DEMAND DRIVEN ADAPTIVE TRANSMISSION FOR WIRELESS SYSTEMS WITH HETEROGENEOUS TRAFFIC

by

Shyh-hao Kuo

B.E. (Hons. I), University of Canterbury, NZ

M.E., University of Canterbury, NZ

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
in the School
of
Engineering Science

SIMON FRASER UNIVERSITY
Fall 2008

APPROVAL

Name: Shyh-hao Kuo

Degree: Doctor of Philosophy

Title of thesis: Demand Driven Adaptive Transmission for Wireless Systems with Heterogeneous Traffic

Examining Committee: Dr. Mirza Faisal Beg
Chair

Dr. James K Cavers, *Senior Supervisor*

Dr. Daniel Lee, *Supervisor*

Dr. Paul Ho, *Supervisor*

Dr. Jie Liang, *Internal Examiner*

Dr. Abbas El Gamal, *External Examiner*
Professor of Electrical Engineering
Stanford University

Date Approved: 6 - oct - 2008

Abstract

With the increasing consumer demand for high-speed wireless access to the Internet, new technologies have been developed and standards have been proposed to improve the efficiency of wireless devices to provide the bandwidth and to support the quality of service (QoS) required. In this thesis, we investigate one aspect of the problem arose from providing Internet-like services with a wireless physical layer interface, specifically, energy optimal cross-layer designs of packet schedulers under strict, per-packet delay constraints. The design of the schedulers takes into consideration the randomness of the packet arrival as well as time-varying channels.

Firstly, a novel convex optimization formulation is proposed. By assuming prescient knowledge of the channel state information and packet arrival and expiry times, an interesting analytical solution is derived with a novel geometric interpretation, referred to as piecewise water-filling. An efficient algorithm for calculating such a solution is also presented. The problem is considered under both single- and multi-carrier scenarios with simulation results showing improvements due to the channel diversity effect.

In addition to the analysis of the *prescient* scheduler, practical issues are also considered. Specifically, several optimal causal schedulers are proposed, each assuming various degrees of prior knowledge of the system parameters. Through simulations of these causal schedulers, it was established that the optimal packet scheduler with-

out cross-layer knowledge uses 10 dB more energy than optimal *prescient* scheduler. In contrast, a practical causal scheduler with cross-layer knowledge of physical layer channel gains, utilizing an 8 tap Wiener channel prediction filter, can achieve energy usage that is only 3 dB away from the *prescient* scheduler.

For completeness, we also study the modifications required to the scheduling problem formulation for multi-user channels, specifically, under the information theoretical multi-access channel (MAC) and broadcast channel (BC) channel models. For practical systems, we propose a practical scheme for cross-layer design in a multi-user environment based on time division multiple access.

Keywords:

packet scheduling, convex optimization, cross-layer design, quality of service, adaptive modulation

Subject Terms:

Telecommunication – Traffic

Constrained optimization

To my beloved wife, Grace;

and my son, Jonathan.

Contents

Approval ii

Abstract iii

Dedication v

Contents vi

List of Figures ix

Abbreviations xii

List of Symbols xiv

1 Overview of the Thesis 1

1.1 Motivation 1

1.2 Organization of the Thesis 2

2 Background 5

2.1 Adaptive Transmission Technology 6

2.2 Cross Layer Designs 8

2.3 Convex Optimization 11

2.4 System Level Considerations 13

2.5 Contribution 17

3 Prescient Schedulers **19**

3.1 A Toy Problem 19

3.1.1 System Model 20

3.1.2 Convex Formulation 23

3.1.3 A Numerical Example 26

3.1.4 KKT Conditions and Their Interpretations 29

3.2 Diversity Channels 33

3.2.1 System Model 34

3.2.2 Convex Optimization Formulation 36

3.2.3 KKT Conditions and Their Interpretations 37

3.2.4 Numerical Result 40

4 Efficient Optimization Algorithms **42**

4.1 Shortest Path Property 43

4.1.1 String Pulling Algorithm 44

4.2 Information and Noise Rate 48

4.2.1 Bounding the Information and Noise Rate 50

4.2.2 Iteratively Finding the Active Tones 54

4.3 Simulation Result 55

5 Single User Online Scheduling Algorithm **58**

5.1 Optimal Online Scheduler 59

5.1.1 Optimal Casual Scheduler with Conditional Future Averages . . 61

5.1.2 Optimal Causal Scheduler with MMSE Channel Prediction . . . 66

5.1.3 The OSI Layered Scheduler 69

5.2 Qualitative Comparison of Online Schedulers 71

5.3 Performance Comparison 76
5.3.1 Normalized Traffic and Channel Model 77
5.3.2 Defining the Simulation Scenarios 77
5.3.3 Performance in Static Flat Channel 79
5.3.4 Performance in Fading Channels 80

