

University of Alberta

SUPPORTING PROPORTIONAL DELAY DIFFERENTIATION IN CDMA  
CELLULAR WIRELESS ENVIRONMENTS

by

Liu Wu ©

A thesis submitted to the Faculty of Graduate Studies and Research in partial  
fulfillment of the requirements for the degree of **Master of Science**.

Department of Computing Science

Edmonton, Alberta

Fall 2002



National Library  
of Canada

Acquisitions and  
Bibliographic Services

395 Wellington Street  
Ottawa ON K1A 0N4  
Canada

Bibliothèque nationale  
du Canada

Acquisitions et  
services bibliographiques

395, rue Wellington  
Ottawa ON K1A 0N4  
Canada

*Your file Votre référence*

*Our file Notre référence*

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-81500-5

University of Alberta

Library Release Form

Name of Author: Liu Wu

Title of Thesis: Supporting Proportional Delay Differentiation in CDMA Cellular Wireless Environments

Degree: Master of Science

Year this Degree Granted: 2002

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Liu Wu .

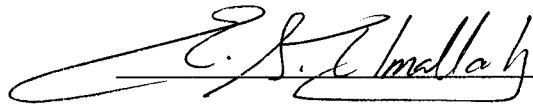
Liu Wu  
505-8515, 112 Street  
Edmonton, AB  
Canada, T6G 1K7

Date: Aug. 26, 2002.

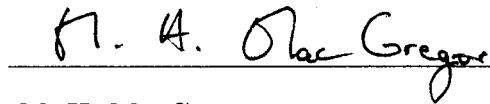
University of Alberta

Faculty of Graduate Studies and Research

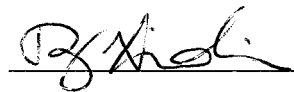
The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled **Supporting Proportional Delay Differentiation in CDMA Cellular Wireless Environments** submitted by Liu Wu in partial fulfillment of the requirements for the degree of **Master of Science**.



Ehab S. Elmallah



M. H. MacGregor



Xiaodai Dong

Date: August 21, 2002

# Abstract

The IETF's Differentiated Services (*DiffServ*) architecture provides a general framework for provisioning quality of service (*QoS*) over the Internet. Extending the DiffServ architecture to mobile users of the third generation (*3G*) wireless systems is expected to provide a low cost means of running guaranteed-service applications on personal communication devices.

In this thesis, we extend the proportional delay differentiation to 3G mobile systems, and provision such service on the downlink of wide-band code-division multiple access (*W-CDMA*) air interface. We seek to devise a set of admission control and scheduling mechanisms that aim at maximizing the effective throughput of the system. Taking user mobility and CDMA soft capacity into consideration, two proportional delay differentiation schedulers and an admission control scheme integrated with power prediction are proposed. Our simulation results show that significant improvement in the average delay, fairness among the DiffServ classes, and the total effective throughput can be attained using the predictive admission control scheme.

# Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor, Dr. Ehab S. Elmallah, for his continuous support and help. His guidance and advice during the research help me overcome many technical obstacles, which would take much more effort. He has spent much time and effort on this thesis research, which deserves my special appreciation.

I am grateful to all the friends and colleagues with whom I have spent my time as a graduate student in the University of Alberta.

I would like to express my earnest gratitude to my parents for their always love and support that made everything I have possible. I am deeply indebted to my dear husband for his love, support, and encouragement.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	An Overview of Future Mobile Cellular Environments . . . . .	1
1.2	An Overview of Future Internet Quality of Service Architectures	4
1.3	Thesis Organization and Contributions . . . . .	5
<b>2</b>	<b>An Overview of Some Related Work</b>	<b>7</b>
2.1	General Features of the DiffServ Architecture . . . . .	8
2.2	An Overview of Forwarding Per-hop Behaviours . . . . .	10
2.2.1	Expedited Forwarding (EF) . . . . .	11
2.2.2	Assured Forwarding (AF) . . . . .	11
2.2.3	Proportional Differentiated Services . . . . .	13
2.2.4	LIRA . . . . .	15
2.2.5	SCORE (Scalable Core) . . . . .	15
2.2.6	Loss Guaranteed Service . . . . .	17
2.3	Some Relevant Aspects of CDMA Systems . . . . .	17
<b>3</b>	<b>Provisioning Proportional Delay Differentiation on the Down- link of a Cellular CDMA Environment</b>	<b>23</b>
3.1	Design Objectives . . . . .	24
3.2	Proportional Delay Differentiation Schedulers . . . . .	27
3.2.1	A Basic Scheduling Mechanism . . . . .	28
3.2.2	Scheduling Based on Head-of-Queue Delay . . . . .	29
3.2.3	Scheduling Based on Most-Delayed Packets . . . . .	31
3.2.4	Downlink Scheduling Procedure . . . . .	33

3.3	A Call Admission Control Scheme . . . . .	35
3.4	Integrating Admission Control with Power Prediction . . . . .	37
3.5	Summary . . . . .	40
<b>4</b>	<b>Downlink Performance Evaluation</b>	<b>41</b>
4.1	Overview of the Simulation Parameters . . . . .	42
4.1.1	Cell and User Mobility Parameters . . . . .	42
4.1.2	Radio Propagation Path Loss and CDMA Parameters . . . . .	43
4.1.3	Traffic parameters . . . . .	44
4.1.4	Summary of Simulation Parameters . . . . .	48
4.2	Performance results on Aggregate Delay Calculation Schemes . . . . .	50
4.3	Performance Results without Call Admission Control . . . . .	53
4.3.1	Average class delay . . . . .	53
4.3.2	Average delay ratio . . . . .	54
4.3.3	Effective throughput . . . . .	57
4.4	Performance Results Using Call Admission Control . . . . .	58
4.4.1	Average class delay . . . . .	59
4.4.2	Effective throughput . . . . .	60
4.5	Performance Results Using Power Prediction . . . . .	61
4.5.1	Average class delay . . . . .	62
4.5.2	Effective throughput . . . . .	63
4.5.3	System total effective throughput . . . . .	64
4.6	Conclusions . . . . .	65
<b>5</b>	<b>Conclusions and Future Work</b>	<b>68</b>
5.1	Conclusions . . . . .	68
5.2	Future Work . . . . .	70

# List of Figures

1.1	UMTS system architecture . . . . .	3
2.1	The architecture of Differentiated Services . . . . .	9
2.2	CDMA downlink and uplink interference . . . . .	18
3.1	Queue organization in CDMA . . . . .	30
3.2	A numerical example for the HOQ scheduler . . . . .	31
3.3	A numerical example for the WIN scheduler . . . . .	32
4.1	19-cell simulation model . . . . .	43
4.2	WIN & HOQ comparison – average class delay . . . . .	51
4.3	WIN & HOQ comparison – average delay ratio . . . . .	51
4.4	No admission control – average class delay . . . . .	54
4.5	No admission control – average delay ratio . . . . .	55
4.6	No admission control – effective throughput . . . . .	57
4.7	Call admission control – average class delay . . . . .	59
4.8	Call admission control – effective throughput . . . . .	61
4.9	Power prediction – average class delay . . . . .	62
4.10	Power prediction – effective throughput . . . . .	63
4.11	System total effective throughput . . . . .	64

# List of Tables

4.1	Maximum number of mobile users for different rates . . . . .	47
4.2	CDMA and power parameters . . . . .	48
4.3	Other simulation parameters . . . . .	49

# Chapter 1

## Introduction

This thesis deals with methods for extending future Internet Quality of Service architecture to mobile users in future generation wireless cellular environments. In particular, the thesis considers extending the Differentiated Services (*DiffServ*) model to mobile users in the proposed Universal Mobile Telecommunication Systems (*UMTS*) cellular environment. Towards this goal, this chapter motivates the general research direction. The chapter is organized as follows. Section 1.1 gives an overview of the UMTS system and the IMT-2000 requirements. Section 1.2 discusses two prominent quality of service (*QoS*) architectures for future Internet structures, in connection with the goal of extending such architectures to the wireless cellular domain. Finally, Section 1.3 concludes by outlining the thesis organization and main contributions.

### 1.1 An Overview of Future Mobile Cellular Environments

Personal mobile communication systems have evolved in three distinct stages: analog, digital, and multimedia. The first generation (*1G*) analog mobile communication systems provided mobile users with voice only services. Then, in the 1990s, the second generation (*2G*) digital mobile communication sys-

tems were developed. The second generation mobile devices offer users digital voice services, as well as data and short message capabilities, at data transmission rates from 9.6 Kbps to 14.4 Kbps. The most successful of the 2G cellular systems is the Global System for Mobile Communications (*GSM*), which supports over 50% of the world's cellular subscribers.

The evolution from the second generation personal wireless systems to the third generation (*3G*) systems began with the introduction of the IMT-2000 requirements, as defined by the International Telecommunications Union (*ITU*). IMT-2000 defines the common standard, which all 3G mobile communication systems should conform to. The main characteristics of IMT-2000 are high transmission data rates and multimedia communication. IMT-2000 is intended to provide up to 2 Mbps data rate for stationary users, 384 Kbps for low-speed moving users, and 144 Kbps for high-speed moving users. Wide-band code-division multiple access (*W-CDMA*) air interface has emerged as the most important air interface for 3G systems. The CDMA architecture is different from the time-division multiple access (*TDMA*) architecture which operates on a time slot basis, and the frequency-division multiple access (*FDMA*) architecture which divides the whole spectrum into frequency bands. The CDMA architecture spreads each user's data over the full allocated bandwidth and transmits different data streams simultaneously by encoding each stream with a particular code. This technology brings higher capacity, improved call quality, and enhanced privacy over other multiple access technologies. Universal Mobile Telecommunication Systems (*UMTS*) [1] architecture using the W-CDMA and CDMA-2000 [7] are two of the most important proposals for achieving the IMT-2000 requirements.

Figure 1.1 illustrates some of the main building blocks in the proposed UMTS architecture. The UMTS system architecture consists of two network elements: the UMTS terrestrial radio access network (*UTRAN*), which is responsible for all radio-related functionality, and the Core Network (*CN*). The CN is responsible for all radio interface independent functions such as call control and mobility management. In the UMTS architecture, the CN

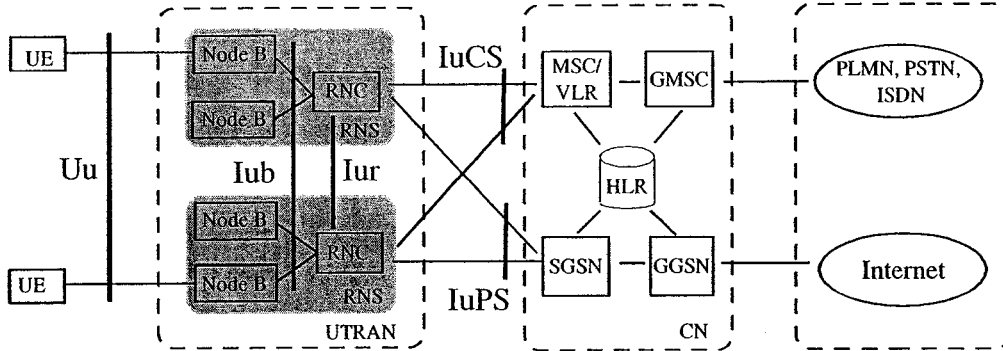


Figure 1.1: UMTS system architecture

is connected to the outside world using two network domains: the Circuit Switched domain centered at the Mobile Switching Centers (*MSCs*) and the Packet Switched domain centered at the GPRS Support Nodes (*GSNs*). Each of the two domains connects to different network backbones. The first backbone carries voice traffic coming from the UMTS Public Land Mobile Network (*PLMN*) peers, the Public Switched Telephone Network (*PSTN*), or the Integrated Service Digital Network (*ISDN*). The second backbone relies on current Internet IP technology, and carries data traffic.

The demand for provisioning quality of service (QoS) over future Internet infrastructure is dramatically increasing. Extending such QoS to mobile users of the UMTS environments becomes of particular importance. We note that UMTS attempts to fulfill the QoS requirements by the introduction of a few QoS classes [17, 29]. The *conversational* class and the *streaming* class are two classes proposed for serving real-time delay-sensitive traffic. The conversational class is intended to serve voice over IP (*VoIP*) and conferencing applications. Uni-directional streaming video and audio applications, on the other hand, can be served using the streaming class. Traffic belonging to both of the above two classes is real-time traffic and always requires a high level of service guarantee. For web-browsing and other messaging applications, the *interactive* class and the *background* class are defined to serve traffic that has no tight delay constraints, but requires high reliability.

## 1.2 An Overview of Future Internet Quality of Service Architectures

Currently, the Integrated Services (*IntServ*) architecture and the Differentiated Services (*DiffServ*) are two prominent models for supporting QoS over future Internet infrastructure.

The IntServ architecture [39] is based on resource reservation. The Resource ReSerVation Protocol (*RSVP*) [45] is proposed as a signaling protocol to reserve all the resources that each flow needs along its path. Thus, the IntServ architecture needs each router along the path to keep per-flow state information. Maintaining per-flow information for a large number of flows gives rise to a scalability problem in the IntServ architecture.

The DiffServ architecture [3, 32], on the other hand, attempts to solve the scalability problem of the IntServ architecture. The DiffServ architecture is based on classifying the traffic at the edge of the network into a few classes and performing class-based forwarding in the core. In each router, each class is associated with some allocated resource and forwarding behaviour. There is no per-flow absolute bandwidth guarantee or delay bounds. Resources are allocated on a class basis, and forwarding assurance is also guaranteed on a class basis. Since DiffServ architecture's conception in 1998, researchers have developed various types of forwarding behaviours that are designed to achieve various QoS aspects.

In this thesis, we choose to consider investigating methods for extending the DiffServ architecture to future generation of wireless cellular systems because of its favourable scalability properties. Challenges arise with the unpredictable nature of wireless communication. It is not like wired networks, where the amount of available bandwidth is always known exactly. Especially in CDMA transmission, one flow's transmission depends on other flows' transmission rate, transmission power, and mobiles' location. The capacity of a CDMA system is called *soft capacity* because it is determined by many run-time factors such as mobility locations, transmission rates, and

power. The difficulty, of reserving a certain amount of bandwidth during a certain period in wireless environments, is also one reason for us to choose the DiffServ architecture. Mechanisms for extending the DiffServ architecture to future generation of wireless cellular systems should take both user mobility and the CDMA's soft capacity into consideration.

### 1.3 Thesis Organization and Contributions

The main focus of the thesis is on developing and investigating the performance of call admission control and scheduling mechanisms to extend the DiffServ architecture to mobile wireless users in future generation cellular systems. Only transmission on the downlink (i.e., from the base station to mobile users) is considered in the thesis. We use simulation as the main tool in our performance study. Toward achieving the above goal, the rest of the thesis is organized as follows.

Chapter 2 gives a background on the IETF's DiffServ architecture with particular emphasis on explaining the different per-hop forwarding behaviours that have been proposed recently in the literature. Chapter 2 also gives a background on the basic characteristics of the W-CDMA air interface that have an impact on our design.

Chapter 3 is devoted to the development of the basic mechanisms underlying our design. The chapter starts by identifying a set of general design objectives and guidelines, and then maps the identified guidelines into a more well defined set of design requirements and objectives. Basic scheduling mechanisms and call admission control mechanisms are then developed to meet the derived objectives. A novel aspect of the work done in this thesis is the investigation of methods that aim at predicting the required base station transmission power, while taking user mobility within one cell into consideration.

Chapter 4 introduces the parameters used in our simulation realizations, and reports on the results obtained when the system operates

- (a) with a suitable scheduler but without call admission control,
- (b) with a suitable scheduler and call admission control, and
- (c) with a suitable scheduler and a call admission control that tries to predict future power requirements.

The main findings are summarized in Chapter 5, along with some directions of future research.

The main contributions of this thesis are in identifying and formalizing a problem in the rapidly evolving field of wireless-wireline integration with QoS provisioning, proposing solution strategies to the problem, and studying the performance of the proposed solutions using simulation. In order to undertake the above tasks, the thesis identifies user mobility and the soft capacity aspect of the CDMA cellular environment as the most critical aspects to consider.

