterion is that when C is about to handoff into cell $N+i$, $(i \geq 1)$, the connections resident in $N+i$ are likely to be those in region A of cell N and those in region B of cell $N+1$. More precisely, these regions are defined as follows:

- a connection is in region A if at time t_C its residual residence time in cell N is less than t_f
- a connection is in region B if at time t_C its residual residence time in cell $N+1$ is larger than t_f.

As illustrated in Figure 32, these two regions define a *window* of size t_s anchored at t_C. Let R_{N+i} be the projected bandwidth that has been reserved in cell $N+i$ for the connections in this window. The task of estimating R_{N+i} can be handled by the satellite in, essentially, two ways. First, if enough on-board computational power is available, such a sliding window is maintained for each cell in the footprint. On the other hand, in the presence of reduced on-board capabilities, R_{N+i} can be estimated as follows. Let D_N and D_{N+1} be the total projected bandwidth requirements for Class

I connections currently in cells N and $N+1$, respectively. It is natural to approximate R_{N+i} by the following *weighted average* of D_N and D_{N+1}:

$$R_{N+i} = \frac{t_f}{t_s} \cdot D_N + (1 - \frac{t_f}{t_s}) \cdot D_{N+1}. \tag{11}$$

To justify (11), let S denote the area of a cell. Clearly, the areas covered by regions A and B are, respectively, $\frac{t_f}{t_s} \cdot S$ and $(1 - \frac{t_f}{t_s}) \cdot S$. Now, assuming that D_N and D_{N+1} are uniformly distributed in cells N and $N+1$, respectively, it follows that the projected bandwidth reserved for Class I connections in A and B is, respectively, $\frac{t_f}{t_s} \cdot D_N$ and $(1 - \frac{t_f}{t_s}) \cdot D_{N+1}$ and that their sum is exactly R_{N+i}.

Observe that if the residual bandwidth available in cell $N+i$ once R_{N+i} has been committed is less than the projected bandwidth needs of connection C, it is very likely that C will be dropped. In the face of this bleak outlook, connection C is not admitted into the system. Thus, the second admission criterion acts as an additional safeguard against a Class I connection being accepted only to be dropped at some later point.

VI.3 EXPERIMENTAL RESULTS

Using simulation, we assessed the performance of our predictive bandwidth reservation and call admission scheme in terms of keeping the CDP low and bandwidth utilization high. For this purpose, we compared our predictive scheme to a fixed-rate scheme that sets aside 5 percent of a cell's bandwidth for the exclusive use of Class I handoffs.

In our simulation model we adopted some of the parameters of the Iridium system [46], where the radius of a cell is 212.5Km and the orbital speed of the satellites is 26,000Km/h. This implies that $t_s \approx$ 65sec. Residence times in the originating cells are uniformly distributed between 0 and 65sec.

We have simulated a one-dimensional array of 36 cells each with a static bandwidth allocation of 30Mbps. Thus, in the fixed reservation scenario, an amount of bandwidth equal to 1.5Mbps (i.e. 5% of 30Mbps) is set aside for Class I handoffs.

The results of the simulation are shown in Figures 33, 34, and 35. First, Figure 33 plots the CDP for Class I connections against call arrival rate. The graphs in the figure confirm that our predictive handoff management scheme offers a very low CDP, outperforming the fixed reservation policy.

Next, Figure 34 illustrates the combined call dropping probability of Class I and Class II traffic against the call arrival rate.

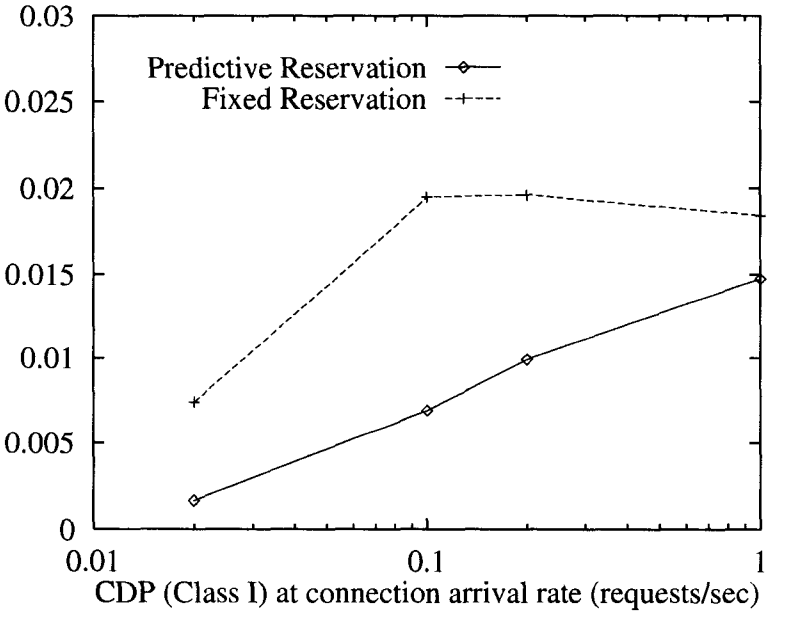

Fig. 33. Call dropping probabilities for Class I traffic.