6 Effect of Queueing Disciplines **85**
6.1 Non Preemptive Priority Queue 86
6.2 Preemptive EDF Queue 90
6.3 Conclusion 94

7 Multi-user Schedulers **96**
7.1 Background on User Admission Control 98
7.2 Prescient Problem Formulation 100
7.2.1 Multi-access Channel 100
7.2.2 Broadcast Channel 103
7.3 Multi-User Piecewise Waterfilling 104
7.4 Practical Considerations 106
7.5 Conclusion 107

8 Conclusions and Future Directions **108**
8.1 Summary 108
8.2 Future Directions 109

A Channel Power Gain Prediction **111**

B The Power Formula for the Multi-access Channel **114**

List of Figures

2.1 A simplified model of a trunked communication system. 16

3.1 The system block diagram of the optimal cross-layer scheduler. 21

3.2 Diagram showing cumulative arrivals and expiries as the bound on the admissible cumulative rate profile. The solid dot (•) at the epoch boundaries shows the actual value taken for each of the parameters in discrete notation. In this diagram, the height of eack box corresponds to the size of the packet. 24

3.3 The optimal cumulative rate profile for the numerical example (a) and the corresponding water-filling diagram showing the piecewise water-filling property in (b). The optimal transmit power is the difference of the two curves shown in (b), $f(r,g) = (e^r - 1)/g$. 28

3.4 The optimal cumulative rate profile for a two-carrier system and the respective water-filling in rate diagram. 41

4.1 An illustration of the string-pulling process. 45

4.2 Diagram showing the relationship between the cumulative arrivals and expiries (a) and the water-filling diagram (c). The transformation required to show the shortest path property is shown in (b). Idle epochs are highlighted by the vertical strips. The cumulative information and noise rate (cINR) is piecewise linear outside of the idle epochs (shaded region). 49
4.3 Graphs showing the cumulative arrivals and noise rate (cANR), cumulative expiries and noise rate (cENR) and cINR curves in (a). The shortest path property can be best visualized by piecing together consecutive segments of active tones as shown in (b). 53
4.4 The water-filling diagram and the cINR plot for the first iteration of the iterative string pulling (ISP) algorithm. 56
4.5 The water-filling diagram and the cINR plot for the second (final) iteration of the ISP algorithm. 57

5.1 Diagram showing the traffic constraints known to the causal schedulers at the time instant as indicated by the vertical cursor. 62
5.2 Diagram showing the definition of the actual and causal expiry and arrival curves with reference to the current time indicated by the vertical cursor. 63
5.3 The system diagram for the layered scheduler. Note that the channel state information is not used in determining the service rate. 69
5.4 The shortest path under only cumulative expiry bounds must form the convex hull over the constraint point as shown in the diagram. 70
5.5 The cumulative rate plot and the rate water-filling diagram of the layered scheduler. 72

5.6 The cumulative rate plot and the rate water-filling diagram for the one tap prediction scheduler. 74
5.7 The cumulative rate plot and the rate water-filling diagram for the 8 tap prediction scheduler. 75
5.8 Comparison of the average signal to noise ratio (SNR) per bit ($\bar{\gamma}_b$) for the causal and prescient schedulers under static channel conditions. . . 80
5.9 Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 1$ subcarriers at spectral efficiency $\rho = 1$ bits per second per Hz. 81
5.10 Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 2$ subcarriers at spectral efficiency $\rho = 1$ bits per second per Hz. 81
5.11 Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 8$ subcarriers at spectral efficiency $\rho = 1$ bits per second per Hz. 82
5.12 Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 1$ subcarriers at spectral efficiency $\rho = 2$ bits per second per Hz. 83
5.13 Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 1$ subcarriers at spectral efficiency $\rho = 2$ bits per second per Hz. 84