## Chapter 2

# An Overview of Some Related Work

The previous chapter has outlined the motivation for extending the Differentiated Services (DiffServ) model to future generation wireless cellular environments. This chapter serves two purposes: firstly, it provides background information on various proposals for implementing the DiffServ model; secondly, it discusses the main characteristics of the code division multiple access (CDMA) air interface (that is being proposed for use in the third generation wireless systems) that impact our design decisions.

In particular, Section 2.1 discusses the main characteristics of the DiffServ model from a system architecture point of view. Section 2.2 presents a number of per-hop forwarding behaviours that have been proposed for use in implementing the DiffServ model, and gives detailed information on studies done on three of them. Section 2.3 discusses some relevant aspects of the CDMA scheme as it applies to mobile cellular users in the proposed UMTS system.

## 2.1 General Features of the DiffServ Architecture

This section gives an overview of the DiffServ architecture. We recall that DiffServ [3] has been developed to alleviate the scalability problem of the Integrated Services (IntServ) model and, at the same time, to support quality of service (QoS) capabilities.

In IntServ, we focus on individual flows. RSVP provides per-flow reservation and each router keeps per-flow state. In contrast, DiffServ classifies traffic into a small number of classes, and allocates resources for each class. Packets are forwarded based on the class information encoded in the packet header. There is no need to reserve resources for individual flow or to keep per-flow state in each router. As a result, DiffServ does not suffer from the scalability problem. By allocating different resources to each class, DiffServ provides each class with different levels of service. In DiffServ, there is a contract between the customer and the service provider, which defines what kind of service the customer wants to receive. This contract is called the *service level agreement (SLA)*.

A differentiated services domain (*DS domain*) normally consists of one or more networks under the same administration, such as an organization's intranet or an ISP. There are two types of routers in each domain: edge routers and core routers. The edge routers interconnect the DS domain to other domains, either DS or non-DS. The core routers only connect to other routers in the same domain. One of the key features of DiffServ is the different functionality of edge routers and core routers. In DiffServ, traffic policing is done at the edge and class-based forwarding is done in the core.

Figure 2.1 shows the DiffServ architecture. DiffServ edge routers classify and possibly condition ingress traffic to ensure that packets traversing the domain are appropriately marked. The edge routers have three functions: classification, marking and, possibly, conditioning. The classifier model at the edge routers selects packets and maps each packet to a particular for-

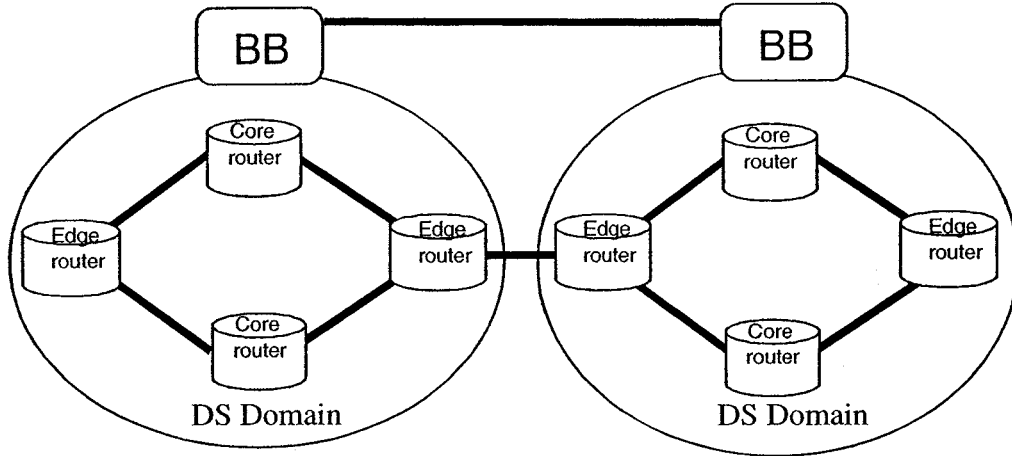


Figure 2.1: The architecture of Differentiated Services

warding behaviour. The selected packets are marked with a particular *differentiated service codepoint (DSCP)* [31]. This codepoint is a 6-bit number encoded in the packet header, and it is used by the core routers to decide the forwarding treatment for this packet. Finally, the conditioner in the edge routers will measure the traffic stream against a traffic profile, which comes from the SLA. Depending on whether the traffic stream meets with the SLA or not, the packets are classified as either in-profile or out-of-profile. Different actions such as dropping, shaping, or remarking could be performed on out-of-profile packets.

The core routers perform fewer operations than the edge routers; they apply the forwarding behavior by mapping the DS codepoint marked by the edge routers to one of the implemented per-hop behaviours (*PHBs*).

In each DiffServ domain, there is another component called bandwidth broker (BB) [32]. BBs have two responsibilities — resource management for each domain, and message passing to the adjacent domain's BB. If an allocation is desired for a flow, a request will be sent to the BB. The BB will then verify that there exists sufficient unallocated resources to meet the request. If so, the BB informs the leaf router with the flow's information to deliver the service to the flow at the time the service is needed. If the

destination is outside the requester's domain, the requester domain's BB informs an adjacent domain's BB that it will use this amount of resource allocation.

Based on the above discussion, we can see that the essential characteristics of DiffServ include:

- **Handling of aggregated traffic instead of individual flow.** Resources are allocated to each class. For individual flow, there is no absolute guarantee.
- **Traffic policing at the edge.** The edge routers perform all the complicated operations, such as classification, marking, and conditioning. This ensures that all the packets coming into the core network are marked properly for class-based forwarding.
- **Class-based forwarding in the core.** In the core routers, there is no per-flow state kept, as there is in IntServ. Core routers forward the packets, by mapping the codepoint encoded in the packet header, to one of the implemented PHBs. The selected PHB decides which kind of forwarding treatment the packet can receive.
- **Resource provisioning rather than reservation.** In IntServ, the RSVP protocol will reserve the resource along the path for each flow if the resource is available. However, in DiffServ, the resource is allocated to each class, and all the packets belonging to the same class will compete for the total resource allocated to this class.

## 2.2 An Overview of Forwarding Per-hop Behaviours

A number of DiffServ per-hop forwarding behaviours have been studied. In this section, some of these models are discussed. Although they all conform to the same DiffServ architecture, they differ in the specific services provided,

and in their implementation mechanisms. We will give a detailed view of the research done on three of the introduced per-hop forwarding behaviours, in terms of their basic mechanisms and simulation results. These three per-hop behaviours are the Assured Forwarding (*AF*), the proportional delay differentiation, and the Scalable Core (*SCORE*).

### 2.2.1 Expedited Forwarding (EF)

There are two PHBs standardized by IETF. One is Expedited Forwarding (EF) [18], which is also called the premium service. It is defined as a forwarding treatment for a traffic aggregate, where the departure rate of the aggregate's packets from any DS node must equal or exceed a configurable rate. Moreover, the EF traffic should always receive the configurable rate independent of the intensity of other traffic types. EF service actually provides the equivalent of a dedicated link with a fixed bandwidth. A lower-bound of service rate is guaranteed for EF packets. As a result, EF service could provide effective support for real-time applications.

### 2.2.2 Assured Forwarding (AF)

The Assured Forwarding [16] is the other IETF standardized PHB group intended to provide different levels of forwarding probabilities. In AF, four forwarding classes, along with three drop precedences in each class, have been proposed. Each class is allocated a minimum amount of resources (bandwidth or buffer spaces). Resource sharing among classes is possible: at any moment, each class may utilize the unused resources from other AF classes. However, whenever the other class needs its own resources back, the class which is currently using this resource has to give it up. The customer's packets are assigned to one of these classes according to the specified SLA.

Within each AF class, three drop precedences are possible. Each packet is marked with one of these three drop precedence values. The drop precedence of a packet determines the relative importance of the packet within the class.

In the case of congestion, packets with the highest drop precedence will be dropped first. However, the dropping among different classes is totally independent.

There are a number of factors affecting the performance of Assured Forwarding including bandwidth management, buffer management, and fairness between different traffic types. Considerable research has been done relating to these aspects.

Buffer management is one of the most important issues in Assured Forwarding. It manages the selection of packets to be dropped during congestion. The most widely used buffer management mechanism in Assured Forwarding is random early drop (RED) [11]. Several variations of RED are used in Assured Forwarding for multiple drop precedence. This is the subject of [28].

Another widely studied issue in Assured Forwarding is the number of drop precedences and the assignment of different drop precedence to the incoming packets. This is examined in [12, 13, 15].

TCP and UDP are two important transport protocols in the Internet, which respond differently to packet losses. Upon packet dropping, TCP flows will reduce their packet rates, while UDP flows will keep their rates and use the excess bandwidths. How to guarantee fairness between different traffic types is the subject of [2, 12, 34].

[12] provides a broad study of most aspects of Assured Forwarding services, including the RED buffer management scheme, the number of drop precedences, and the fairness issue. In [12], NS2 simulator is used and there are a total of 10 data sources. Nine of these are Reno TCP sources and the other one generates UDP data at a rate of 1.28 Mbps. All of them belong to the same class. There are two drop precedence numbers examined: 2(green, red) and 3(green, yellow, red). [12] evaluates two performance measures: the utilization of reserved bandwidth and the fairness achieved in allocation of excess bandwidth. The simulation results show that 3 level of drop precedence performs better than 2 level of drop precedence in terms of fairness between different traffic types. In addition, RED parameters and implementations

have significant impact on the performance.

### 2.2.3 Proportional Differentiated Services

The proportional differentiated services model is introduced in [10], where the performance measures at each hop are proportional to certain class differentiation parameters.

Proportional delay differentiation [10] is a type of proportional differentiated services which uses average delay as the performance measure. In proportional delay differentiation, all network traffic is divided into  $N$  classes. For differentiation, each class is assigned a delay differentiation parameter (*DDP*). The delay differentiation parameter determines the proportional relation between each class's average delay. Specifically, if the average queuing delay for class  $i$  packets is  $\bar{d}_i$ , [10] states that:

$$\frac{\bar{d}_i}{\bar{d}_j} = \frac{\delta_i}{\delta_j} \quad (2.1)$$

where  $\delta_i$  is the DDP of class  $i$  and, because higher classes are assumed better, all the DDPs are ordered such that  $\delta_1 > \delta_2 > \dots > \delta_N$ . Thus, the class with a small DDP experiences less delay than a class with a large DDP, and the delay is proportional to the DDPs.

[10] also proposes two packet scheduling schemes for proportional delay differentiation. One is called the Backlog-Proportional Rate (BPR) scheduler; it determines the link sharing for each class, using the backlog of each class queue and the class DDP.

$$\frac{r_i(t)}{r_j(t)} = \frac{q_i(t)/\delta_i}{q_j(t)/\delta_j} \quad (2.2)$$

where  $q_i(t)$  is the backlog of queue  $i$  at time  $t$  and  $r_i(t)$  is the service rate assigned to queue  $i$  at time  $t$ . The other scheduler is called the Waiting-Time Priority (WTP) scheduler, and it uses the head-of-queue packet delay as the load of each class, instead of the queue backlog used in BPR.  $q_i(t)$  in

Equation 2.2 is substituted by  $w_i(t)$ , which is the waiting-time of the first packet in queue  $i$  at time  $t$ .

In the experiment described in [10], there are 4 classes. The interarrival time between packets of the same class is determined by Pareto distribution with a shape parameter  $\alpha = 1.9$ . The packet length is the same for all classes: 40% of the packets are 40 bytes, 50% are 550 bytes, and 10% are 1500 bytes. Each simulation result is an average over 10 random runs with different seeds. For performance measures, average delay ratio is used to evaluate the delay proportion between classes, and average queuing delay is also examined for microscopic views.

The simulation in [10] shows that both the BPR and WTP schedulers maintain the proportional delay differentiation under heavy load conditions, even in short time scales. For moderate load traffic conditions, when the utilization is 70% or under, the differentiation ratio is under expected. The WTP scheduler outperforms the BPR scheduler in terms of providing consistent delay differentiation independent of class load distribution.

[30] discusses the implementation of delay differentiation among classes, and also proposes a dynamic delay class adjusting mechanism. By dynamically adjusting the delay class of a flow, the end-to-end delay bound for the flows could be provided. Each flow specifies the average end-to-end delay requirement and also the maximum price that it is willing to pay. The delay class adaptation placed on the access routers is responsible for adjusting the delay class of a flow so that the current delay class is the lowest possible that satisfies the flow's end-to-end delay requirement. Dynamic class selection is also discussed in [9].

The other direction for proportional differentiated services is Loss Rate Differentiation [8]. Instead of using current existing buffer management schemes, such as complete buffer partitioning, partial buffer sharing, or multiclass RED [11, 35], two Proportional Loss Rate Droppers (*PLR*) are proposed. They determine the dropping of packets in each queue by loss rate differentiation; that is, the loss rate in each class queue is proportional to

the loss rate differentiation parameters of this class. The simulation results show that both of the droppers can, to some extent, meet the proportional loss rate constraints.

Additional research on proportional differentiated services is presented in [4, 19, 23, 25, 26].

## 2.2.4 LIRA

In [42], Stoica and Zhang propose another Assured Service model called Location Independent Resource Accounting (*LIRA*). In *LIRA*, service profiles between users and ISP are defined in units of resource tokens, instead of absolute bandwidth. Each user is assigned a service profile specified by a resource token bucket  $(r, b)$ , where  $r$  is the resource token rate and  $b$  is the bucket depth. The number of tokens consumed by a single bit is not a constant; it is a dynamic function of the path it traverses.

Experiments in [42] show that *LIRA* is effective in providing service differentiation at user level, and provides high probability for the marked packets to be delivered.

## 2.2.5 SCORE (Scalable Core)

Stoica and Zhang later propose the Scalable Core (*SCORE*) architecture [41, 43], which combines the best of both IntServ and DiffServ.

*SCORE* is intended to provide IntServ like service. For each individual flow, there is absolute bandwidth guarantee and delay bound, while it is not necessary to keep per-flow state information in routers along the path. Thus, *SCORE* is, at the same time, as scalable as the stateless DiffServ architecture. The service is achieved by encoding packet state information in the packet's header. In other words, the packet itself carries its state information along the path. *SCORE* pushes the idea of DiffServ in that the packet itself can play an important role in state information passing.

The key technique used in *SCORE* is the Dynamic Packet State (*DPS*), in

which each packet carries in its header its own state information, initialized by the ingress node and updated by each core router passed. This state information carried in the packet's header (instead of kept in each core router), is used for admission control and packet scheduling during transmission.

For packet scheduling, each packet is assigned an eligible time and a deadline upon its arrival. The packet is held until the eligible time and transmitted according to its deadline. For the  $k^{th}$  packet of flow  $i$ , at the  $j^{th}$  node on its path, its eligible time  $e_{i,j}^k$  and deadline  $d_{i,j}^k$  is computed as [43]:

$$\begin{aligned} e_{i,j}^1 &= a_{i,j}^1 \\ e_{i,j}^k &= \max(a_{i,j}^k + g_{i,j-1}^k, d_{i,j}^{k-1}) \\ d_{i,j}^k &= e_{i,j}^k + \frac{l_i^k}{r_i} \end{aligned} \tag{2.3}$$

where  $a_{i,j}^k$  is the arrival time for the  $k^{th}$  packet of flow  $i$  at node  $j$ ,  $l_i^k$  is the length of the packet,  $r_i$  is the reserved rate for the flow and  $g_{i,j-1}^k$  is stamped to the packet header from the previous node, as the difference of the packet deadline and its departure time at the  $(j-1)^{th}$  node. We can see from Equation 2.3, that in the packet header,  $r_i$ ,  $g_{i,j-1}^k$  and  $d_{i,j}^{k-1}$  should all be encoded, as well as one admission control parameter.

[43] proposes using 4 bits from the type of service (TOS) field and 13 bits for packet fragmentation field, for packet state encoding in DPS. A floating point-like format encoding scheme is used.