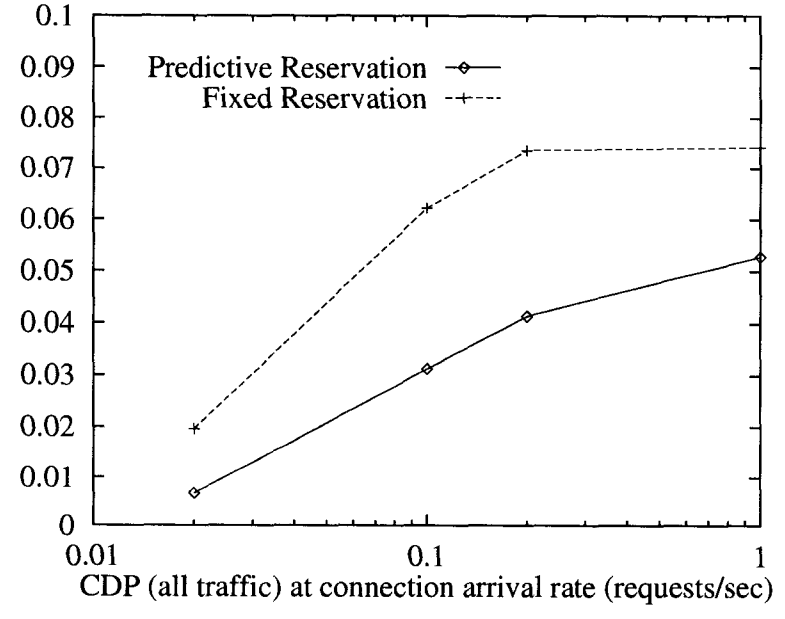

Fig. 34. Call dropping probabilities for Class I and Class II traffic combined.

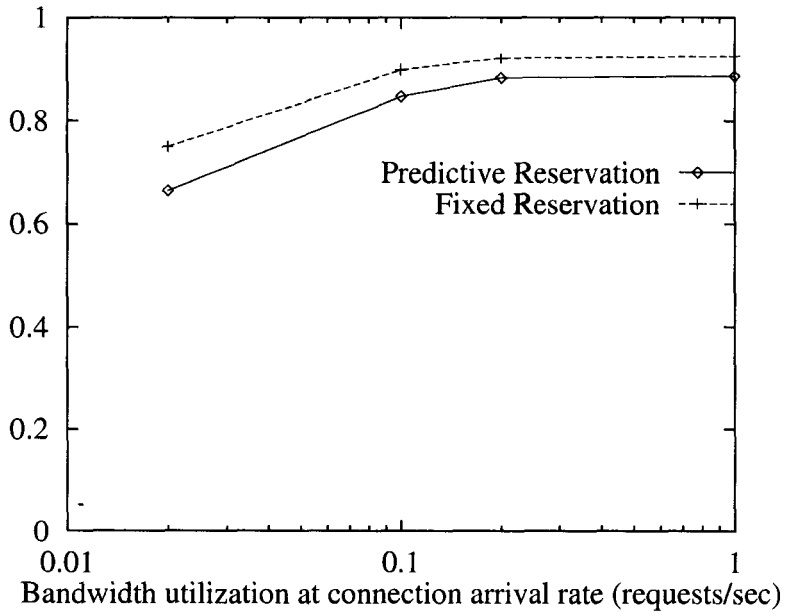

Fig. 35. A comparison of the bandwidth utilization.

It is clear that the goals of keeping the call dropping probability low and keeping the bandwidth utilization high are conflicting. It is easy to ensure a low CDP at the expense of bandwidth utilization and, similarly, it is easy to ensure a high bandwidth utilization at the expense of call dropping probability [47, 17]. The challenge, of course, is to come up with a handoff management protocol that strikes a sensible balance between the two. As Figure 35 shows, our scheme features a high bandwidth utilization in addition to keeping the call dropping probability low.