6.1 A close up view of the cumulative arrivals and expiries showing the relaxation of the expiry constraint when queue reordering takes place. Each of the queue diagram shows the state of the system assuming that the actual cumulative rate curve intersects the vertical axis at the point range shown. 88
6.2 Diagram showing optimal power allocation for a preemptive priority queueing scheme. 92

A.1 Plot of the power gain of a Rayleigh fading channel and various minimum mean square error (MMSE) predictor outputs for 1, 2, 8, and 16 taps. . 112

Abbreviations

AWGN additive white Gaussian noise

BC broadcast channel

BER bit error rate

CLD cross layer design

CLO cross layer optimization

CSI channel state information

cANR cumulative arrivals and noise rate

cENR cumulative expiries and noise rate

cINR cumulative information and noise rate

EDF earliest deadline first

FIFO first in first out

INR information and noise rate

ISI inter symbol interference

ISP iterative string pulling

KKT Karush-Kuhn-Tucker

LHS left hand side

MAC multi-access channel

MIMO multi-input multi-output

MMSE minimum mean square error

OFDM orthogonal frequency division multiplexing

OSI open system interconnection

PWF piecewise water-filling

QoS quality of service

mQAM m-ary quadrature amplitude modulation

RHS right hand side

SISO single input single output

SNR signal to noise ratio

TDM time devision multiplexing

TDMA time division multiple access

WiMAX worldwide interoperability for microwave access

List of Symbols

chapter 3

i, j	general discrete dummy variables for indexing purposes.
$T_a(i)$	arrival time of packet i.
$T_e(i)$	expiry time of packet i.
$\tau(i)$	maximum delay requirement of packet i.
$D(i)$	length of packet i.
t	continuous time.
m	discrete time index (epoch or symbol).
n	subcarrier index (for multi-carrier system only).
N	total number of epochs or symbol.
M	total number of subcarriers.
$g(t), g(m), g_n(m)$	channel power gain at time t, or for epoch m, in subcarrier n.
$r(t), r(m), r_n(m)$	transmission rate profile at time t, or for epoch m, in subcarrier n.
$R(m)$	transmission rate profile for epoch m summed over all subcarriers.
$R_{\text{cum}}(t)$	cumulative transmission rate.
$A_{\text{cum}}(t)$	cumulative arrivals.
$E_{\text{cum}}(t)$	cumulative expiries.
W	frequency bandwidth of the channel.
$x(m)$	normalized bandwidth of epoch m.
$f(r, g)$	required transmit SNR as a function of transmission rate r and channel power gain g.
$p(r, g)$	power required to transmit at rate r over channel gain g.

$\lambda_m, \lambda_n(m)$	Lagrangian multipliers for the rate positivity constraint.
$\mu_m, \mu(m)$	Lagrangian multipliers for the cumulative arrival constraints.
$\nu_m, \nu(m)$	Lagrangian multipliers for the cumulative expirie constraints.
$\omega, \omega(m)$	power waterlevel of the water-filling solution.
K	SNR modifier constant for adaptive modulation and coding.
ω	power water level.
N_0	Noise power.

chapter 4

$\eta_n(m)$	noise rate for epoch m in subcarrier n.
$\zeta_n(m)$	Information and noise rate for epoch m in subcarrier n.
Ω_m	Set of active channels for epoch m.
Ψ_m	An arbitrary set of subcarriers.
$Z^{\Psi_m}(m)$	sum of $\zeta_n(m)$ over all $n \in \Psi_m$.
$Z^{\mathbf{\Psi}}_{\text{cum}}(m)$	A cumulative information and noise rate (calculated over some estimate of the active set).
$Z^{\mathbf{\Omega}}_{\text{cum}}(m)$	The optimal cumulative information and noise rate (calculated over the active set).
$\square^{[i]}$	interime values of the quantity $\square$ for iteration i of the ISP algorithm.