The experiments are carried out on two FreeBSD v2.2.6 machines with point-to-point 100 Mbps Ethernets. The first machine is configured as an ingress router and, the second is configured as a core router. All traffic is UDP traffic and the packet length is 1000 bytes. Packet arrival and departure time, as well as, flow rate are used for evaluation. The results show that it is indeed possible to apply DPS techniques to current networks.

### 2.2.6 Loss Guaranteed Service

The Loss Guaranteed (*LG*) service model is proposed in [5]. The LG per-hop behaviour aims at providing a loss bound service through flow admission control. The admission control admits a new flow if the following two conditions hold:

- The measured current link usage does not exceed the targeted link utilization level.
- The measured loss rate after accepting the new flow does not exceed the targeted loss rate.

The LG admission control scheme relies heavily on measuring aggregate traffic rate, packet loss, and queue length. The measured quantities are communicated among the BBs. The LG service provides a quantitative QoS guaranteed service in terms of loss bound. However, the measurement based admission control scheme may not provide a completely reliable loss bound. It is assumed that applications requesting LG service are tolerant to some occasional loss bound violation.

## 2.3 Some Relevant Aspects of CDMA Systems

To satisfy the IMT-2000 requirements, many proposals have been considered [1, 7]. Of these, the W-CDMA and the CDMA-2000 proposals have gained wide acceptance. Both proposals incorporate many optimized engineering designs to perform modulation, interleaving, coding, and spreading functions. Transmission on the downlink is characterized by the use of orthogonal codes to separate the users, and the use of dedicated transport channels to serve users receiving data at relatively high speeds.

From a networking perspective, an important characteristic of data transmission in the above proposals is the *soft capacity* aspect of the resulting

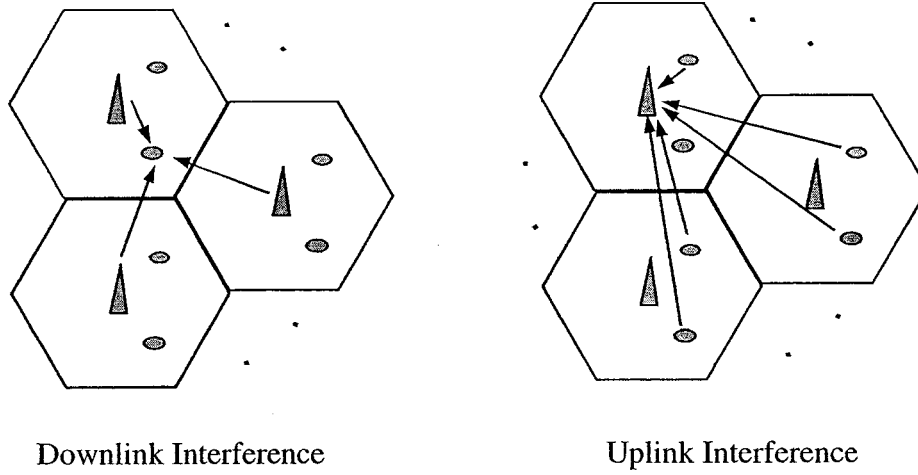


Figure 2.2: CDMA downlink and uplink interference

cellular system. The soft capacity aspect refers to the limiting role that interference levels play in determining the available bandwidth for information transmission from the base station to mobile users (the downlink direction), and from the mobile users to the base station (the uplink direction). Figure 2.2 sketches the sources of interference in both directions when the system operates in the FDD mode.

In the FDD mode, the transmission from each base station to its served mobile users shares the same carrier frequency. Hence, the receiver of a mobile deals with transmission intended to other mobiles as interference. On the other hand, transmission from mobile users on the uplink uses a different carrier frequency. Hence, the receiver in each base station receives transmission from the mobile devices in the same cell as well as neighbouring cells.

In the thesis, we are interested in provisioning DiffServ over the downlink. A key relation that determines the required power to the  $i$ th mobile user is given by:

$$\left( \frac{E_b}{I_0} \right)_i \leq \frac{\text{energy per bit}}{\text{interference} + \text{noise power}} \quad (2.4)$$

Rewriting Relation 2.4 using more specific symbols gives the following formula [20]:

$$\left(\frac{E_b}{I_0}\right)_i \leq \frac{\frac{W}{R} \times \frac{P_{0,i}}{L_{0,i}}}{\sum_{j \neq i} \gamma \frac{P_{0,j}}{L_{0,i}} + \sum_{c \neq 0} \frac{P_c}{L_{c,i}} + \eta_0 W} \quad (2.5)$$

Relation 2.5 is subject to the constraint that the sum of power assigned to all users does not exceed the total available transmission power of the base station. In the above relation:

- $W$  is the CDMA chip rate.  $R$  is the data transmission rate.  $\frac{W}{R}$  is called the spreading gain.
- $P_{0,i}$  is the transmission power from mobile  $i$ 's serving base station (with index 0) to mobile  $i$ .  $P_c$  ( $c > 0$ ) denotes the total transmission power from base station  $c$ .
- $L_{0,i}$  is the path loss from mobile  $i$ 's serving base station (with index 0) to mobile  $i$ .  $L_{c,i}$  is the path loss from base station  $c$  to mobile  $i$ .
- $\gamma$  is an orthogonality factor, which defines the orthogonality between different users in downlink transmission.
- $\eta_0$  is the noise density.

Thus, in Relation 2.5, for any mobile  $i$ ,  $\frac{P_{0,i}}{L_{0,i}}$  is the received power at the desired mobile  $i$  from its own cell's base station — base station 0.  $\sum_{j \neq i} \gamma \frac{P_{0,j}}{L_{0,i}}$  is all the received interference at mobile  $i$  which is dedicated to other users in this cell.  $P_{0,j}$  is the transmission power from base station 0 to other users, except mobile  $i$ . We have to note that the path loss here is the path loss from base station to mobile  $i$ , instead of other mobiles.  $\sum_{c \neq 0} \frac{P_c}{L_{c,i}}$  is the interference power received at mobile  $i$  from other cells. Each user has a target signal-to-interference ratio (*SIR*) —  $\left(\frac{E_b}{I_0}\right)_i$ , which comes from the users' bit error rate (*BER*) requirement. The obtained signal-to-interference ratio should not be less than the target  $\left(\frac{E_b}{I_0}\right)_i$  in order to guarantee that the required bit error rate is maintained.

A few remarks about Relation 2.5 are now in order.

1. Increasing the transmission rate of the encoded data to the  $i$ th user results in decreasing the spreading gain ( $W/R$ ) experienced by the user's receiver. In W-CDMA, for example,  $W = 4.096M$  chips per second. A user requesting data at a speed of 128 Kbps will be assigned a spreading gain of 32.
2. Although the downlink transmission is characterized by the use of orthogonal codes to separate the users, the orthogonality advantage deteriorates as interference increases. The orthogonality factor  $\gamma$  may approach 1 in worst case scenarios.
3. In the special case where only 1 user exists in a cell, and assuming the remaining parameters in Relation 2.5 are set to some practical values (e.g.,  $W = 4.096$  Mcps,  $P_{0,i} = P_c = 25$  watts,  $\left(\frac{E_b}{I_0}\right)_i = 7$  dB,  $\eta_0 = 3.98e-18$  mwatts, and the user is at 700 meters from the base station), the maximum speed that the system can offer is 1 Mbps. This is consistent with the limits mentioned in the W-CDMA proposal.
4. For any given rate  $R$ , for which the system is designed to serve, the maximum number of users that the system can serve at this speed is finite, regardless of the total available base station power (that is, the capacity of the system is *interference limited*).
5. The minimum amount of power that should be allocated to each user can be computed by solving a linear system of equations [37].

Numerous studies have been done on the efficient management of the CDMA soft capacity aspect to maximize throughput, while satisfying mobile users requests [6, 14, 21, 22, 24, 27, 36, 37, 38, 40, 44]. The amount of literature in this direction is growing at a fast pace.

The main differences among such studies lie in considering particular combinations of the following factors:

1. The particular aspects of the CDMA environment under consideration: this involves the transmission direction (downlink versus uplink), whether the systems operates in the TDD mode or the FDD mode, in addition to other detailed link level aspects (e.g., type of modulation, coding, error correction, etc.)
2. The particular mix of heterogeneous traffic considered in the study. Voice traffic is characterized by its delay sensitivity, low rate (and consequently high spreading gain), low required  $E_b/I_0$  ratio, and the applicability of variable encoding techniques to increase the system capacity (i.e., the number of served users). Data traffic, on the other hand, generally requires higher transmission speeds (and consequently can be served using lower spreading gains), and higher required  $E_b/I_0$  ratio, and a bursty nature.

Earlier work, for example, considered a mix of voice traffic and data traffic [37]. More recent work considered more heterogeneity in the data traffic with various ways of generating the traffic and the various QoS requirements [22].

3. The particular QoS aspects involved in the study (e.g., delays, rates, or losses), the granularity of the data size to which the aspect is being applied to (e.g., link level protocol data unit (PDU) transmitted during a few milliseconds time slot, or an IP data packet), and the required performance level (e.g., minimizing certain averages, or maintaining certain fair queuing levels).
4. Whether handoff is considered or not, and if considered, the particular method used in reserving resources to the handoff.
5. The methods used to design the required call admission control, load control, and scheduling mechanisms.

The main factors that distinguish the study done in the thesis in comparison to previous studies can be summarized as follows:

- (a) the performance measures of interest here apply to aggregate classes of traffic (in a general case, each class may contain heterogeneous traffic),
- (b) the admission control mechanisms apply to flows (not individual packets), and
- (c) we aim at developing strategies that attempt to predict the future state of the system with some knowledge of the expected mobile users behaviour. In contrast, many existing studies rely on keeping track of the immediate past behaviour of the system to make suitable decisions.

## Chapter 3

# Provisioning Proportional Delay Differentiation on the Downlink of a Cellular CDMA Environment

The previous chapters motivated the goal of extending the Differentiated Services architecture to cellular wireless domains. This chapter aims at achieving a step towards this goal by investigating and proposing methods for provisioning the proportional delay differentiation (PDD) per-hop behaviour (PHB) on the downlink of a cellular CDMA environment, operating in the frequency-division duplex (FDD) mode. The chapter is organized as follows. Section 3.1 outlines our design objectives: we first identify a set of high-level guidelines and design goals, and then map this set of guidelines to a set of more specific design requirements; then, we formalize more specific research goals in the context of using PDD PHB. In Section 3.2, two delay differentiation schedulers are proposed. Section 3.3 explores a basic call admission control mechanism that works in conjunction with the scheduling algorithms. Section 3.4 proposes a novel mechanism for integrating admission control with power prediction, which uses mobility prediction to deal

with the CDMA soft capacity aspect. Section 3.5 summarizes the chapter.

## 3.1 Design Objectives

In the previous chapters, we motivated the need of extending the DiffServ architecture to wireless cellular domains. In the remaining part of the thesis, we consider a challenging problem in this context: serving delay-critical traffic over the downlink of a CDMA environment operating in the FDD mode.

Currently, the area of provisioning end-to-end delay-critical traffic on the wired part of the Internet, through the utilization of some per-hop forwarding behaviours, is an active topic of research. Extending the existing approaches to mobile users in a CDMA environment adds two new layers of complexity: the need to harness mobility, and the need to harness the CDMA's soft capacity. In this thesis, we hope to achieve a step in this direction.

To this end, the following issues are addressed in this section:

- i. identifying a general set of guidelines and design objectives for this research direction,
- ii. mapping the guidelines to a set of more specific requirements and design parameters,
- iii. identifying a suitable per-hop behaviour (PHB) for satisfying the derived requirements, and lastly,
- iv. taking all the above into consideration, formalizing specific goals for the control and scheduling mechanisms that are used to steer the development proposed in the remaining part of the chapter.

We start by identifying a set of general guidelines that are in line with the main objective of provisioning delay-critical traffic. Satisfying all aspects of the guidelines mentioned below is more ambitious than what the thesis

achieves. However, they serve as a basis for deriving a more manageable set of design requirements. The guidelines are as follows:

- a. It is of interest to consider flows that are sufficiently long so as to allow the user to travel a considerable distance from the base station while receiving the flow. This requirement captures the challenging mobility aspect.
- b. It is of interest to serve delay-bounded traffic.
- c. The end user should not be charged on the basis of the sheer traffic volume received. Rather, the basis of charging the user should be on the amount of *useful* received traffic. (Ideally, the user should be charged on the information content requested, rather than the traffic volume required to deliver the requested information. This aspect, however, is left as a future research topic.)

When mapped to specific design requirements and parameters, the above guidelines give rise to consideration of a broad spectrum of traffic workloads, mobility scenarios, and delay requirements. Of this broad spectrum, we choose to consider the following set of simplified requirements.

1. To account for the first guideline on flows that are long enough to allow significant user mobility, we consider cases where the ratio of a flow duration to user speed allows a unidirectional user to traverse at least half the cell radius. For example, one may consider flows of time duration that is uniformly distributed between 60 and 90 seconds, delivered to users that move randomly at an average speed of 10 meters per second, in a cell that has a radius of 1000 meters. Here, a unidirectional user may traverse at least 0.6 of the cell radius while receiving a flow.
2. Achieving the goal underlying the second guideline is a challenging networking problem. A simple conceptual framework is used. In this framework, we choose to associate each packet with a maximum acceptable delay interval. The maximum acceptable delay is decreased

each time a packet is delayed in an internal router between the source and the destination. Upon arrival to the Internet-wireless gateway, if the packet is delayed more than the preset expiry time, the packet is considered useless.

3. To account for the third guideline, we keep track of the fraction of packets in each flow that have been delivered within the acceptable time delay discussed in (2). If the fraction of successfully delivered packets is below a certain threshold value (e.g., below 90% of total number of packets in a flow), then we count this as a failure in delivering the entire flow. All packets in such a failed flow do not contribute to the *effective throughput* of the system.

The above considerations motivate the use of proportional delay differentiation (PDD) per-hop behaviour (PHB), discussed in Section 3.2, to achieve our goals. The PDD per-hop behaviour uses two complementary mechanisms for its operation:

- a. during congestion time, the PDD PHB aims at managing the bandwidth so as to make the average delays perceived by packets in any two delay classes in the inverse ratios of the corresponding delay weights, and
- b. if the resulting end-to-end flow delay is not as desired, then a per-flow end-to-end delay class adaptation mechanism dynamically adjusts the delay class of that particular flow.

Currently, for the wired Internet, research work has focused on the first aspect of the PDD PHB. The second aspect has recently been considered in [9, 30]. In this thesis, the main focus is on maximizing the effective throughput without using any flow adaptation mechanism.

We also note that, at any interval of observation, one can identify the following distinct behaviours of a router utilizing the PDD PHB:

1. No queue builds up for any class: this behaviour occurs when the allocated bandwidth to each class suffices to serve all packets in the class as they arrive.
2. Queues build up for the classes, but most packets do not miss their expiry time: in such cases, the scheduling algorithm should succeed in enforcing the preset delay weights of the PDD PHB.
3. Queues build up for the classes, and most packets miss their expiry time: here, the router is heavily congested, and the average delay in each class approaches the preset maximum expiry time.

A network operator attempts to implement a set of call admission control and scheduling mechanisms that keep the system away from exhibiting the third behaviour.

Based on the above framework, our objective is to devise a set of control and scheduling mechanisms that aim at maximizing the effective throughput of the system in a high mobility environment, subject to constraints on the total available base station transmission power, and the requirement of achieving target proportional delays among the various DiffServ classes at moderate congestion times. The devised control mechanisms require knowledge of the current allocated base station power to active users. This can be done by requiring the base station to periodically update the Internet-wireless gateway with the average power allocated to each user, where the average is computed over an immediate past interval of time.

## 3.2 Proportional Delay Differentiation Schedulers

This section deals with basic issues in constructing a suitable scheduler for implementing the PDD PHB. The methods proposed here are for use in an Internet-wireless gateway; they are different from the methods proposed in [10, 30] for use in a conventional wired router.

The main difference is that in the wired case, one can assign one queue to all flows belonging to each delay class. In each time slot during congestion time, as many packets are transmitted from each queue as are required to satisfy the relative delay constraints. Many packets may belong to the same flow and each flow is destined to some user.