CHAPTER VII

CONCLUSIONS AND FUTURE WORK

The contributions of this thesis include four novel admission control and resource allocation schemes for QoS provisioning in wireless networks. The schemes addresss QoS provisioning in the local wireless link, not end-to-end, and they focus on call-level QoS metrics such as CBP and CDP. The predictive reservation scheme for LEO satellite networks uses knowledge of the distinctive mobility pattern of MHs in the networks in order to reserve bandwidth for MHs in upcoming cells and lower the CDP.

The schemes for cellular networks feature the use of resource reallocation in order to free bandwidth for arriving connections and lower both the CBP and CDP. The max-min fair borrowing scheme, while it produces better results for CDP and CBP than RBBS, suffers from a few problems that make it a less attractive scheme for QoS provisioning. Unlike RBBS, it does not borrow in a smooth or regular manner - the amount of bandwidth borrowed from or returned to a connection at any given time depends on the other connections in the cell and is unpredictable. In the max-min fair scheme, it is possible that connections requiring a large amount of bandwidth will be rejected even when the required bandwidth is available. The frequent rejection of high-bandwidth connections clearly raises the scheme's performance with respect to CBP and CDP.

RBBS is fair in a different sense than the max-min fair scheme; it is proportionally fair to connections with respect to their stated adaptivity. The concept of shares ensures that fluctuations in bandwidth caused by borrowing are small and gradual. The scheme is simple and inexpensive to implement, especially when compared with schemes that require communication between base stations.

It is certain that multimedia sessions will be part of the traffic mix in future cellular networks. We have shown that call-level QoS provisioning schemes must be aware of the relationship between the streams in multimedia sessions in order to provide good performance in the face of this type of traffic. MSAS is a multimedia session-aware QoS provisioning scheme based on RBBS. MSAS understands the relationship between streams in a session, and allows a simple prioritization between streams – suspendability. We simulated this new scheme against two versions of a session-unaware scheme, also based on the RBBS. In the simulation results, the MSAS produced good

results for CDP, especially when suspending was used. In the session-only experiment, the effects of session awareness on CDP were illustrated more clearly. This was one of the key benefits of session awareness - reducing the CDP.

In this thesis, we introduced two new QoS parameters, DBS and DBT, which represent different views on supplied bandwidth. These metrics are especially important when analyzing schemes that use resource reallocation, in order to understand how much the schemes themselves, rather than network conditions, are affecting the QoS in terms of supplied bandwidth. DBS and DBT can give protocol designers a better perspective on their schemes than CDP and CBP alone. Furthermore, users might set thresholds for DBS and DBT in order to express limits on their willingness to be adaptive.

In experimenting with MSAS, we looked at DBS and DBT to understand effects of borrowing and suspending. We found that MSAS does reasonably well with respect to supplied bandwidth. Even class II connections, which can be subjected to aggressive borrowing, fare surprisingly well. The fluctuation rates demonstrated another effect of borrowing, and MSAS held an advantage in this metric over the unaware schemes. Also, with an average recovery time of just 17 seconds in a busy network, the benefits of using suspendability seem to outweigh the drawbacks. By simulation, we demonstrated the other two key benefits of session awareness - eliminating the overhead caused by processing streams that will not survive and eliminating the possibility that such streams will cause other connections to be dropped (UDR). Finally, we envisioned the role of MSAS as a component of an end-to-end QoS architecture based on the SIP and SDP protocols, in order to demonstrate that its assumptions about available data and about application adaptation are realistic.

The future work in this area should involve priority classes. In our schemes, we give priority based on only two classes of traffic, real-time and non-real-time. In MSAS, we allow the user to define a simple prioritization between streams in a multimedia session. Curescu [43] and Ibrahim [44] have developed some new schemes that define priorities based on utility and price, respectively. Pricing is a very important issue, and more work needs to be done specifically in that area. Future work should explore more ways for users and service providers to express their needs and goals, and search for methods to balance them.

REFERENCES

[1] S. Kalyanasundaram, *Call-Level and Class-Level Quality-of-Service in Multiservice Networks.* PhD thesis, Purdue University, 2000.

[2] C. Oliviera, J. B. Kim, and T. Suda, "An adaptive bandwidth reservation scheme for high-speed multimedia wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 16, Aug 1998.

[3] D. C. Cox and D. O. Reudink, "Increasing channel occupancy in large-scale mobile radio systems: Dynamic channel reassignment," *IEEE Transactions on Communications*, vol. Com-21, Nov 1973.

[4] S. Oh and D. Tcha, "Prioritized channel assignment in a cellular radio network," *IEEE Transactions on Communications*, vol. 40, Jul 1992.

[5] J. Tajima and K. Imamura, "A strategy for flexible channel assignment in mobile communication systems," *IEEE Transactions on Vehicular Technology*, vol. 37, May 1998.