chapter 5

k	the optimization horizon of the online schedulers measured in either epochs or symbols.
κ	symbol or epoch index of the online schedulers.
$\tilde{E}^{[m]}_{\text{cum}}(\kappa)$	partially known cumulative expiry bounds at epoch/symbol m.
$E^{[m]}_{\text{cum}}(\kappa)$	a random vector denoting cumulative expiries of packets not yet arrived at epoch/symbol m.
$A^{[m]}_{\text{cum}}(\kappa)$	a random vector denoting cumulative arrivals of future packets at epoch/symbol m.
$Q^{[m]}$	queue length (measured in *nats*) at epoch/symbol m.
$\mathbf{g}^{[m]}_{\text{past}}$	a vector of past and current channel power gains up to and

	include epoch/symbol m.
$\boldsymbol{g}_{\text{future}}^{[m]}$	a vector of future channel power gains.
$\hat{\boldsymbol{g}}_{\text{future}}^{[m]}(\kappa)$	estimates of future channel power gains.
$T^{[m]}$	future traffic states at symbol/epoch m.
$h_n(m)$	complex channel gain at symbol/epoch m.
δ	sampling spacing of the channel prediction filter measured in symbols.
t_{IA}	a random variable denoting the packet interarrival times.
θ	a random variable denoting the packet length.
ρ	the spectral efficiency measured in bits/sec/Hz
t_e	a random variable denoting the maximum delay constraint of the packets.
f_d	maximum Doppler frequency.
$\bar{\gamma}_b$	average SNR per bit.

chapter 7

N	redefined in this chapter as the number of users in a multi-user system.
n	user index.
$\boldsymbol{g}(m)$	length N vector of channel power gains for epoch/symbol m.
$\boldsymbol{A}_{\text{cum}}(m)$	length N vector of cumulative arrivals for epoch/symbol m.
$\boldsymbol{E}_{\text{cum}}(m)$	length N vector of cumulative expiries for epoch/symbol m.
$\boldsymbol{r}(m)$	length N vector of transmission rates for epoch/symbol m.
$H_{\text{abc}}(m)$	an indicator function identifying the order of channel gains for user a,b, and c.

Chapter 1

Overview of the Thesis

1.1 Motivation

Guaranteed quality of service (QoS) is a difficult but necessary requirement in modern wireless communication systems. Many real-time media services, such as realtime voice or video, require that the packetized data be properly received and decoded before a specified deadline. This deadline is the time instant after which the data packet is no longer useful to the receiver. In this thesis, we consider the issue of providing guaranteed packet delivery within the specified deadline for each packet over some commonly encountered wireless channels.

Conventionally, multimedia data are only transported through wire-line channels where bandwidth can be expanded by adding an extra pair of wires and where power usage is not a limiting factor. However, in recent years, we see a proliferation of mobile communication devices with multi-media capabilities and there is an increasing user demand for wireless systems with real-time streaming capability. With the additional constraints of limited bandwidth and the time-varying nature of the mobile channel, the inefficiency inherent in the layered open system interconnection (OSI) reference

model of networking becomes more apparent. In recent years, many researchers have studied adaptive transmission that jointly considers both the queueing and physical aspects of a wireless system in an attempt to optimize the performance of such wireless systems. A comprehensive survey of pioneering research in this area can be found in [1].

This thesis investigates the design of an adaptive modulation and power control system that is capable of meeting the individual packet transmission deadlines with the minimum amount of energy usage. The emphasis is on minimizing the energy usage under the influence of both a time-varying traffic load and time-varying channels. A variety of channel models is considered in this thesis, specifically, we consider flat fading, multi-carrier, as well as multi-access channel (MAC) and broadcast channel (BC).

1.2 Organization of the Thesis

This thesis is organized around the following main contributions of my Ph.D. research. These are:

1. the derivation of the optimal prescient schedule and the *piecewise water-filling* (PWF) property of the optimal rate profile.
2. the development of the iterative string pulling (ISP) algorithm as an efficient method for finding the prescient optimal schedule.
3. the development of the channel predicting *causal schedulers* as practical schedulers to be used in real world wireless communication systems.
4. the analysis of the optimal prescient scheduler for a variety of queueing disciplines.
5. the extension of the scheduling algorithms to the multi-user channels.

The problem investigated and the solution obtained is applicable to a wide variety of channel and traffic models. In the most general case, the solution presented here can be applied to multi-user, multi-carrier channels with a variety of queueing disciplines. In this thesis, instead of presenting the most general system model and the associated solution, we present firstly the full details of the analysis of a simpler single user, multi-carrier scheduler with a first in first out (FIFO) queueing discipline. This is presented in Chapter 3 to 5 and forms the main part of this thesis. The variation in queueing disciplines and the consideration of multi-user channels are discussed in Chapter 6 and Chapter 7 respectively as modifications to the single user, multi-carrier system. Numerical results are used throughout the thesis as specific examples to demonstrate the important properties of the analytical solutions. It is not intended to simulate the performance of any particular real world application nor does it provide a full coverage of all combinations of system parameters.