In a CDMA cellular environment, flows are transmitted over dedicated or shared transport channels. There is a maximum volume of traffic that can be transmitted to a single user in any time slot; the volume depends on the transmission speed required by the SLA. Hence, multiple concurrent transmissions to different users should take place within one time slot. This gives rise to the need for a scheme to schedule transmissions for multiple queues belonging to different users in each delay class, and an accompanying scheme to assess the average delays encountered in each delay class.

The above aspects are investigated in this section. To start, in Section 3.2.1, we review the basic proportional delay differentiation scheduling mechanism. In the next two sections, we explore two possible ways of estimating the average delays experienced in each delay class. Then, in Section 3.2.4, we present the main steps involved in implementing a PDD scheduler.

### 3.2.1 A Basic Scheduling Mechanism

Our basic scheduling policy is based on delay differentiation. Its goal is to make the average delay experienced by packets in each class inversely proportional to the delay weight of the class.

Suppose we have  $N$  delay classes. The  $i$ th delay class is associated with a delay weight  $\Delta_i$ , where  $\Delta_1 > \Delta_2 > \dots > \Delta_N$ . So that, for  $i < N$ , class  $i$  is required to experience less delay than class  $(i + 1)$ . Specifically, if  $\bar{d}_i$  is the average delay for class  $i$ , we want to achieve the following goal:

$$|\bar{d}_i \Delta_i - \bar{d}_j \Delta_j| \rightarrow 0 \quad (3.1)$$

We call the product  $\bar{d}_i \Delta_i$  of average delay and delay weight the *normalized delay* of class  $i$  and denote it by  $p_i$ . That is,

$$p_i = \bar{d}_i \times \Delta_i \quad (3.2)$$

$p_i$  is the value that we consider when it comes to transmission, and it determines the relative priority among classes during transmission. We want to achieve almost the same normalized delay for each class, such that a class with a higher delay weight experiences less delay than a lower delay weight class.

When it comes to packet transmission, classes with higher normalized delay  $p_i$  are given higher priority for transmission than classes with smaller values of  $p_i$ . Because the wireless environment supports simultaneous multiple transmission, and the exact number of multiple transmitted packets can only be determined at run time, we can only say that classes with higher normalized delay have higher transmission priority. Due to the limited total amount of the base station transmission power budget, in the case of network congestion, packets belonging to a lower priority class have a higher chance of being postponed than packets belonging to a higher priority class based on the computed normalized delays.

### 3.2.2 Scheduling Based on Head-of-Queue Delay

As mentioned earlier, we need to support simultaneous multiple transmission of flows on the downlink. Figure 3.1 sketches a conceptual organization of flows at the Internet-wireless gateway. In this section, we discuss the suitability of using the delays incurred by the head-of-queue packets to derive scheduling decisions.

The rationale of using the head-of-queue delays is as follows: If a queue receives a small amount of service relative to its arrived packets in a recent past interval, the head-of-queue packet will experience a large delay. Thus, the load of a queue is reflected by the waiting-time of the packet at the head of the queue.

Specifically, we consider an approach to deriving scheduling decisions from the head-of-queue packet delays using the following calculations. At any

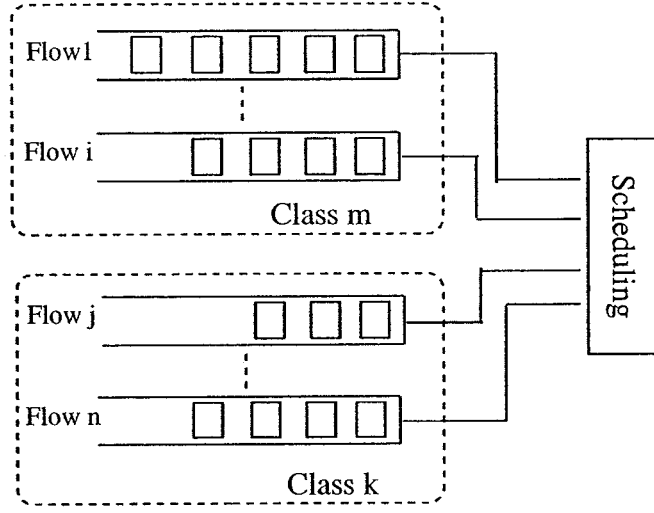


Figure 3.1: Queue organization in CDMA

scheduling instant  $t$ , let  $n_i$  be the number of queues belonging to class  $i$ , and let  $h_{ij}(t)$  be the head-of-queue packet delay of the  $j$ th flow in class  $i$  at time  $t$ . Compute the normalized delay  $p_i$  of class  $i$  as:

$$p_i = \bar{h}_i(t) \times \Delta_i \quad (3.3)$$

where  $\Delta_i$  is delay weight of class  $i$ , and  $\bar{h}_i(t)$  is calculated as:

$$\begin{aligned} \bar{h}_i(t) &= \frac{h_{i1}(t) + h_{i2}(t) + \dots + h_{in_i}(t)}{n_i} \\ &= \frac{\sum_{j=1}^{n_i} h_{ij}(t)}{n_i} \end{aligned} \quad (3.4)$$

The obtained relative values of normalized delays are then used to determine the classes to be served.

Figure 3.2 illustrates a simple numerical example. Suppose there are two classes, Class 1 and Class 2, where  $\Delta_1$  is 2 and  $\Delta_2$  is 1. At time  $t$ , the contents in each queue are as shown in Figure 3.2. Each rectangle represents one packet, and the number in each rectangle is the packet delay. The above approach gives the following normalized delays:

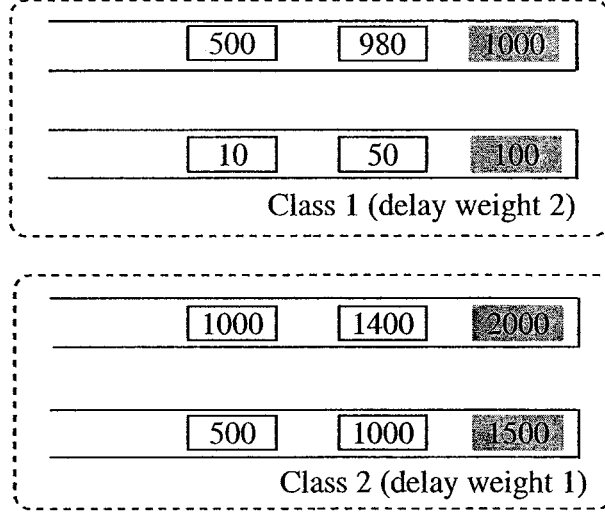


Figure 3.2: A numerical example for the HOQ scheduler

$$p_1 = 2 \times \frac{(1000 + 100)}{2} = 1100$$

and

$$p_2 = 1 \times \frac{(2000 + 1500)}{2} = 1750$$

Hence,  $p_2$  is greater than  $p_1$  which means that for the next transmission, Class 2 packets will have higher priority than Class 1 packets using the above HOQ scheduler.

The above numerical example also illustrates a head-of-queue delay distribution that results in a scheduling decision that favours the class with very close head-of-queue delays.

### 3.2.3 Scheduling Based on Most-Delayed Packets

In this section, we consider another approach for deriving scheduling decisions based on the average delay of the most-delayed packets in each class. We denote this approach as WIN, since there is one parameter that specifies

the window-size which is the number of most-delayed packets that should be considered in each class.

The WIN approach performs the following calculations. Suppose the window-size is  $n_{win}$ . We select the first  $n_{win}$  most-delayed packets. Let  $w_{ij}(t)$  be the delay of the  $j$ th longest delayed packet in class  $i$  at time  $t$ . The normalized delay  $p_i$  of class  $i$  is expressed as:

$$p_i = \bar{w}_i(t) \times \Delta_i \quad (3.5)$$

where  $\Delta_i$  is the delay weight of class  $i$ , and  $\bar{w}_i(t)$  is calculated as:

$$\begin{aligned} \bar{w}_i(t) &= \frac{w_{i1}(t) + w_{i2}(t) + \dots + w_{in_{win}}(t)}{n_{win}} \\ &= \frac{\sum_{j=1}^{n_{win}} w_{ij}(t)}{n_{win}} \end{aligned} \quad (3.6)$$

The obtained relative values of normalized delay are then used to determine the transmission priority among classes, and the classes to be served.

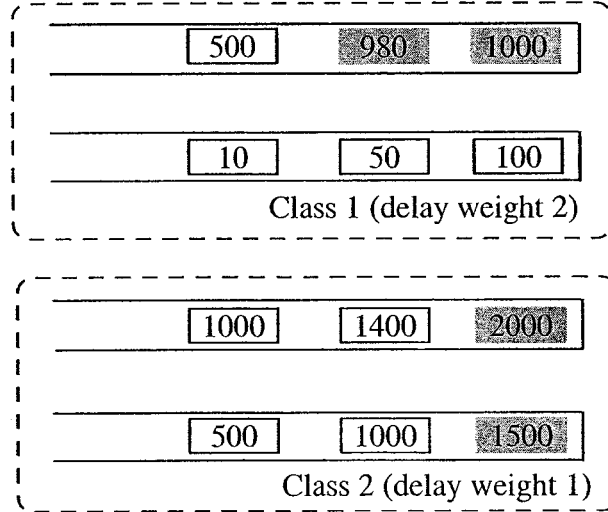


Figure 3.3: A numerical example for the WIN scheduler

Figure 3.3 uses the same numerical example as Figure 3.2. We set the

window-size to be 2 in this example. Packets in gray are those packets selected for calculation. The obtained normalized delays are:

$$p_1 = 2 \times \frac{(1000 + 980)}{2} = 1980$$

and

$$p_2 = 1 \times \frac{(2000 + 1500)}{2} = 1750$$

The above example illustrates that WIN gives priority to the class with highly delayed packets, without being influenced by the delay distribution of the head-of-queue packets.

### 3.2.4 Downlink Scheduling Procedure

This section describes the main steps in constructing the scheduling algorithm (Algorithm 1) used in our performance study. The scheduling algorithm may use either the HOQ or the WIN approach to compute the normalized delays (Recall that the higher the computed normalized delay, the higher priority a class should receive). The algorithm is organized around a main loop that postpones transmission of flows from classes with low computed normalized delays, when the base station is perceived to have insufficient power.

Algorithm 1 shows the scheduling algorithm for one power update interval. The head-of-queue packets from each active flow are considered for transmission. The *while* loop in Step 1 determines whether or not there exists a feasible power assignment. If a feasible allocation does exist, which means all the head-of-queue packets can be transmitted under current power budget while achieving the specified signal to interference ratio, the *while* loop is ended, and Step 5 is executed. In Step 5, all the head-of-queue packets in each active flow will be transmitted using the power assignment allocated in this power update interval.

---

**Algorithm 1** Downlink scheduling procedure

---

**scheduling parameters:**

- $q_{active}$ : the set of active queues of flows,  
(i.e., queues with packets waiting for transmission)  
 $N$ : the number of classes  
 $\Delta_i$ : the delay weight for class  $i$ ,  $i=1, 2, \dots, N$ .

**algorithm:**

- 1: **while** (there is no feasible power assignment to serve all the head-of-queue packets in  $q_{active}$ ) **do**
  - 2:   based on  $\Delta_1$  to  $\Delta_N$  and packet delays from  $q_{active}$ , use *HOQ scheduler* (or *WIN scheduler*) to decide the priority among classes based on the computed normalized delays
  - 3:   postpone the transmission of the flow with the smallest head-of-queue delay in the lowest priority class, and remove that flow from  $q_{active}$
  - 4: **end while**
  - 5: transmit the head-of-queue packets of all the flows in  $q_{active}$  using the allocated power assignment
-

Otherwise, in Step 2, we perform the scheduling program using either the HOQ or the WIN scheduler to determine the relative priority among classes based on computing the normalized delays. Then, in Step 3, one flow from the lowest priority class with the smallest head-of-queue delay is postponed to the next transmission. The while loop continues until a feasible power assignment is perceived to exist. The scheduled packets are then transferred to the RNC for subsequent transmission to the user. This way, we guarantee that a higher priority class has higher forwarding probability than a lower priority class in terms of the class normalized delay. The final goal is to achieve the class average delay inversely proportional to the class's delay weight.

### 3.3 A Call Admission Control Scheme

In this section, we devise a suitable call admission control (CAC) scheme, that is intended to work in conjunction with the scheduling algorithms devised in the previous section to implement the PDD PHB.

In our design, the admission control procedures run at the Internet-wireless gateway, and determine whether or not to admit a new request at the beginning of every new *admission cycle*. Call admission control limits the number of flows accepted and aims at avoiding heavily congested situations.

Each admission cycle has a fixed time length  $\tau$ , where the time required to process all outstanding requests in all queues establishes a lower bound on  $\tau$ . For each admission cycle, the flows are considered for admission in the order of class priority, from the highest class (Class 1) to the lowest class (Class N) (Note, in contrast, the scheduling algorithm discussed in the previous section uses an order based on the computed normalized delays.). In each class, the flows are considered for admission on a first-come-first-served basis. Once a flow is admitted, the system uses the scheduling algorithm (Algorithm 1) to schedule the delivery of packets of the admitted flows.

The call admission control makes the admission decision based on current

available base station power. If there exists a feasible power allocation for all the flows already admitted and the newly incoming flow, the new flow is admitted. Otherwise, the system rejects the new flow because admitting it may cause power shortage at the base station. Algorithm 2 shows the procedure for this admission control scheme. Although admitting a new flow means sufficient base station power is currently available, this admission control scheme does not guarantee that sufficient base station power will continue to exist in the future; this motivates us, in the next section, to seek methods for predicting power requirements at some instants in the future.

---

**Algorithm 2** A call admission control algorithm for each admission cycle  $\tau$   
**admission control parameters:**

---

$N$ : the number of classes

**algorithm**

```

1: for (class  $i$  from 1 to  $N$ ) do
2:   for (each flow in class  $i$  waiting for admission) do
3:     if (there exists a feasible power allocation for all flows already ad-
        mitted and the newly incoming flow) then
4:       admit the new flow
5:     else
6:       reject the new flow
7:     end if
8:   end for
9: end for

```

---

### 3.4 Integrating Admission Control with Power Prediction

An important aspect of this thesis is the consideration of flows that have time duration long enough for the target end users to travel a considerable distance from the base station while receiving the flow. It is possible that, during such a time period, the users change their locations, and thereby require a total transmission power that exceeds the available base station power. If such power shortage occurs frequently, many packets are likely to miss their delay expiry time, which will cause a loss in the system's effective throughput. We recall that effective throughput is defined as follows: given a prescribed fraction  $\rho$ , we count a flow to be successfully delivered if the system manages to deliver at least  $\rho$  of the flow's packets prior to their expiry time; otherwise, the system fails to deliver the flow. The effective throughput is then obtained by restricting our attention to the successfully delivered flows and ignoring the failed flows.

To minimize the risk of admitting flows that are likely to cause the base station to have a power shortage at some instants in the future, we adopt a simple mechanism that aims at predicting the base station power requirements as mobile users move near or away from the base station. The mechanism is applied to a random mobility model with parameters set to induce a high mobility environment. More specifically, in the adopted random mobility model, each user travels for a certain amount of time (e.g., 3 seconds), at a relatively high speed (e.g., 10 meters/second) before picking a new direction from a well defined set of directions (e.g., the 4 directions: north, east, south, and west).