[6] D. Everitt and D. Manfield, "Performance analysis of cellular mobile communication systems with dynamic channel assignment," *IEEE Journal on Selected Areas in Communications*, vol. 7, Oct 1989.

[7] R. Prakash, N. G. Shivaratri, and M. Singhal, "Distributed dynamic fault-tolerant channel allocation for mobile computing," *IEEE Transactions on Vehicular Technology*, vol. 48, Nov 1999.

[8] S. K. Das, S. K. Sen, and R. Jayaram, "A dynamic load-balancing strategy for channel assignment using selective borrowing in a cellular mobile environment," *Wireless Networks*, vol. 3, pp. 333–347, 1997.

[9] H. Jiang and S. Rappaport, "CBWL: A new channel assignment and sharing method for cellular communication systems," *IEEE Transactions on Vehicular Technology*, vol. 43, pp. 313–322, May 1994.

[10] S. Papavassiliou, L. Tassiulas, and P. Tandon, "Meeting QoS requirements in a cellular network with reuse partitioning," *IEEE Journal on Selected Areas in Communications*, vol. 12, Oct 1994.

[11] S. Tekinay and B. Jabbari, "Handover and channel assignment in mobile cellular networks," *IEEE Communications*, vol. 29, Nov 1991.

[12] S. Tekinay and B. Jabbari, "A measurement-based prioritization scheme for handovers in mobile cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 10, Oct 1992.

[13] P. Agrawal, D. K. Anvekar, and B. Narendran, "Channel management policies for handovers in cellular networks," *Bell Labs Technical Journal*, Autumn 1996.

[14] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks," in *Proceedings of ACM SIGCOMM*, 1998.

[15] R. Ramjee, R. Nagarajan, and D. Towsley, "On optimal call admission control in cellular networks," in *Proceedings of IEEE INFOCOM*, 1996.

[16] O. T. W. Yu and V. C. M. Leung, "Adaptive resource allocation for prioritized call admission over an ATM-based wireless PCN," *IEEE Journal on Selected Areas in Communications*, vol. 15, Sep 1997.

[17] D. Levine, I. Akyildiz, and M. Naghshineh, "A resource estimation and call admission algorithm for wireless multimedia networks using the shadow cluster concept," *IEEE ACM Transactions on Networking*, vol. 5, Feb 1997.

[18] A. Aljadhai and T. F. Znati, "A framework for call admission control and QoS support in wireless environments," in *Proceedings of IEEE Infocom*, 1999.

[19] W. Su and M. Gerla, "Bandwidth allocation strategies for wireless ATM networks using predictive reservation," in *IEEE GLOBECOM*, 1998.

[20] W. W. Su, *Motion Prediction in Mobile/Wireless Networks.* PhD thesis, University of California Los Angeles, 2000.

[21] M. Naghshineh and M. Schwarz, "Distributed call admission control in mobile/wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 14, May 1996.

[22] S. Choi and K. G. Shin, "Predictive and adaptive bandwidth reservation for hand-offs in QoS-sensitive cellular networks," *Wireless Networks*, vol. 6, 2000.

[23] J. M. Peha and A. Sutivong, "Admission control algorithms for cellular systems," *Wireless Networks*, 1999.

[24] B. Sadeghi and E. W. Knightly, "Architecture and algorithms for scalable mobile QoS," *Wireless Networks*, vol. 9, pp. 7–20, Jan 2003.

[25] J. Misic, S. T. Chanson, and F. A. Lai, "Admission control for wireless multimedia networks with hard call level quality of service bounds," *Computer Networks*, vol. 31, pp. 125–140, 1999.

[26] C. Chao and W. Chen, "Connection admission control for mobile multiple-class personal communications networks," *IEEE Journal on Selected Areas in Communications*, vol. 15, Oct 1997.

[27] B. M. Epstein and M. Schwarz, "Predictive QoS-based admission control for multi-class traffic in cellular wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 18, Mar 2000.

[28] M. Naghshineh and M. Willebeek-LeMair, "End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework," *IEEE Communications Magazine*, Nov 1997.

[29] S. K. Das, M. Chatterjee, and N. K. Kakani, "QoS provisioning in wireless multimedia networks," in *Wireless Communications and Networking Conference (WCNC '99)*, Oct 1999.

[30] M. El-Kadi, S. Olariu, and H. Abdel-Wahab, "A rate-based borrowing scheme for QoS provisioning in multimedia wireless networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 2, pp. 156–166, 2002.

[31] A. Malla, M. El-Kadi, and P. Todorova, "A fair resource allocation protocol for multimedia wireless networks," *IEEE Transactions on Parallel and Distributed Systems*, vol. 13, no. 10, 2002.