The rest of this thesis is organized as follows. After a brief background and literature survey in Chapter 2, we present a deterministic problem formulation in Chapter 3 for a point-to-point system, i.e., a single user channel. This deterministic view of the system assumes all future traffic and channel states are known *a priori* and it is from this deterministic view that we derived the optimal solution. To the best of the author's knowledge, this is the first time such a solution was observed. The treatment of the deterministic problem begins with a small toy formulation using a single user, point-to-point flat fading link, followed by the generalization of the *piecewise water-filling property* to multi-carrier channels. From this *piecewise water-filling* observation, an efficient numerical algorithm for computing the optimal solution was designed and is presented in Chapter 4. This algorithm exploits the piecewise water-filling property and has computational complexity much lower than that of using generic convex optimization routines.

After the treatment of the prescient scheduling problem, the causal formulation is considered (Chapter 5). Several optimal causal schedulers are formulated that are different in the causality assumptions. Specifically, we consider the cases where only limited knowledge of the future channel gains and packet arrivals are available through prediction. An arbitrarily chosen set of traffic scenarios were used to simulate the performance of these causal schedulers and the resulting energy usage is compared against that of the prescient solution. A brief discussion on the relative merits of these causal schedulers is also provided.

Chapter 6 and Chapter 7 present extensions to the single user scheduling problem. Firstly, a qualitative discussion of the effect of utilizing earliest deadline first (EDF) priority queues is given. Both preemptive and non-preemptive queues are considered and the differences in performance are discussed. While it is obvious that a preemptive EDF queue can achieve the lowest energy usage and the FIFO queue has the worst performance, the performance gap is not quantified since it is highly dependent on the traffic model.

Finally, the problem formulations for multi-user channels are presented in Chapter 7. In this chapter, we show how the *piecewise water-filling solution* is to be applied to the multi-user channels. Only a theoretical treatment would be given. Testing and comparison of the various proposed algorithm under multi-user channels are planned as future work.

Chapter 2

Background

In designing a mobile wireless network, we are faced with different challenges than designing for the traditional media for the internet, the wire-line channels. Mobile wireless channels are often characterized by the rapid constructive and destructive interference of the received signal (Rayleigh fading) as well as a slower variation in signal strength due to blockage of the signal path (shadowing). In addition, mobile devices are often equipped with a limited energy source. In order to use this limited energy efficiently, various techniques are developed to compensate for channel variations, ranging from simply allocating enough power margin to ensure there is always enough received signal power, to a more sophisticated method of dynamically allocating communication resources such as transmission power or information rate based on knowledge of the channel's state. Some of these adaptive resource allocation technologies are being adapted by the third-generation cellular standards (see e.g. [2]) and will feature even more strongly in future broadband wireless access technologies such as the mobile worldwide interoperability for microwave access (WiMAX).

During the time when the main application of mobile communication devices is real-time voice communications, channel fading was compensated by keeping a con-

stant received signal to noise ratio (SNR) by adjusting the transmit power to be inversely proportional to the channel gain, a technique known as channel inversion. While this provides guaranteed performance, it is more energy efficient if the data can be transmitted in high rate bursts during good channel conditions while staying idle in periods of bad channel condition [3].

To provide the necessary background on designing the optimal packet scheduler for heterogeneous traffics, we start with an historic overview of the development of adaptive transmission technology that only concerns itself with physical layer parameters for power control purposes in Section 2.1. In Section 2.2, a brief literature survey is provided into recent research into joint consideration of communication delay and information theoretic power control, which is the main area of research of this thesis. We then digress into introducing the mathematical terminologies used in formulating the problem in Section 2.3. Finally, we conclude this chapter with a description of the wireless communication system model in Section 2.4 and a summary of contribution in Section 2.5.

2.1 Adaptive Transmission Technology

The idea of adapting transmission parameters dynamically to the changing channel state can be dated back to the pioneering work in the late 60's and early 70's by [4, 5]. In these early works, a single end to end link with time varying channel is considered and the aim is to produce a time-varying modulation scheme (i.e. varying rate [5] or varying power [4]) that maintains a certain link quality measure (e.g. bit error rate (BER)). However, dynamic adaptation of transmission parameters requires feedback of the channel states and large computational power that were not available at the time and work in the area has not been taken seriously by the industry until more recently.