Now, suppose we would like to assess whether or not the base station will suffer from a power shortage during an interval  $t_{pred}$  in the future. We sample the space of outcomes in the following way: we conduct  $n_{trial}$  trials (e.g.,  $n_{trial} = 3$ ). Each trial involves a number of checkpoints (the checkpoints are equally spaced, and separated by a time interval  $t_{interval}$  of prescribed length;

---

**Algorithm 3** A predictive call admission control algorithm

---

**prediction parameters:**

- $t_{pred}$ : total prediction duration
- $t_{interval}$ : checkpointing is done every  $t_{interval}$  seconds
- $n_{trial}$ : the number of trials conducted
- $p_{success}$ : a threshold for considering a successful power allocation

**algorithm**

- 1:  $n_{success} = 0$
  - 2: **for** ( ;  $n_{trial} > 0$ ;  $n_{trial} = n_{trial} - 1$ ) **do**
  - 3:   **for** ( $t = t_{interval}$ ;  $t \leq t_{pred}$  ;  $t = t + t_{interval}$ ) **do**
  - 4:     perform mobility prediction: update all mobiles' locations up to time  $t$  in the future; if feasible power assignment does not exist for all flows already admitted and the newly incoming flow, exit the loop (i.e., go to step 6).
  - 5:   **end for**
  - 6:   **if** (feasible power assignments exist for all  $\lfloor t_{pred}/t_{interval} \rfloor$  checkpoints) **then**
  - 7:      $n_{success} = n_{success} + 1$
  - 8:   **end if**
  - 9: **end for**
  - 10: **if** ( $(n_{success}/n_{trial}) < p_{success}$ ) **then**
  - 11:   reject the new flow
  - 12: **else**
  - 13:   admit the new flow
  - 14: **end if**
-

thus,  $\lfloor t_{pred}/t_{interval} \rfloor$  checkpoints are examined during a prediction interval of length  $t_{pred}$ ).

At each checkpoint, we simulate the random movements of the users, and check whether there exists a feasible assignment of the base station power to each user. If there is no feasible power assignment at some checkpoint, then the corresponding trial fails (and there is no need to evaluate any remaining checkpoints in the trial). On the other hand, if a feasible power assignment exists for all  $\lfloor t_{pred}/t_{interval} \rfloor$  checkpoints in a trial, then the trial succeeds. We consider each trial to be one sample, and design our predictive call admission control to accept a flow if the ratio between the number of successful trials (denoted as  $n_{success}$ ) and the total number of trials  $n_{trial}$  exceeds a certain threshold, denoted as  $p_{success}$ .

The modified predictive call admission control works as follows: at the beginning of each admission cycle, the admission control considers flows from the highest priority (Class 1) to the lowest priority class (Class N). Within each class, flows are considered on a first-come-first-served basis. Each flow is tested by simulating the system, with the flow under test being added to the system. Algorithm 3, which performs the predictive sampling approach described above is then applied to the augmented system state.

We now illustrate the operation of Algorithm 3 using some numerical values. Let us assume that the flow under test has a total length of  $t_{flow} = 60$  seconds. Moreover, let us assume that the algorithm is set to sample the system for  $t_{pred} = 0.25 * t_{flow} = 15$  seconds. If Algorithm 3 performs checkpointing every  $t_{interval} = 0.4$  seconds, then the algorithm considers approximately 37 checkpoints in each trial. If at least 2 out of the 3 trials succeed, the modified CAC accepts the new flow. Subsequently, the scheduling algorithm (Algorithm 1) is applied to schedule packet transmission for all admitted flows.

## 3.5 Summary

In this chapter, we examined some basic mechanisms for provisioning proportional delay differentiation on the downlink of a cellular CDMA environment. Our objective is to devise a set of call admission control and scheduling schemes that maximize the effective throughput of the system, subject to a constraint of the total base station power budget, and the requirement of achieving target proportional delays among DiffServ classes at moderate congestion times.

The proportional delay differentiation scheme is identified in the thesis as a suitable PHB for satisfying the derived requirements, and two PDD schedulers are proposed. We also examine call admission control schemes running at the Internet-wireless gateway that work in conjunction with the PDD scheduling algorithms to implement PDD PHB. A basic call admission control scheme is reviewed first; we then propose an integrated admission control and power prediction mechanism that uses mobility prediction to estimate the base station power required by all mobiles during a specified future period.

## Chapter 4

# Downlink Performance Evaluation

In Chapter 3, we have discussed methods of estimating the average delay incurred by packets in each DiffServ class, while awaiting downlink transmission from the Internet-wireless gateway to end users. Chapter 3 also describes two call admission control approaches for provisioning proportional delay differentiation on the downlink, while taking the user mobility and CDMA soft capacity into consideration.

In this chapter, we use simulation to analyze the performance of the above approaches. The simulation realization considers a challenging situation, where the required quality of service applies to flows that are sufficiently long to allow the corresponding mobile users to travel a significant distance relative to their initial positions when the flows are first admitted.

This chapter is organized as follows. Section 4.1 outlines the main simulation parameters used in the study. In Section 4.2, we investigate the performance of the two proportional delay differentiation schedulers proposed in Section 3.2. Section 4.3 examines the performance of proportional delay differentiation scheduling without call admission control in terms of the achieved class delay, delay ratios, and effective throughput. In Section 4.4, we incorporate the call admission control scheme presented in Section 3.3 and

investigate its performance. Section 4.5 presents performance results when integrating admission control with power prediction, as proposed in Section 3.4. Section 4.6 summarizes the chapter.

## 4.1 Overview of the Simulation Parameters

This section presents the main parameters used in the simulation study. The parameters can be conveniently classified into the following categories:

- (a) cell and user mobility parameters,
- (b) radio propagation path loss and CDMA parameters, and
- (c) traffic parameters.

Tables 4.2 and 4.3, at the end of the section, summarize the important values.

### 4.1.1 Cell and User Mobility Parameters

Our simulation is based on the widely used 19 cells model. We assume a central cell surrounded by two tiers of neighboring cells. We focus on analyzing the admission control and scheduling for the central cell and consider signal interference coming from the other 18 cells. Figure 4.1 shows the geometry structure of the 19 hexagonal cells.

Handoff flows are not simulated in our evaluation. All the end users only move around in the central cell, which has a radius of 1000 meters. In the simulation, the maximum number of mobiles in each cell is 40.

During the simulation, mobiles are randomly generated. Each mobile is associated with its current position and speed. For each flow, the associated mobile is assumed to be located at an initial location. We also assume a high mobility environment, where all mobiles are moving at the speed of 10 meters per second. The direction of mobiles' movement is selected randomly from 8 possible angles,  $0, \frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi, \frac{5\pi}{4}, \frac{3\pi}{2}, \frac{7\pi}{4}$ . The distance of one movement is

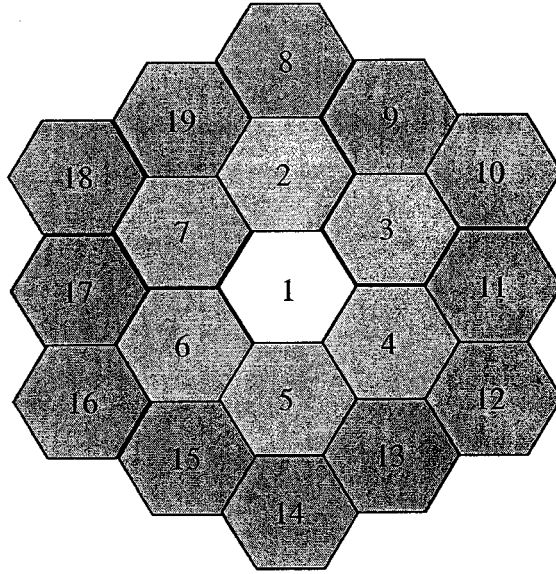


Figure 4.1: 19-cell simulation model

calculated as the product of mobile speed and the past time interval since last move.

The mobile's location is updated every a prescribed update interval, which is 3 seconds in our simulation. In our simulation realizations, the Internet-wireless gateway is assumed to receive power usage information at intervals shorter than the time intervals during which users change their locations.

#### 4.1.2 Radio Propagation Path Loss and CDMA Parameters

Path loss decides the received signal strength, which depends on the distance between the receiver and transmitter and also the environment. Propagation models focus on predicting signal strength at the receiving end at a given distance from the transmitter [33]. There exists a number of theoretical path loss models including the Friis free space model, the 2-ray ground reflection model, and the knife-edge model. The path loss model used in our simulation is the *log-normal shadowing model*.

Both theoretical and practical findings indicate that the average received signal strength decreases logarithmically with distance independent of the environment. Measurements have also shown that, at distance  $d$ , the path loss of a particular location is distributed log normally in dB. The log-normal shadowing model [33] states that:

$$PL(d)[dB] = \overline{PL}(d) + X_\sigma = \overline{PL}(d_0) + 10n \log \left( \frac{d}{d_0} \right) + X_\sigma \quad (4.1)$$

where  $d_0$  is a reference point,  $\overline{PL}(d_0)$  is the average path loss (in dB) at  $d_0$ ,  $n$  is an exponent varying with different environments, and  $X_\sigma$  is a 0 mean Gaussian distributed random variable with standard deviation  $\sigma$  in dB.

The exponent  $n$  varies with different environments. For example, for free space,  $n = 2$  has been found to be a suitable value. For environments obstructed with buildings,  $n = 4$  to 6 fit the empirical observations. In our simulation, the reference point  $d_0$  and its path loss in Equation 4.1 is respectively 711 meters and 142 dB (see [7], for example). The exponent  $n$  is of value 4 and the standard deviation  $\sigma$  is 5 dB.

Other CDMA parameters in our simulation are shared with the CDMA-2000 proposal [7]. We use typical W-CDMA parameters. The chip rate parameter is 4.096 Mcps; convolutional coding rate is 1/3; and  $E_b/I_0$  requirement is 7 dB. We assume all mobile users require the same  $E_b/I_0$  target ratio. Other radio link parameters are also commonly used ones, including the noise spectral density which is  $3.98e-18$  mwatt and the orthogonality factor which is 0.2.

### 4.1.3 Traffic parameters

In our simulation study, traffic generation is not intended to capture flows from any realistic application. Rather, it is intended to generate a workload that is likely to cause no congestion at lower speeds (e.g. less than 64 Kbps), and a definite congestion at higher speeds (e.g. greater than 128 Kbps). This section describes the traffic generation process and the setting

of the associated parameters in our simulation study. We also use a simple numerical argument to show that the chosen parameter values satisfy the following property: as the user requested data rates increase in the range from 16 Kbps to 192 Kbps, the offered traffic workload drives the system from a no congestion state to a heavily congested state.

At the Internet-wireless gateway, each received flow encapsulates variable number of packets. The duration of each flow is uniformly distributed from 60 seconds to 90 seconds, which allows significant user mobility. For one mobile, the interarrival time between flows is exponentially distributed with a mean value of 30 seconds.

If the entire flow is admitted, the transmission is in terms of packets. All packets of the flow are generated by Poisson process including packet length and packet interarrival time. The length of the packet is generated using Poisson process with mean packet length of 420 bytes. The interarrival time of between packets is also generated using Poisson process with mean packet interarrival time of 1 second. To keep the analysis simple, it is assumed that the packet expiry time is fixed for all packets. If a packet is not delivered within the maximum delay limit 6 seconds, the packet is dropped.

We now argue that the above traffic parameters generate a workload that satisfies the property mentioned above. In the first part of the argument, we estimate the maximum number of users  $N_{max}$ , that can be served by a base station transmitting to all users continuously at the same data rate  $R$ . To estimate  $N_{max}$  as a function of  $R$ , we assume some reasonable values for the remaining parameters (e.g., all users are located at a distance of 700 meters from the base station, the total base station transmission power is 25 watts, and the  $\frac{E_b}{I_0}$  requirement is 7 dB for all mobiles).

The second part of the argument adjusts the derived  $N_{max}$  figures by taking into account the fact that users in our present context do not receive data continuously. Rather, packets are separated by some interarrival times. The modified maximum number of users, denoted as  $N_{real,max}$ , that can be accommodated by the system (as a function of  $R$ ), is then used to reason

about the offered workload.

To estimate  $N_{max}$  under the above circumstances, we only consider the first tier of other cell interference, which is roughly 1300 meters away from the target mobile based on the cell radius of 1000 meters. The SIR of the target mobile should satisfy the following inequality:

$$\frac{E_b}{I_0} \leq \frac{\frac{W}{R} \times \frac{P_{tr}}{PL(700m)}}{\sum_{j \neq i} \gamma \frac{P_{tr}}{PL(700m)} + \sum_{tier \#1} \frac{P_{max}}{PL(1300m)} + \eta_0 W} \quad (4.2)$$

where  $W$  is the chip rate,  $R$  is the data transmission rate,  $P_{tr}$  is the transmission power,  $PL$  is the path loss,  $\gamma$  is the orthogonality factor, and  $\eta_0$  is the noise spectral density.  $P_{max}$  is the base station maximum transmission power which is 25 watts.

If we multiply the numerator and the denominator simultaneously with  $PL(700m)$  and replace each symbol in Relation 4.2 with the real simulation values, we obtain the following inequality.

$$(N - 1)P_{tr} + 105 \leq \frac{W}{R}P_{tr} \quad (4.3)$$

where  $N$  is the total number of mobile users in the cell. We assume the transmission power is the same for all mobiles, because they all ask for the same data rate, they are located at the same distance from the base station, and they require the same  $\frac{E_b}{I_0}$  value in our assumption. Thus, the following inequality should also hold.

$$NP_{tr} \leq 25 \quad (4.4)$$

where 25 is the total base station power budget.

The maximum number of users  $N_{max}$  in the cell is achieved when inequalities 4.3 and 4.4 hold as equations. Table 4.1 tabulates  $N_{max}$  as a function of  $R$ , for  $R = 32K$ ,  $64K$ , and  $128K$  bps.

$N_{max}$  is the maximum number of users if the base station keeps transmitting to mobile users continuously. However, in our simulation, each flow is transmitting in the format of packets, and packets are arriving with a mean

Rate (Kbps)	$\frac{W}{R}$	$P_{tr}$ (watt)	$N_{max}$	$N_{real\_max}(N_{max} \times \frac{1}{0.4})$
32	128	1.008	24.8	62
64	64	2	12.5	31
128	32	4	6.25	15

Table 4.1: Maximum number of mobile users for different rates

interarrival time. The power update interval, in the simulation, is 0.4 seconds. If we set the mean packet interarrival time to be 1 second, the real maximum number of users the system can support should be  $N_{max} \times \frac{1}{0.4}$ , which are the values in the last column of Table 4.1. We can see that for 32 Kbps, the maximum number of users the system can afford is 62, which is far more than 40 – the maximum number of users allowed in our simulation. For 64 Kbps,  $N_{real\_max}$  is 31, which is a little bit smaller than 40. However, because not all the users have traffic for transmission at any instant, for 64 Kbps, the system still can work well, but it is approaching the system capacity limit. However, for 128 Kbps, the number of users the system can afford is only 15, which is much less than 40. At this transmission rate, the system is definitely under congestion. Thus, by setting the packet interarrival time to be 1 second, our traffic generation objective is reached. For lower transmission rates, we generate a region of almost no congestion. For medium speeds, the system is moderately congested when the proportional delay differentiation target ratios should be achieved. For high rates, the system is heavily congested to the extent that the average delay in each class approaches the preset maximum expiry time. In the simulation, we have 8 transmission rates ranging from the low 16 Kbps to the high 192 Kbps.

#### 4.1.4 Summary of Simulation Parameters

All the CDMA and power related parameters in the simulation are listed in Table 4.2. Other simulation parameters including cell, mobile, flow, and mobility parameters are listed in Table 4.3.

Base station has a total power budget of 25 watts in the simulation. All the downlink transmission is limited by this power bound. The system updates its power allocation each 0.4 seconds. Each run in the simulation takes 2 simulation hours. Each data in the remaining sections is an average of 3 random simulation runs. When calculating the effective throughput, the fraction  $\rho$ , which is the fraction of successfully delivered packets prior to their expiry time in a flow, is 0.9.