[32] F. Fluckiger, *Understanding Networked Multimedia Applications and Technology.* Prentice–Hall, 1995.

[33] K. Lam, J. Yuen, S. H. Son, and E. Chan, "Scheduling video stream transmissions for distributed playback over mobile cellular networks," in *International Conference on Parallel and Distributed Systems (ICPADS'02)*, Dec 2002.

[34] R. S. Tupelly, J. Zhang, and E. K. P. Chong, "Opportunistic scheduling for streaming video in wireless networks," in *Proceeding of the 37th Annual Conference on Information Sciences and Systems*, Mar 2003.

[35] V. T. Vakili and A. Aziminejad, "A novel DCA scheme for resource sharing in cellular environments with heterogeneous traffic," in *6th IFIP/IEEE International Conference on Management of Multimedia Networks and Services*, Sep 2003.

[36] M. M. Islam, M. Murshed, and L. S. Dooley, "Directional probability function estimation for QoS provisioning in mobile networks using Laplace PDF," in *Australian Telecommunications, Networks and Applications Conference (ATNAC 2003)*, Dec 2003.

[37] M. M. Islam, M. Murshed, and L. S. Dooley, "A direction-based bandwidth reservation scheme for call admission control," in *International Conference on Computers and Information Technology 2002 (ICCIT 2002)*, pp. 345–349, Dec 2002.

[38] M. M. Islam, M. Murshed, and L. S. Dooley, "New mobility based call admission control with on-demand borrowing scheme for QoS provisioning," in *IEEE International Conference on Information Technology: Computers and Communications 2003 (ITCC 2003)*, pp. 263–267, Apr 2003.

[39] M. M. Islam, M. Murshed, and L. S. Dooley, "Enhanced cell visiting probability for QoS provisioning in mobile multimedia communications," in *IEEE International Conference on Information Technology: Coding and Computing 2004 (ITCC 2004)*, Apr 2004.

[40] M. M. Islam, M. Murshed, and L. S. Dooley, "A novel mobility support resource reservation and call admission control scheme for quality-of-service provision in wireless multimedia communications," in *IEEE International Conference on Communications 2004 (ICC 2004)*, Jun 2004.

[41] M. M. Islam and M. Murshed, "Velocity and call duration support for QoS provisioning in mobile wireless networks," *IEEE Wireless Communications*, to appear.

[42] R. Zander and J. M. Karlsson, "A rate-based bandwidth borrowing and reservation scheme for cellular networks," in *IEEE Vehicular Technology Conference (VTC Fall 2004)*, Sep 2004.

[43] C. Curescu and S. Nadjm-Tehrani, "Time-aware utility-based QoS optimisation," in *Proceedings of the 15th Euromicro Conference on Real-time Systems*, Jul 2003.

[44] W. Ibrahim, J. Chinneck, and S. Periyalwar, "A QoS-based charging and resource allocation framework for next generation wireless networks," *Wireless Communications and Mobile Computing*, vol. 3, pp. 895–906, 2003.

[45] E. Del Re, R. Fantucci, and G. Giambene, "Efficient dynamic channel allocation techniques with handover queuing for mobile satellite networks," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 333–347, Feb 1995.

[46] E. Del Re, R. Fantucci, and G. Giambene, "Characterization of user mobility in low earth orbit mobile satellite systems," *Wireless Networks*, vol. 6, 2000.

[47] S. Cho, "Adaptive dynamic channel allocation scheme for spotbeam handover in LEO satellite networks," in *Proceedings of VTC*, 2000.

[48] S. Cho, I. F. Akyildiz, M. D. Bender, and H. Uzunalioglu, "A new spotbeam handover management technique for LEO satellite networks," in *Proceedings of IEEE GLOBECOM*, 2000.

[49] I. Mertzanis, R. Tafazolli, and B. G. Evans, "Connection admission control strategy and routing considerations in multimedia (non-GEO) satellite networks," in *IEEE VTC*, 1997.

[50] M. El-Kadi, S. Olariu, and P. Todorova, "Predictive resource allocation in multimedia satellite networks," in *Proceedings of IEEE GLOBECOM*, Nov 2001.

[51] P. Todorova, S. Olariu, and H. N. Nguyen, "A selective look-ahead bandwidth allocation scheme for reliable handoff in multimedia LEO satellite networks," in *Second European Conference on Universal Multiservice*, Apr 2002.

[52] S. Olariu, S. R. A. Rizvi, R. Shirhatti, and P. Todorova, "Q-Win - a new admission and handoff management scheme for multimedia LEO satellite networks," *Telecommunication Systems*, vol. 22, pp. 151–168, Jan.–Apr. 2003.