Most of the modern adaptive resource allocation schemes are based on the work by [3] that shows the water-filling power allocation scheme is optimal in achieving capacity under average power constraint. It, [3], also compares the water-filling scheme with other suboptimal schemes that strives to maintain a constant received SNR at a constant rate and show that there is a large power penalty. While [3] provides the informational theoretical framework, [6, 7] shows how variable information rate can be achieved through a family of coded m-ary quadrature amplitude modulation (mQAM) constellations and that at any specified BER target, the rate and power relationship follows closely to the logarithmic form of the capacity equation [8, 9] with a constant SNR penalty. A similar work using turbo coding to approach capacity can be found in [10]. A unified treatment of adapting various physical layer parameters, namely power, BER, and modulation can be found in [11], which provides a solid framework for designing optimal power control and adaptive modulation scheme for the physical layer. Variation on the water-filling adaptive scheme by [3] have been proposed. In [12], a peak power constraint is imposed on the system and it has been shown that with reasonable selection of the peak power, the performance penalty is minimal. A more severely limited transmitter with a constant on/off power control has been considered in [13] with a very-low complexity logarithm-free power allocation algorithm.

While all of the previously presented references assume perfect channel state information (CSI) at the transmitter, in practical systems, this CSI has to be estimated at the receiver and communicated back to the transmitter. Thus, there is an estimation error and delay associated with the actual CSI being used. This effect is considered in [14] where the effect of estimation error on the optimal algorithm is analyzed. In other work, [15] has proposed the use of channel prediction filters to provide a more up to date channel estimate at the transmitter by extrapolating the outdated channel estimate. A more sophisticated approach where uncertainty of channel esti-

mation is considered and compensated for in the problem formulation can be found in [16, 17]. In these works, a *strongly robust signal* is introduced that performs well under a selected set of channel autocorrelation functions. The idea of robustness and the use of a channel prediction filter will be used in this thesis to formulate the online algorithm where future channel gains and traffic are unknown.

2.2 Cross Layer Designs

The above-mentioned researches into adaptive modulation and coding within the physical layers, while complete and can be shown to be optimal, fail to consider the delay aspect of the data transmission. For example, if one were to use the water-filling power and rate control scheme in a shadowing situation, a user would experience no communication for the period in the shadow and a burst of activity when the channel became good again. This type of bursty transmission is energy efficient and can be used if the application is delay insensitive such as e-mail or large file transfer. However, it would not be acceptable for surfing the web nor real-time voice communications which has been the traditional market for wireless services. On the other hand, a channel inversion or constant receive SNR power control scheme offers guaranteed signal strength everywhere for voice communications but is extremely inefficient for delay insensitive traffics. Traditionally, the open system interconnection (OSI) reference model of communication systems separates the consideration of delay and modulation into two independent layers to be designed separately so that the physical layer power control is not traffic type aware and cannot adapt its power control strategy to the traffic load.

Cross layer design (CLD) is a name used to describe communication system designs that do not conform to the OSI reference model and allow different layers to communicate and optimize the overall performance. There are many different system parameters to be considered and we will only consider single hop delay in this

thesis. An overview of recent development on CLD in general can be found in [18]. The first challenge is to formulate the relationship between delay and power. While the information theoretical approach to power control yields a closed form solution, the codeword length and the associated delay are difficult to control. Several related capacity measures such as outage capacity [19] and delay limited capacity [20] have been proposed to provide a more meaningful measure of performance under delay constraints. Work in joint consideration of information theoretical capacity and delay are characterized by [21–23] and many more are mentioned in the survey by [1]. However, most of this work considers optimization for average delay only and does not cater for different delays for multiple classes of traffic nor does it guarantee a finite maximum delay.

For this thesis, the system considered is one that guarantees a maximum delay on a per-packet basis while minimizing transmission energy for time-varying channels. To the best of the author's knowledge, this is the first attempt at characterizing optimal transmission with per-packet, hard deadline constraints with a realistic continuous time-varying channel. Under this formulation, a communication link supporting multiple classes of traffic with different delay constraints can be modelled. Furthermore, the solution obtained is optimal in the sense that there exists no rate and power allocation scheme with lower energy usage under the same set of traffic constraints and channel conditions.