Parameters	Value	Unit
Chipping rate	4.096	Mcps
Noise spectral density ( $\eta_0$ )	$3.98e-18$	mwatt
Orthogonality factor ( $\gamma$ )	0.2	
Convolutional coding rate	1/3	
$E_b/I_0$ requirement	7	dB
Log-normal shadowing exponent ( $n$ )	4	
Log-normal shadowing reference point ( $d_0$ )	711	meters
Reference point average path loss	142	dB
Log-normal shadowing standard deviation ( $\sigma$ )	5	dB
Base station power budget	25	watts
Power update interval	0.4	sec

Table 4.2: CDMA and power parameters

Parameters	Value	Unit
Cell radius	1000	meters
Maximum number of mobiles per cell	40	
Average mobile speed	10	m/sec
Number of directions a mobile can follow	8	
Interval before a mobile picks a new direction	3	sec
Flow duration	[60 - 90]	sec
Mean flow interarrival time	30	sec
Mean packet interarrival time	1	sec
Mean packet length	420	bytes
Maximum packet delay limit	6	sec
Simulation time per run	2	hours
Number of random simulation runs for each parameter set	3	
Predictive CAC number of trials ( $n_{trial}$ )	3	
Predictive CAC success probability threshold ( $p_{success}$ )	2/3	
Effective throughput success packets fraction threshold ( $\rho$ )	0.9	

Table 4.3: Other simulation parameters

## 4.2 Performance results on Aggregate Delay Calculation Schemes

In Section 3.2, we outlined two possible methods for estimating the average delay incurred by the queued packets in each DiffServ class at any instant  $t$ . The two methods give rise to two different scheduling schemes for utilizing the downlink bandwidth.

The first method, called the head-of-queue (HOQ) method, computes the average delay of the  $i$ th DiffServ class by averaging the time delays incurred by the first packet in each queue of the class. The second method, denoted as WIN, computes the average delay of the  $i$ th class by averaging the window-size number of most-delayed packets in each DiffServ class. From the data structure point of view, implementing the second method requires more storage and computations than the first method.

In this section, we investigate the performance of the resulting scheduling methods using the achieved per-class average delay, as well as the achieved delay ratios. We recall that the simulation parameter set is that there are three DiffServ classes in the simulation, denoted as Class 1, Class 2, and Class 3. We want to provide the highest forwarding assurance to Class 1 and the lowest forwarding assurance to Class 3. The delay weights for Class 1, Class 2, and Class 3 are 4, 2, and 1 respectively. Traffic is equally distributed among the three classes. Each class's traffic occupies 33.3% of the total traffic volume. In this section, no admission control scheme is used. The WIN scheduler uses window-size of 4 when compared with the HOQ scheduler.

Figure 4.2 displays the average Class 1, Class 2, and Class 3 delay for both the HOQ and the WIN schedulers. Figure 4.3 examines the delay ratios among classes for each transmission rate. There are two series of ratios. One is Class 2 to Class 1 delay ratio which is expected to be around 2, and the other is Class 3 to Class 1 delay ratio which is expected to be around 4. The results for the WIN scheduler are in solid lines and the HOQ scheme is represented in dotted lines.

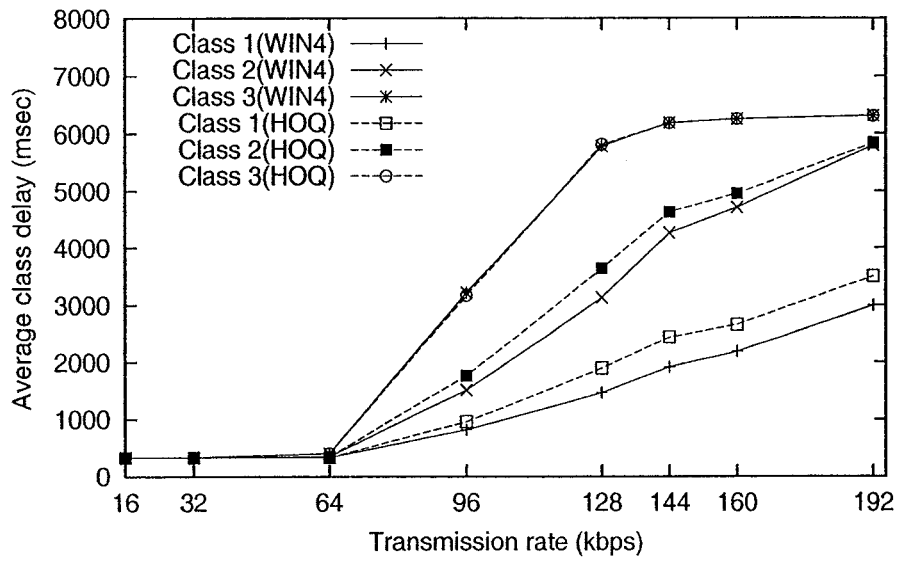


Figure 4.2: WIN & HOQ comparison – average class delay

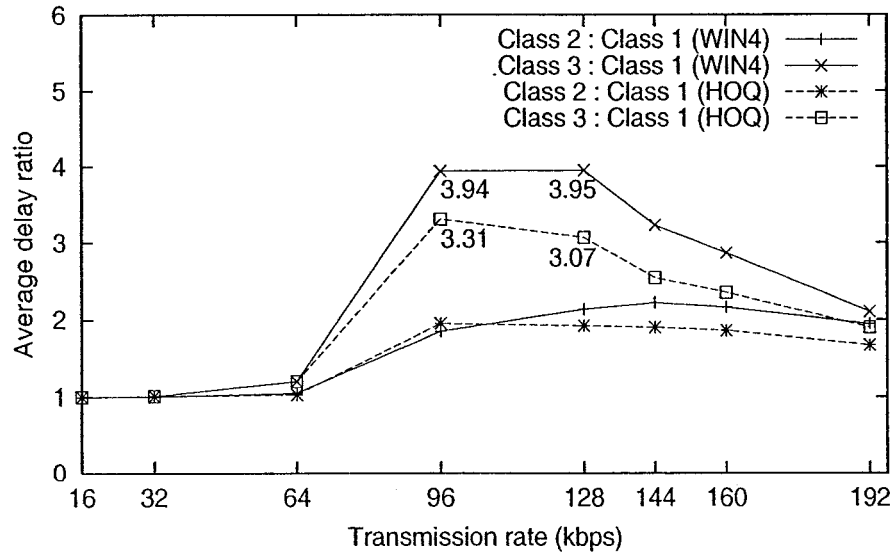


Figure 4.3: WIN & HOQ comparison – average delay ratio

The WIN scheme outperforms the HOQ scheme in terms of Class 1 and Class 2 average delay in Figure 4.2. For Class 1 and Class 2, the WIN scheme generates lower delay than the HOQ scheme. From 96 Kbps to 192 Kbps, the Class 1 delay difference between the WIN and the HOQ schedulers is about 0.5 seconds, and the Class 2 delay difference is around 0.3 seconds. Furthermore, in Figure 4.3, the WIN scheme brings more exact delay ratios than the HOQ at moderate congestion times. For rates 96 Kbps and 128 Kbps, the HOQ scheme can only generate Class 3 to Class 1 delay ratio of 3.31 and 3.07 respectively which is lower than the expected ratio 4, while the WIN scheme can reach the target ratio better (3.94 and 3.95).

The performance difference is because the HOQ takes the average of head-of-queue packet's delay in each queue. Back to CDMA queue organization, we recall that one flow is kept in one queue. If one flow is coming much later than the other ones, when we take its head-of-queue delay into consideration, it actually plays an negative effect on the calculated average aggregate delay. This calculated average aggregate delay blocks the representation of the delay characteristics for this class. However, in the WIN scheme, it only considers a set of most-delayed packets in each class, no matter which flow they belong to. Thus, in the long run, HOQ brings higher average class delay and less accurate delay ratio among classes at moderate congestion times, because some highly delayed packets may not be transmitted due to the influence caused by the delay distribution of head-of-queue packets.

For all the experiments of the WIN scheme, we use window size of 4 packets. We also test other window sizes including 1, 2, 6, 8, 10, and 30. There is not much difference among the results except that, when the window size increases, the observed results are not as accurate as the results from window size 1, 2, and 4. We select 4 as our default experiment window size because we want to average the delay to some extent, and at the same time, keep it unaffected by less-delayed packets.

## 4.3 Performance Results without Call Admission Control

In Section 3.1, we outlined three distinct behaviours of a router using the PDD PHB. The first is that no queue builds up for either class, and the base station power suffices to serve all the packets as they arrive, where there is almost no congestion. The second behaviour is that queues build up, and most packets do not miss their expiry time, where the system is moderate congested. The last situation is that the router is heavily congested, and most packets miss their expiry time. We also stated that the objective of the thesis is to maximize the effective throughput under the requirement of achieving target proportional delays among the various DiffServ classes at moderate congestion times.

In this section, we investigate the proportional delay differentiation performance without call admission control under the three different system congestion levels in terms of the achieved average class delay, delay ratios among classes, and effective throughput. We examine whether the proportional delay differentiation can achieve the target proportional delays among difference classes at moderate congestion level, and its performance under the other two congestion levels.

The simulation parameters used in this section is as follows. Delay weights for Class 1, Class 2, and Class 3 are 4, 2, and 1 respectively. The incoming traffic is uniformly distributed among the three classes. Proportional delay differentiation is using WIN scheduler of window-size 4. The simulation, in this section, is performed without call admission control.

### 4.3.1 Average class delay

Figure 4.4 shows Class 1, Class 2, and Class 3 average class delay with time changing. The data transmission rate in this figure is 128 Kbps. We sample a total of 150 seconds (2.5 minutes) of the whole 2 hour simulation time from 150 second to 300 second. Each data point in the figure is the

average aggregate queuing delay of all packets in that class at the time instant specified.

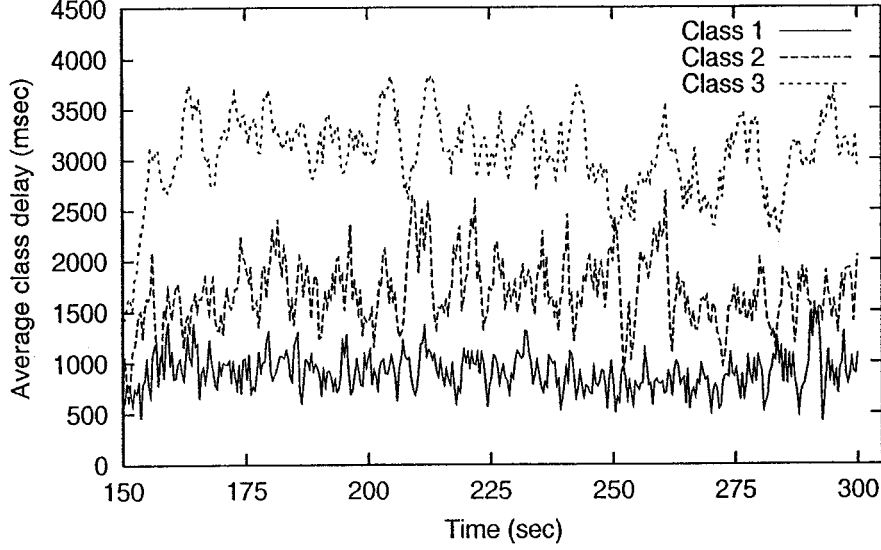


Figure 4.4: No admission control – average class delay

From Figure 4.4, we have the following observations: Class 1's delay is mostly in the range of [800 - 900] msec; Class 2's delay is mostly in the range of [1600 - 1800] msec; Class 3's delay is mostly in the range of [3000 - 3500] msec.

According to the delay weights of 4:2:1, this result is just what we expect. Class 1's delay is approximately half of Class 2's delay, which is in turn approximately half of Class 3's delay. The expected proportional delay among classes is achieved at 128 Kbps transmission rate when we take a detailed look during the simulation.

### 4.3.2 Average delay ratio

Figure 4.5 is the average delay ratio achieved under different data transmission rates. According to the analysis in Section 4.1.3, different transmission rates create different congestion levels. In Figure 4.5, we explore how delay

differentiation scheme performs under different system congestion. The dotted line represents Class 3 to Class 1 delay ratio, which is expected to be around 4. The solid line is Class 2 to Class 1 delay ratio, which is expected to be around 2.

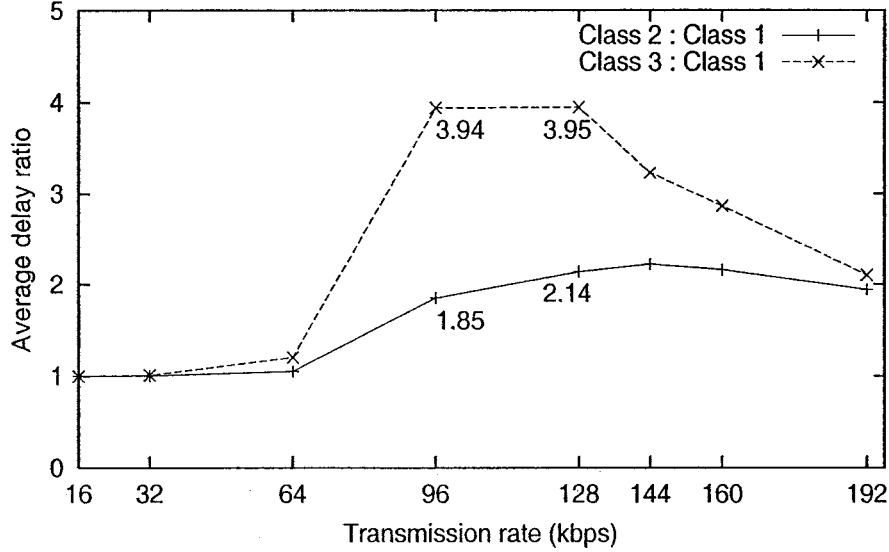


Figure 4.5: No admission control – average delay ratio

Firstly, each data in Figure 4.5 is above 1 and the Class 3 to Class 1 delay ratio line is always above the Class 2 to Class 1 delay ratio line for each rate. Thus, for each transmission rate, proportional delay differentiation guarantee that Class 1's delay is always lower than Class 2's delay, which is in turn lower than Class 3's delay. In general, a higher forwarding assurance class is guaranteed a lower average class delay.

There are three regions of different characteristics in Figure 4.5. The first region ranges from 16 Kbps to 64 Kbps. The second region is from 96 Kbps to 128 Kbps. 144 Kbps to 192 Kbps form the last region. These three regions of different delay ratio characteristics are a result of different congestion levels caused by different downlink transmission rates.

For the low rate part from 16 Kbps to 64 Kbps, the expected delay ratios can not be achieved. All the delay ratios are a little bit above 1. At low

transmission rates, the system capacity is high, delay differentiation scheduler does not play as much an important role as in the case of network congestion. Most of the time, there is no need of delay differentiation scheduling because all of the traffic can be transmitted as they arrive under the current base station power budget, and few packets experience serious delay. Thus, the class average delay is almost the same for all classes and the expected delay ratio can not be achieved. Take the speed of 64 Kbps as an example, the Class 1, Class 2, and Class 3's delay is 0.33, 0.35, and 0.4 seconds respectively. In the simulation, the mean packet interarrival time is 1 second, which indicates that almost no queue builds up for any class at 64 Kbps transmission rate.

However, for the medium rate part ranging from 96 Kbps to 128 Kbps, the target proportional delays among the 3 DiffServ classes are achieved. The ratios are marked in Figure 4.5. The delay ratios between Class 3 and Class 1 are almost 4 and the delay ratios between Class 2 and Class 1 are near 2. For 128 Kbps, the average class delay from Class 1 to Class 3 is 1.5, 3.1, and 5.8 seconds respectively. Queues build up because the average delay is higher than the packet interarrival time 1 second. However, many packets do not miss their 6 seconds expiry time. Downlink transmission rates 96 Kbps and 128 Kbps generate a moderate congestion level, under which the target proportional delays are achieved as the objective of the thesis states.

For the high rate part from 144 Kbps to 192 Kbps, the system is heavily congested and most packets miss their expiry time. In high rate transmission, the system capacity is relatively low. The scarce resource leads to high class delay, which approaches the maximum delay limit 6 seconds. For example, the average class delay for 160 Kbps, in our simulation is 2 seconds for Class 1, 5 seconds for Class 2, and 6 seconds for Class 3. We observe from Figure 4.5 that Class 2 to Class 1 delay ratio can still be maintained because Class 1 always has priority over Class 2 and Class 3, and its delay still can be guaranteed relatively low under this base station power budget. However, for Class 2, its delay is already very high, as well as Class 3's delay. Their delays are both around the 6 seconds limit. There is no way to differentiate

Class 2 and Class 3's delay under such circumstance. This is why the Class 3 to Class 1 target delay ratio can not be maintained from 144 Kbps to 192 Kbps.