[53] M. Handley and V. Jacobson, "SDP: Session Description Protocol." IETF RFC 2327, Apr 1998.

[54] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler, "SIP: Session Initiation Protocol." IETF RFC 3261, Jun 2002.

[55] M. Handley, C. Perkins, and E. Whelan, "Session Announcement Protocol." IETF RFC 2974, Oct 2000.

[56] N. Greene, M. Ramalho, and B. Rosen, "Media gateway control protocol architecture and requirements." IETF RFC 2805, Apr 2000.

[57] C. Groves, M. Pantaleo, T. Anderson, and T. Taylor, "Gateway control protocol version 1." IETF RFC 3525, Jun 2003.

[58] "IETF MMUSIC working group website." www.ietf.org/html.charters/mmusic-charter.html.

[59] G. Camarillo and A. Monrad, "Mapping of media streams to resource reservation flows." IETF RFC 3524, Apr 2003.

[60] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource ReSerVation Protocol (RSVP) - version 1 functional specification." IETF RFC 2205, Sep 1997.

[61] G. Camarillo, W. Marshall, and J. Rosenberg, "Integration of resource management and Session Initiation Protocol (SIP)." IETF RFC 3312, Oct 2002.

[62] M. A. Siddiqui, K. Guo, S. Rangarajan, and S. Paul, "End-to-end QoS support for SIP sessions in CDMA2000 networks," *Bell Labs Technical Journal*, vol. 9, May 2004.

[63] A. S. Youssef, *A Quality of Session Control Framework for Multimedia Multicast Systems.* PhD thesis, Old Dominion University, 1998.

[64] S. Chen and K. Nahrstedt, "Distributed quality-of-service routing in ad-hoc networks," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1488–1505, 1999.

[65] H. Kanakis, P. P. Mishra, and A. Reibman, "An adaptive congestion control scheme for real-time video packet transport," *IEEE/ACM Transactions on Networking*, vol. 3, 1996.

[66] S. V. Raghavan and S. K. Tripathy, *Networked Multimedia Systems.* Prentice–Hall, 1998.

[67] N. Tran and K. Nahrstedt, "Adaptive adaptation by program delegation in VOD," in *Proceedings of the Int'l. Conference on Multimedia Computing and Systems*, pp. 96–107, 1998.

[68] B. J. Vickers, M. Lee, and T. Suda, "Feedback control mechanism for real-time multipoint video services," *IEEE Journal on Selected Areas in Communications*, vol. 15, 1997.

[69] U. Black, *Second Generation Mobile and Wireless Networks.* Prentice–Hall, 1999.

[70] D. Bertsekas and R. Gallager, *Data Networks.* Prentice–Hall, 1992.

[71] L. Kalampoukas, A. Varma, and K. K. Ramakrishnan, "An efficient rate allocation algorithm for ATM networks providing max-min fairness," in *Proceedings of the 6th IFIP Int'l Conference on High Performance Networking*, pp. 143–154, Sep 1995.

[72] J. Rosenberg and H. Schulzrinne, "An offer/answer model with the Session Description Protocol (SDP)." IETF RFC 3264, Jun 2002.

[73] D. Prokopp, M. Matthes, O. Drobnik, and U. Krieger, "Integration of mobility-, QoS-, and CAC-management for adaptive mobile applications," in *Proceedings of Architecture for QoS in the Internet (Art-QoS 2003)*, LNCS 2698, pp. 29–48, Springer-Verlag, 2003.

APPENDIX

THE CELLULAR NETWORK SIMULATOR

In order to experiment with the QoS provisioning schemes we devised, we developed a simulation environment, the Cellular Network Simulator (CNS), in which to implement and run them. This software went through many iterations - one with a user interface showing the traffic in the cells in a network, one that was highly threaded and ran in real-time, another less-threaded that ran in simulated time. It is written in Java and its input is provided in XML. Besides being used for our own work, a version of it was used by Prokopp et. al. [73] and by Olariu et. al. [52].

It is currently a robust simulator for admission control and bandwidth reservation schemes with a goal of providing call-level QoS in cellular and satellite networks. New schemes can be easily swapped in and out for testing against the same input, and data collection and processing mechanisms are also easily interchangeable. CNS allows complex network topologies to be modelled, including the ability to define roads or heavily populated areas, and to mix different sized cells in the same network. In this chapter, we will describe the design of the current version, which was used to run the experiments on the session-aware scheme.