### 4.3.3 Effective throughput

Figure 4.6 examines throughput performance of delay differentiation model. The effective throughput is calculated as the total throughput of flows, whose successfully delivered packet ratio is above  $\rho$ , which is 0.9 in the simulation.

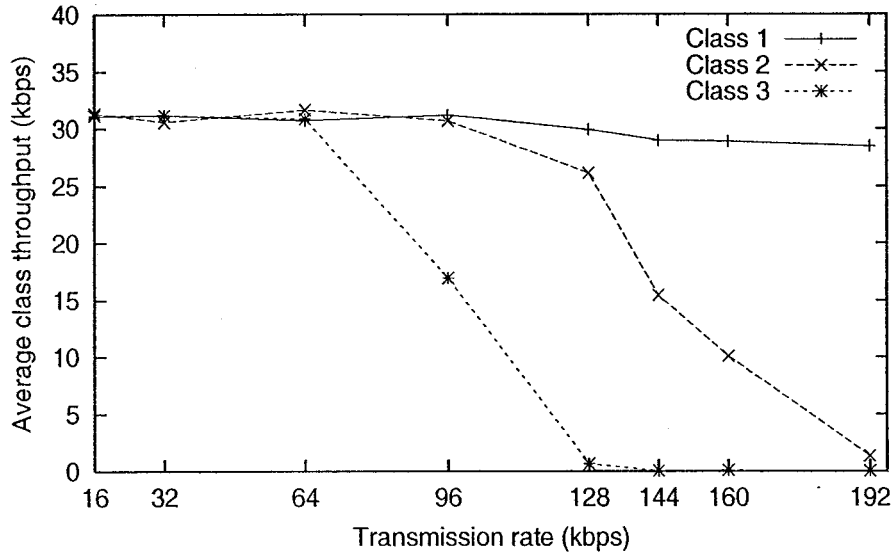


Figure 4.6: No admission control – effective throughput

From 16 Kbps to 64 Kbps, almost all of the packets can be transmitted as they arrive. In our simulation result, for these three rates, no packet is dropped due to delay. Thus, the throughput does not make big difference for different classes (Recall that traffic is evenly distributed among classes).

However, for 96 Kbps to 192 Kbps, the system capacity becomes lower and lower. Delay differentiation model determines the transmission priority among different classes and, in turn, the throughput. The Class 1's throughput is always maintained at the same level because it has the highest forwarding assurance. While, Class 2 and Class 3's throughput is sacrificed in

order to maintain the lower delay in Class 1. And with the transmission rate increasing, Class 2 and Class 3's traffic begins starving. Without admission control, all traffic is competing for the limited resource. Under this circumstance, when performing delay differentiation scheduling, a higher class is guaranteed a lower delay at the cost of lower classes' starving. Throughput unfairness among classes happens when the system is congested, and this unfairness is more notable when the system is heavily congested.

## 4.4 Performance Results Using Call Admission Control

This section presents results on average delays and effective throughput when the system incorporates the call admission control scheme presented in Section 3.3.

In the simulation realization, a newly incoming flow is assumed to be buffered at (or beyond) the edge of the distinguished DiffServ domain containing the target RNC, base station, and user of interest. Buffering delays outside the target RNC are not accounted in the results presented in this section. The results show delays incurred by the packets after admitting the flow.

Upon receiving a request to admit a new flow, the call admission control scheme makes a decision based on available base station power at the request arrival time, as described in Section 3.3. The new flow is admitted if there exists sufficient base station power to transmit all already admitted flows and the newly incoming flow at current instant.

Intuitively, the anticipated effects of incorporating call admission control include bringing down the system's congestion level, which in turn reduces the delay incurred by packets. Through admission control, the amount of traffic coming into the system is decreased, less traffic is competing for the same total amount of base station power budget. Theoretically, the delay incurred by packets in each class will be lower than that without admission

control. We also anticipate that the effective throughput can be increased by incorporating call admission control. Admission control brings less system congestion, which means more packets in a flow will be delivered successfully to end users. Thus, the effective throughput of the system should be improved.

In this section, our objective is to investigate the performance of call admission control in terms of the achieved average class delay and effective throughput.

We recall the simulation parameters in this section that there are three DiffServ classes with delay weights of 4, 2, and 1 respectively. Traffic is evenly distributed among the various classes. WIN scheduler of window-size 4 is used. We compare the results with admission control to those without admission control.

#### 4.4.1 Average class delay

Figure 4.7 is the result of average class delay without admission control, and with admission control.

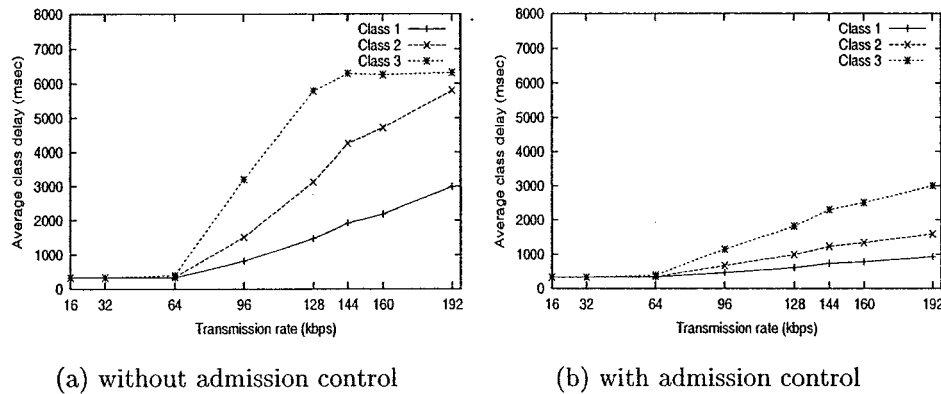


Figure 4.7: Call admission control – average class delay

Beginning from 96 Kbps, the average class delay incurred by packets in each class with admission control in Figure 4.7(b) is less than the average class

delay without admission control in Figure 4.7(a). This observation proves part of our anticipation. Call admission control reduces the average delay incurred by packets in each class, because it limits the amount of traffic in the system, and brings the system congestion level down. The delay incurred by packets is lower with admission control.

In Figure 4.7(b), there is no heavily congested region as there is without admission control (i.e., from 144 Kbps to 192 Kbps in Figure 4.7(a)). None of the average class delay in Figure 4.7(b) is approaching the 6 seconds limit. This observation shows that call admission control can keep the system away from exhibiting the heavily congested behaviour. Furthermore, in Figure 4.7(b), the average delay of Class 1 for each transmission rate is always lower than 1 second, which means that most of the Class 1 packets can be transmitted within the mean packet interarrival time 1 second before the next packet arrives. Almost no queue builds up for Class 1 packets with admission control.

#### 4.4.2 Effective throughput

Figure 4.8 investigates the performance in terms of effective throughput for each DiffServ class. There are two sub-figures for comparison. Figure 4.8(a) is the result without call admission control, which is the same as Figure 4.6 that we have analyzed. Figure 4.8(b) is the performance result with admission control.

With admission control in Figure 4.8(b), the effective throughput for Class 1 and Class 2 is almost the same. However, Class 3's throughput is still lower than Class 1 and Class 2's. Admission control brings more fairness among classes than without admission control, where lower classes' traffic has to sacrifice themselves to guarantee lower delay in higher classes. With admission control, every class always has a share of the total throughput, and lower classes' starving is avoided.

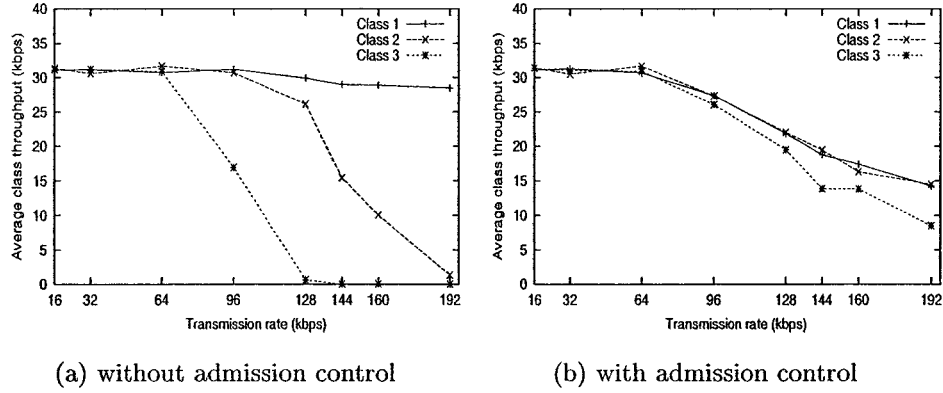


Figure 4.8: Call admission control – effective throughput

## 4.5 Performance Results Using Power Prediction

This section presents the performance results on average class delays and effective throughput when integrating admission control with power prediction proposed in Section 3.4.

The call admission control scheme proposed in Section 3.3 is evaluated in the previous section and it is shown that significant improvement in average class delay and throughput fairness is achieved by this admission control mechanism. However, due to the mobility nature of a cellular CDMA environment, admitting a flow, based on the current availability of sufficient base station power, may lead to future power shortage when current available power can not afford the transmission to all mobiles after they have moved to new locations. A novel feature of this work is to integrate admission control with power prediction, as described in Section 3.4. Mobility prediction is adopted to estimate the base station power required by all mobiles in the future.

By integrating admission control with power prediction, we expect the average class delay and throughput will be further improved, because power

prediction gives a future estimate of base station power availability, and lowers the chance of power shortage in the future.

We have two data series in this section. One is power prediction with a duration of 5% flow duration, and the other is power prediction with a duration of 10% flow duration. 5% of the flow duration (60 seconds to 90 seconds) is 3 seconds to 4.5 seconds, which allows one mobility update in the simulation with mobility update interval of 3 seconds. This allows a mobile user to move 30 meters to 45 meters with a speed of 10 meters per second in a cell of radius 1000 meters. 10% time prediction allows 2 to 3 mobility updates, when a mobile user can traverse 60 to 90 meters during the prediction time.

#### 4.5.1 Average class delay

Figure 4.9 shows the performance results of average class delay with 5% time prediction and 10% time prediction.

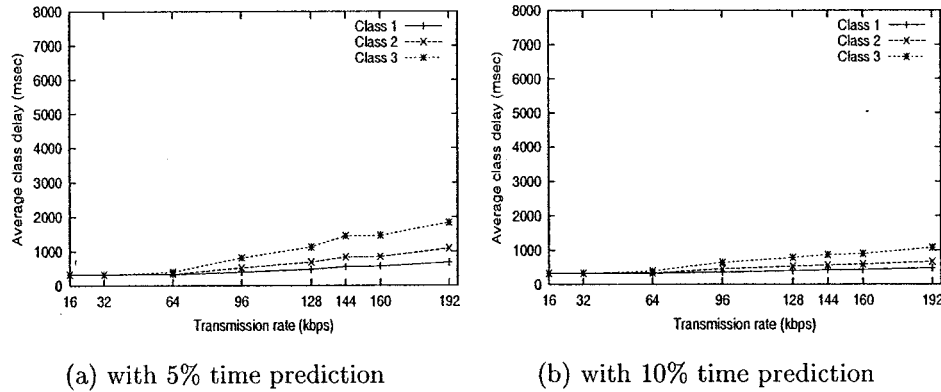


Figure 4.9: Power prediction – average class delay

From the above figures, we observe that power prediction brings even lower average class delay than the admission control method presented in Section 3.3 (Figure 4.7(b)). The longer prediction time 10% (4.9(b)) generates even lower delay than 5% (4.9(a)) time prediction. By performing

mobility prediction to estimate the base station power availability, power prediction outperforms the simple admission control scheme in terms of average class delay. With 5% time prediction, there is almost no queue builds up for either Class 1 or Class 2 since their average class delays are both under 1 second. With 10% time prediction, almost no queue builds up for any class. Prediction can drive the system from moderate congested state to almost no congestion state as in Figure 4.9(b).

### 4.5.2 Effective throughput

In this section, we review the performance results of effective throughput when integrating admission control with power prediction. Figure 4.10 shows the effective throughput for each class with 5% and 10% time power prediction.

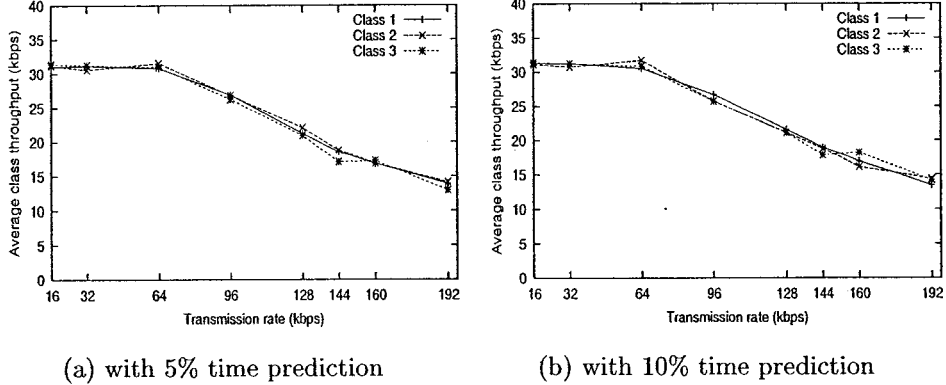


Figure 4.10: Power prediction – effective throughput

Compared with Figure 4.8, power prediction brings more fairness among different classes than the admission control scheme in Figure 4.8(b). In Figure 4.10(a) and 4.10(b), the throughput for each class is almost the same. Power prediction gives a way of providing throughput fairness among the various DiffServ classes. Instead of starving lower forwarding assurance classes

in order to guarantee lower delay in higher classes, predictive admission control achieves the same level of effective throughput for all DiffServ classes.

### 4.5.3 System total effective throughput

One of our thesis objective stated in Section 3.1 is to maximize the total effective throughput of the system. Thus, in this section, the total system effective throughput is explored under four different admission control scenarios. One is without admission control, one is with admission control mechanism presented in Section 3.3, the third is with 5% time power prediction, and the last is with 10% time power prediction. Figure 4.11 gives out the performance results.

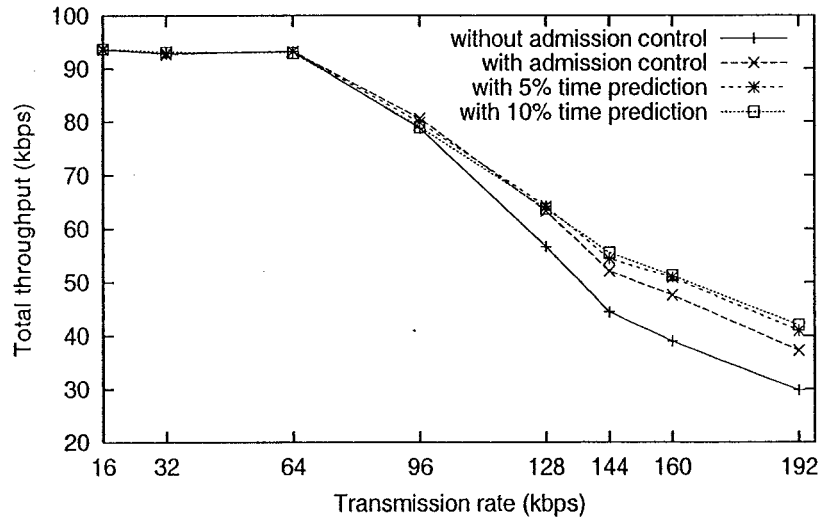


Figure 4.11: System total effective throughput

In Figure 4.11, the total effective throughput decreases with the CDMA downlink transmission rate increasing. Back to the analysis in Section 4.1.3, the number of affordable mobile users decreases with the transmission rate increasing, and the downlink CDMA capacity decreases. Consequently, the total effective throughput decreases.

In Figure 4.11, beginning from 96 Kbps, the effective throughput without

admission control is much lower than those with admission control and with power prediction. The total effective throughput with admission control is approximately 8 Kbps higher than that without admission control. Furthermore, beginning from 144 Kbps, predictive admission control (both 5% and 10% time prediction) has a higher effective throughput than with admission control. And 10% time prediction generates even higher throughput than 5% time prediction. 5% time prediction has an effective throughput which is approximately 3 Kbps higher than that with admission control, and 10% time prediction's throughput is about 1 Kbps higher than that with 5% time prediction.