SIMULATION CONFIGURATION WITH XML

A simulation that is to be run takes most of its configuration information from an XML file. The name of the XML file and the length of the simulation in (simulated) seconds is specified on the command line. The Document Type Definition (DTD) for the simulation configuration information is shown here:

```
<?xml version="1.0" encoding="US-ASCII"?>
<!ELEMENT simulation (network, traffic+, cell+)>
<!ATTLIST simulation
    name                CDATA #REQUIRED
    network_class       CDATA #REQUIRED
    mh_class            CDATA #REQUIRED
    avg_speed           CDATA #REQUIRED
```

```
    min_speed            CDATA #REQUIRED
    max_speed            CDATA #REQUIRED>
<!ELEMENT network (#PCDATA)>
<!ATTLIST network
    name                 CDATA #REQUIRED
    x                    CDATA #REQUIRED
    y                    CDATA #REQUIRED
    cell_class           CDATA #REQUIRED
    cell_collector_class CDATA #REQUIRED
    mh_collector_class   CDATA #REQUIRED
    filetag              CDATA #IMPLIED
    reserve              CDATA #IMPLIED
    bandwidth            CDATA #IMPLIED
    shares               CDATA #IMPLIED
    rate                 CDATA #IMPLIED>
<!ELEMENT traffic (connection+)>
<!ATTLIST traffic
    name                 CDATA #REQUIRED
    avg_duration         CDATA #REQUIRED
    min_duration         CDATA #REQUIRED
    max_duration         CDATA #REQUIRED>
<!ELEMENT connection (layer+)>
<!ATTLIST connection
    name                 CDATA #REQUIRED
    type                 CDATA #REQUIRED
    fixed                CDATA #REQUIRED
    suspendable          CDATA #REQUIRED
    avg_bandwidth        CDATA #REQUIRED
    min_bandwidth        CDATA #REQUIRED
    max_bandwidth        CDATA #REQUIRED>
<!ELEMENT layer (#PCDATA)>
<!ATTLIST layer
    name                 CDATA #REQUIRED
    bandwidth            CDATA #REQUIRED>
```

```
<!ELEMENT cell (#PCDATA)>
<!ATTLIST cell
    x_index             CDATA #REQUIRED
    y_index             CDATA #REQUIRED
    bandwidth           CDATA #IMPLIED
    shares              CDATA #IMPLIED
    reserve             CDATA #IMPLIED
    rate                CDATA #IMPLIED
    busy_index          CDATA #IMPLIED>
```

Simulation definition

A simulation consists of a network definition, a set of traffic definitions and a set of cell definitions. The speed characteristics for mobile hosts are specified at the simulation level. Java dynamic class loading enables the network and mobile host classes to be input parameters as well, so that using a different network model (i.e. satellite instead of cellular) or mobile host model (i.e. directed rather than random) does not require code changes other than the new class file itself.

Network definition

The network definition consists of an x and y value (yielding the number of cells and a topology), as well as class names for the cell class to be exercised and the mobile host and cell data collector classes. These two classes can be used to collect and output the data from the mobile hosts and cells that are of interest in the experiment, (e.g., CDP and CBP). Optionally, a file tag can be specified to perform automated output processing that directly creates **gnuplot** input files. Also optionally, the amount of bandwidth in each cell (e.g., 30000), a fixed bandwidth reservation value (e.g., 0.05), a connection arrival rate (e.g., 1.0), and the number of shares to use in the proportionally fair borrowing scheme (e.g., 10) can be specified globally for the network.

Traffic definition

Each traffic type must have duration characteristics and a distinct name (needed for data collection and output). A traffic definition contains a set of one or more connections. Multimedia session traffic will have more than one connection.

Connection definition

Each connection type has a distinct name and a type (class). It has a flag indicating whether it is suspendable or not and bandwidth requirement characteristics. The fixed flag indicates whether to use the bandwidth characteristics directly as given or calculate different bandwidth requirements for each mobile host using a geometric distribution around the maximum, minimum and average given.

Layer definition

For some types of multimedia connections, such as video and audio, the ability for an application to adapt may be set to distinct bandwidth levels, corresponding to different rates on an adjustable-rate codec, or layers of quality. A connection can have a set of layer definitions that describe the steps in bandwidth between the desired and the minimum levels. This is useful for borrowing schemes, which can borrow with respect to these layers, in order to make best use of application adaptation.

Cell definition

Each cell has an x and y coordinate within the network layout, so that the network topology can be visualized. Using this idea, different parameters may be set for each cell, in order to simulate differing conditions within a network. Optionally, a cell may have its total bandwidth, number of shares (rate-based borrowing scheme), and bandwidth reservation amount set individually. Also the connection arrival rate may be set per cell, to simulate more populated areas in the network. Finally, a *busy_index* can be set, which may be used with a special kind of MH, such that MHs are drawn to move in the direction of cells with a higher *busy_index*. This would help simulate highways or downtowns or other such topological features. If these optional values are not set at the cell level, the value is taken from the network definition.

It is easy to see that this same method could be used to model different sized cells in a network topology, say with micro- or pico-cells in a downtown area, but larger

cells in suburban areas. To implement it, the MH speed values could be set at the cell level rather than the simulation level (a higher speed equaling a shorter cell dwell time).

CLASS DESCRIPTIONS

The *Network* and *NetworkDefinition* classes

The *Network* object holds the collection of *Cells* and arranges them in the correct structure (i.e. assigns the neighbor relationships). The *NetworkDefinition* object holds the configuration information about the network that was read from the xml file.

The *Cell* and *CellDefinition* classes

The *Cell* class represents a cell in a network. A cell knows about its neighbors. This class is abstract and it is extended in order to implement QoS provisioning schemes. For example, the classes *SessionCell*, *FixedReservationCell*, and *SessionUnawareCell* contain the implementations of the similarly-named schemes. The class's primary methods are *request* and *release*. A *MobileHost* calls these methods in order to ask for resources and to free them. Each *Cell* object has its own *TrafficGenerator* object that creates and introduces new *MobileHosts* to it. A *CellDefinition* object holds the configuration information for each *Cell*. Each *Cell* is a thread.

The *TrafficGenerator* class

A *TrafficGenerator* object produces the new traffic for a single cell. It also takes care of making *MobileHosts* that it creates progress through the network. The length of a particular simulation is specified as a number of simulated seconds, which the *TrafficGenerator* steps through, creating traffic according to the rate specified, and checking all its current *MobileHosts* at each step to see if it is time for them to move or if their lives have ended. When it creates a new *MobileHost*, it assigns it a traffic type and then starts its life time and cell dwell time timers. In the latest version of CNS, a traffic type is chosen with equal likelihood from the set of traffic types defined for the simulation. A nice enhancement would be for each traffic type to be assigned a percentage of total traffic occurrence rate. Each *TrafficGenerator* is a thread and steps through the simulation time independently.

The *TrafficDefinition* and *ConnectionDefinition* classes

A *TrafficDefinition* class is used by a *TrafficGenerator* to set up a *MobileHost*. It defines the attributes of a particular type of traffic. It contains a vector of *ConnectionDefinitions*. There is one *TrafficDefinition* object for each traffic type defined in the simulation configuration xml file.

The *Session* and *Connection* classes

The *TrafficGenerator* uses the data from the *TrafficDefinition* and *ConnectionDefinion* to create a *Session* and a vector of *Connection* classes for each *MobileHost* object. These objects hold the actual resource requirements for each MH, as well as the actual resources it is being provided throughout its lifetime.

The *MobileHost* class

The *MobileHost* class represents a mobile host. It is an abstract class that must be extended to define the *move* method, which is called when the *MobileHost* is ready to execute a handoff, and returns the neighboring *Cell* to which the *MobileHost* should move. Some examples are the *RandomMH*, which moves randomly and the *DirectedMH*, which is more likely to continue in the same direction it came from. The *SatelliteMH* moves in one direction across the linear set of *Cells* in the *SatelliteNetwork*.

The *MHDataCollector* and *CellDataCollector* classes

There is one of each of these objects for the simulation. They each have one or more *writeData* methods and a *processData* method. *MobileHosts* use the *MHDataCollector* object to record information about the service they are receiving from the network, and *Cells* record pertinent information about their condition in the *CellDataCollector*. The *processData* methods output the consolidated information to the user or to files.

The *Simulation* and *SimConfig* classes

The *Simulation* object uses the *SimConfig* object to process the xml file and create the definition objects. It then creates the main simulation objects and starts the

TrafficGenerators. When the simulation is over, it calls the *processData* methods of the two data collectors in order to output the experiment results.

VITA

Mona El-Kadi Rizvi
Department of Computer Science
Old Dominion University
Norfolk, VA 23529

Mona El-Kadi Rizvi received her B.S. degree in computer science (summa cum laude) from Old Dominion University (ODU) in 1985, and she returned to ODU to pursue her Ph.D. in 1996. She has worked in industry since 1984 as a software engineer and software manager on various projects for companies and customers that include the Defense Advanced Research Projects Agency (DARPA), the Naval Research Laboratory (NRL), the United States Air Force (USAF), the U.S. Department of Labor, the Service Employees International Union, USA.NET, and United Airlines. She will begin working as an assistant professor at Christopher Newport University in Newport News, VA in August 2004.

Typeset using LaTeX.