From the above observations, we can conclude that admission control improves the total effective throughput of the system. Furthermore, power prediction improves the total effective throughput even more. Call admission control brings less system congestion, under which situation more packets in a flow will be delivered successfully to end users and more useful traffic is received. Thus, the effective throughput of the system is improved. When integrating admission control with power prediction, mobility prediction gives power estimate in a pre-specified near future, which reduces the chance of base station power shortage due to user mobility. This way, the amount of successfully delivered traffic increases and the useful received traffic increases, which in turn, brings even higher total effective throughput to the system. When performing the prediction longer in our simulation with 10% time prediction, the effective throughput is even higher than that with 5% time prediction.

## 4.6 Conclusions

In the first section of Chapter 3, we state that our thesis objective is to devise a set of admission control and scheduling mechanisms that aim at maximizing the effective throughput of the system, subject to the constraints on the total available base station power, and the requirement of achieving

the target proportional delays among the various DiffServ classes at moderate congestion times. In the following sections of Chapter 3, a set of scheduling and admission control mechanisms are devised including two proportional delay differentiation schedulers and two call admission control methods.

In this chapter, we aim at analyzing the performance of the above approaches using simulation, and investigating whether our thesis objective stated in Section 3.1 can be achieved by the scheduling and admission control mechanisms proposed. We summarize the performance results through Chapter 4 as follows:

- The WIN scheduler outperforms the HOQ scheduler in terms of lower average class delay and more accurate delay ratio achieved at moderate congestion times.
- Our traffic generation is intended to produce three different levels of system congestion for different CDMA downlink transmission rates. At the low rate part ranging from 16 Kbps to 64 Kbps, there is likely no congestion in the system. The average class delay and throughput for each class is almost the same because almost all the packets can be transmitted at the time of their arrival. There is little need for proportional delay differentiation scheduling.
- For the medium rates from 96 Kbps to 128 Kbps without call admission control, the system is moderately congested. Proportional delay differentiation scheduling achieves the target proportional delays among the various DiffServ classes, which is one objective of the thesis.
- For the high rate part from 144 Kbps to 192 Kbps without call admission control, the system is heavily congested and most packets miss their expiry time. Lower classes' delays are approaching the preset maximum delay limit, which affects the achieving of target proportional delays.

- Call admission control lowers the system congestion level and reduces the average class delay in each DiffServ class. Integrating admission control with power prediction brings even lower average class delay than the simple admission control proposed in Section 3.3.
- Admission control brings throughput fairness among classes. Without admission control, lower classes' traffic is starving when the system is heavily congested. However, through admission control, every class's traffic can have a share of the total throughput at any transmission rate. Integrating admission control with power prediction brings even more throughput fairness among classes.
- Admission control brings higher total effective throughput than that without admission control. The total effective throughput achieves the maximum value when integrating admission control with power prediction, which is the main objective of our thesis.

In conclusion, by integrating admission control with power prediction to provision proportional delay differentiation on the downlink of a cellular CDMA environment, we maximize the effective throughput of the system with fairness among the various DiffServ classes, and satisfy the requirement of achieving target proportional delays among classes at moderate congestion time with reduced average class delay.

# Chapter 5

## Conclusions and Future Work

### 5.1 Conclusions

Third Generation wireless systems are designed to provide mobile users with a variety of guaranteed services. Provisioning quality of service (QoS) connections to mobile users in W-CDMA environments then arises as a significant issue. DiffServ architecture provides a general framework for provisioning quality of service (QoS) over the Internet. In this work, we have considered provisioning proportional delay differentiation per-hop behaviour over the downlink of a cellular W-CDMA environment. Our objective has been to devise a set of call admission control and scheduling mechanisms that aim at maximizing the effective throughput of the system in a high mobility environment, subject to constraints on the total available base station power, and the requirement of achieving target proportional delays among the various DiffServ classes at moderate congestion times.

We consider serving delay-critical traffic over the downlink of a CDMA environment. The time duration of a single flow in our study is long enough for the associated mobile host to travel a significant distance relative to its initial position at the time the flow is admitted. We also identify the effective throughput concept based on the consideration that the user should only be charged on the amount of useful received traffic.

Two proportional delay differentiation schedulers are proposed for assessing the average delay incurred by the queued packets in each DiffServ class at any instant. The first scheduler, called the HOQ scheduler, computes the average delay of each class by averaging the time delays incurred by the head-of-queue packet in each queue of that class. The second scheduler, called the WIN scheduler, computes the average delay of each class by averaging the time delays incurred by a set of most-delayed packets in each class.

A call admission control scheme is devised, which makes admission decision based on current availability of base station power. If admitting the newly incoming flow does not cause power shortage at current time instant, the new flow is admitted. However, admitting a new flow based on current availability of base station power may lead to future power shortage due to mobiles' movement. Taking the user mobility and CDMA soft capacity into consideration, an admission control scheme integrated with power prediction is proposed. Mobility prediction is adopted to estimate the base station power required by all mobiles in a pre-specified future period.

Towards the thesis objective, we use simulation to analyze the performance of the above mechanisms in terms of the average delays incurred by the packets in each delay class, and the achieved effective throughput. The simulation results show that the WIN scheduler outperforms HOQ scheduler in terms of lower average class delay and more accurate delay ratio achieved. The target proportional delay requirement among the various DiffServ classes is satisfied at moderate congestion times. By integrating call admission control with power prediction, the average delay incurred by packets in each class is reduced and throughput fairness among the various DiffServ classes is achieved. Call admission control brings higher system effective throughput, and the total effective throughput of the system is maximized by power prediction.

## 5.2 Future Work

This section introduces a number of possible future research directions that complement the work done in this thesis.

Firstly, in order to enhance the obtained results, we may consider the following future research directions; these directions include considering a heterogeneous mix of traffic within the same Diffserv class, a heterogeneous mix of user mobility profiles (by varying the travelling speeds), and enhancing the methods used for mobility prediction sampling.

We also note that the call admission control devised in the thesis does not limit the amount of traffic admitted in each class. In our simulation realization, the traffic is assumed evenly distributed among the three delay classes. If the traffic of one class is greedy (i.e., exceeds the traffic of other classes), the greedy traffic will occupy all the system resources and block admission of other classes' traffic. In our future work, resource sharing among the various DiffServ classes during admission control should guarantee a minimum amount of resources for each class.

Lastly, we propose investigating the design of admission control and scheduling mechanisms for the uplink traffic (from mobile users to base stations). Uplink differs from downlink in a number of aspects including flow buffering, CDMA soft capacity, and power budget. In the uplink, traffic is generated and buffered in mobile devices. Accessing and estimating the average delay incurred by the queued packets in each class, and performing central scheduling as in the downlink, need more effort for provisioning proportional delay differentiation. A suitable signaling protocol is necessary for traffic information and power allocation command communication between mobile hosts and base stations. Moreover, in the uplink, each mobile has its own power budget, which is much less than the transmission power budget of the base station. When provisioning proportional delay differentiation in the uplink, this mobile's power budget may limit the central scheduling flexibility, which in turn may influence the quality of the achieved proportional delays.

Once the proportional delay differentiation is implemented in the uplink as in the downlink, we can further investigate dynamic class adaptation mechanisms over CDMA wireless environments for end-to-end flow delay guarantee.

# Bibliography

- [1] N. Anderson. *UMTS Terrestrial Radio Access (UTRA) ITU-R RTT candidate*. ETSI, September 1998.
- [2] I. Andrikopoulos, L. Wood, and G. Pavlou. A fair traffic conditioner for the assured service in a differentiated services internet. In *Proceedings of IEEE International Conference on Communication (ICC '2000)*, volume 2, pages 806–810, New Orleans, June 2000.
- [3] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss. An architecture for differentiated services. RFC 2475, IETF, December 1998.
- [4] S. Bodamer. A scheduling algorithm for relative delay differentiation. In *Proceedings of IEEE Conference on High Performance Switching and Routing (ATM 2000)*, pages 357–364, Heidelberg, June 2000.
- [5] H. Chen, H. Hassanein, and H. Mouftah. Providing packet loss guarantees in differentiated services architectures. In *Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering*, volume 1, Toronto, Ontario, May 2001.
- [6] C. Comaniciu, N. B. Mandayam, D. Famolari, and P. Agrawal. QoS guarantees for third generation (3G) CDMA systems via admission and flow control. In *Proceedings of IEEE Vehicle Technology Conference*, Boston, MA, September 2000.

- [7] S. Dennet. *The CDMA2000 ITU-R RTT Candidate Submission*. TIA, July 1998.
- [8] C. Dovrolis and P. Ramanathan. Proportional differentiated service, part II: Loss rate differentiation and packet dropping. In *Proceedings of the 2000 International Workshop on Quality of Service (IWQoS)*, pages 52–61, Pittsburgh PA, June 2000.
- [9] C. Dovrolis and P. Ramanathan. Dynamic class selection: from relative differentiation to absolute QoS. In *Proceedings of Ninth International Conference on Network Protocols (ICNP) 2001*, pages 120–128, Riverside, California, November 2001.
- [10] C. Dovrolis, D. Stiliadis, and P. Ramanathan. Proportional differentiated services: Delay differentiation and packet scheduling. In *Proceedings of ACM SIGCOMM '99*, pages 109–120, Boston, MA, September 1999.
- [11] S. Floyd and V. Jacobson. Random early detection gateways for congestion avoidance. *IEEE/ACM Transactions on Networking*, 1:397–413, August 1993.
- [12] M. Goyal, A. Durresi, R. Jain, and C. Liu. Performance analysis of assured forwarding. Internet draft, IETF, February 2000.
- [13] M. Goyal, A. Durresi, P. Misra, C. Liu, and R. Jain. Effect of number of drop precedences in assured forwarding. In *Proceedings of IEEE Global Telecommunications Conference (GlobeCom99)*, volume 1(A), pages 188–193, Rio de Janeiro, Brazil, December 1999.
- [14] O. Gurbuz and H. Owen. Dynamic resource scheduling schemes for W-CDMA systems. *IEEE Communications Magazine*, 38:80–84, October 2000.

- [15] A. Habib, S. Fahmy, and B. Bhargava. Design and evaluation of an adaptive traffic conditioner for differentiated services networks. In *Proceedings of International Conference on Computer Communications and Networks*, Arizona, October 2001.
- [16] J. Heinanen, F. Baker, W. Weiss, and J. Wroclawski. Assured forwarding PHB group. RFC 2597, IETF, June 1999.
- [17] H. Holma and A. Toskala. *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications*. John Wiley, 2000.
- [18] V. Jacobson, K. Nichols, and K. Poduri. An expedited forwarding PHB. RFC 2598, IETF, June 1999.
- [19] M. R. Jeong, K. Kakami, H. Morikawa, and T. Aoyama. Wireless scheduler providing relative delay differentiation. In *Proceedings of The Third International Symposium on Wireless Personal Multimedia Communications (WPMC'00)*, pages 1067–1072, Bangkok, Thailand, November 2000.
- [20] L. Jorguseski, E. Fledderus, J. Farserotu, and R. Prasad. Radio resource allocation in third-generation mobile communication systems. *IEEE Communications Magazine*, pages 117–123, February 2001.
- [21] N. Joshi, S. R. Kadaba, S. Patel, and G. S. Sundaram. Downlink scheduling in CDMA data networks. In *Proceedings of MobiCom 2000*, pages 179–190, Boston, MA, August 2000.
- [22] M. Kazmi, P. Godlewski, and C. Cordier. Admission control strategy and scheduling algorithms for downlink packet transmission in WCDMA. In *Proceedings of 52th IEEE Vehicular Technology Conference*, Boston, USA, September 2000.
- [23] A. Kumar, J. Kaur, and H. M. Vin. End-to-end proportional loss differentiation. Technical Report TR-01-33, University of Texas at Austin, February 2001.

- [24] T. H. Lee and J. T. Wang. Admission control for VSG-CDMA systems supporting integrated services. In *Proceedings of IEEE GLOBECOM*, pages 2050–2055, Sydney, Australia, November 1998.
- [25] M. Leung, J. Lui, and D. Yau. Characterization and performance evaluation for proportional delay differentiated services. In *Proceedings of International Conference on Network Protocols (ICNP)*, pages 295–304, Osaka, Japan, October 2000.
- [26] M. Leung, J. Lui, and D. Yau. Adaptive proportional delay differentiated services: Characterization and performance evaluation. *IEEE/ACM Transactions on Networking*, 9(6):801–817, December 2001.
- [27] T. Liu and J. Silverster. Joint admission/congestion control for wireless CDMA systems supporting integrated services. *IEEE selected areas in communications*, 16:845–857, August 1998.
- [28] R. Makkar and I. Lambadaris et al. Empirical study of buffer management scheme for DiffServ assured forwarding PHB. In *Proceedings of Ninth International Conference on Computer Communications and Networks*, Las Vegas, Nevada, October 2000.
- [29] F. Muratore. *UMTS: Mobile Communications for the Future*. John Wiley, 2001.
- [30] T. Nandagopal, N. Venkitaraman, R. Sivakumar, and V. Bharghavan. Relative delay differentiation and delay class adaptation in core-stateless networks. In *Proceedings of IEEE INFOCOM 2000*, pages 421–430, Tel Aviv, Israel, March 2000.
- [31] K. Nichols, S. Blake, F. Baker, and D. Black. Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers. RFC 2474, IETF, December 1998.
- [32] K. Nichols, V. Jacobson, and L. Zhang. A two-bit differentiated services architecture for the internet. RFC 2638, IETF, July 1999.

- [33] T. S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall, 1995.
- [34] G. Rogers, M. Minhazuddin, and R. Liu. Mixing UDP and TCP in a DiffServ assured forwarding PHB - a programmable networks scenario. In *Proceedings of IEEE International Conference on Networks (ICON)*, Bangkok, Thailand, October 2001.
- [35] S. Sahu, P. Nain, D. Towsley, C. Diot, and V. Firoiu. On achievable service differentiation with token bucket marking for TCP. In *Proceedings of ACM SIGMETRICS*, pages 23–33, Santa Clara, CA, June 2000.
- [36] O. Sallent, J. Perez, R. Agusti, and F. Casadevall. A scheduling algorithm for soft-QoS guarantee in 3G systems. In *Proceedings of the IEEE Vehicular Technology Conference 2001*, Greece, May 2001.
- [37] A. Sampath, P. S. Kumar, and J. M. Holtzman. Power control and resource management for a multimedia CDMA wireless system. In *Proceedings of IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pages 21–25, Toronto, Canada, September 1995.
- [38] D. Shen and C. Ji. Admission of multimedia traffic for third generation CDMA network. In *Proceedings of IEEE INFOCOM 2000*, volume 3, pages 1077–1086, Tel Aviv, Israel, March 2000.
- [39] S. Shenker, R. Braden, and D. Clark. Integrated services in the Internet architecture: An overview. RFC 1633, IETF, June 1994.
- [40] S. Singh, V. Krishnamurthy, and H. V. Poor. Integrated voice/data call admission control for wireless DS-CDMA systems. *IEEE Transactions on signal processing*, 50(6):1483–1495, June 2002.
- [41] I. Stoica. *Stateless Core: A Scalable Approach for Quality of Service in the Internet*. PhD thesis, Carnegie Mellon University, 2000.

- [42] I. Stoica and H. Zhang. Lira: A model for service differentiation in the Internet. In *Proceedings of NOSSDAV'98*, pages 115–128, London, UK, July 1998.
- [43] I. Stoica and H. Zhang. Providing guaranteed services without per flow management. In *Proceedings of ACM SIGCOMM*, pages 81–94, Boston, MA, September 1999.
- [44] W. Yang and E. Geraniotis. Admission policies for integrated voice and data traffic in CDMA packet radio networks. *IEEE selected areas in communications*, 12:654–664, May 1994.
- [45] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala. RSVP: A new resource ReSerVation protocol. *IEEE Network Magazine*, 7(5):8–18, September 1993.