The most important aspect of this work is that per-packet maximum delay constraints are considered. It was found through the author's personal experience with engineering design, that the users often specify performance parameters in absolute terms, e.g., maximum delay shall not exceed 150 ms, while the system designers are more inclined to work with probabilistic measures, e.g., the average delay shall be 100 ms. This is perhaps due to the fact that most analysis in queueing theory only

provides results regarding average quantities. Thus, a system is often designed using these average design guidelines in an attempt to met some maximum or minimum specifications. For example, by designing for an average delay of 100 ms with some standard deviation around 10 ms, one could comfortably meet the 150 ms maximum delay constraint in most cases. However, this slight mismatch often results in sub-optimal designs and it is the goal of this research (perhaps due to some personal preference) to design a packet scheduler that meets this maximum delay constraints directly.

There exists only a limited number of previous works that provide solutions with a maximum delay instead of average ones, and even fewer where the delay constraints are explicitly controlled. Previous work that considers hard delay limit can be found in [24–31]. These works consist mainly of two Ph.D. theses and a paper which are described in more details below.

The earliest work is done by E. Uysal-Biyikoglu for her Ph.D. research [24–27] which considers a random arrival problem with a common deadline motivated from satellite communication systems. While it guarantees a maximum delay, it is considered as a system wide parameter and is not constrained nor minimized explicitly in the formulation. The problem formulation can be seen as a special case of the problem considered in this thesis, which allows individual packet deadline constraints. In [24–27] a problem-specific optimization algorithm, known as MoveRight, is proposed. It is similar in purpose to the iterative string pulling (ISP) algorithm developed in this thesis, but it can not be applied directly to the more general problem considered here. In the initial study performed for this thesis, it was observed that the MoveRight algorithm can be modified to find the solution for our problem, by "moving left and right" iteratively. However, it was found that the convergence of this "moving left and right" algorithm to the optimal solution was very slow, perhaps due to the additional

delay constraints. A more direct and general approach is developed in this thesis that exploits the piecewise waterfilling property observed in the analysis of the problem. The resulting ISP algorithm can also be applied to solve the common deadline problem considered in [24–27] and in the author's opinion, represents a more efficient algorithm for solving the optimisation problem.

Another related work is the Ph.D. research by M. Zafer [28–30] which considers a system setup similar to the one considered in this thesis with random arrivlas and expiries, however, the emphasis is quite different. It pursues a discrete Markov modelling approach with a 2 state Gilbert-Elliot channel model which, in the author's opinion, oversimplifies the problem and maskes the interesting property of piecewise water-filling over a continuous time-varying channel.

Finally, for completeness, there is also the work by [31] that presents a novel formulation of the problem as an adaptive filtering problem which can be solved readily with adaptive filtering techniques. However, the filter model can only be applied to the scheduling problem under a static channel.

2.3 Convex Optimization

Despite the variation in the problem formulation and sometimes the conclusions with cross layer optimization (CLO) researches, a common theme is to dynamically adapt certain physical layer resource allocations (rate, power, modulation, etc) in order to extremize some performance measure (error rate, power usage, capacity etc). The formulation of the optimization problem and the associated solution has been well studied and documented both in engineering [9] and as a mathematics discipline under numerical optimization and operations research [32, 33]. In this section, we provide the necessary background for the mathematical tools required to study optimization

problems that falls in the broad category of convex optimization. It does not aim to provide a comprehensive treatment of convex optimization but to provide a quick description of the terminologies and certain useful key results that are used in this thesis. A survey of the use of convex optimization methods for communications and signal processing research can be found in [34].

In general, an optimization problem is stated in the form of minimizing or maximizing some mathematical expression, known as the objective, over an admissible set of parameters. For the problems that we are interested in, the admissible set of parameters are defined as a continuous region in a multi-dimensional space often expressed as a set of inequality expressions. This region is also known as the domain of the problem. The goal is to find the point in this domain with the smallest or largest objective.

Convex optimization problem is a special class of optimization problem such that the domain forms a convex set and the objective is a convex function. The procise mathematical property of convexity can be found in [32]. It suffices to know for now that a convex optimization problem can be solved efficiently with numerical methods and that it can be transformed into another convex optimization problem, known as the dual, from which an analytical solution may be obtained. Furthermore, a set of necessary and sufficient conditions, known as the Karush-Kuhn-Tucker (KKT) conditions, can be stated and solved to aid in the understanding of the general property of the optimal solution. A good example of the use of the KKT condition is the derivation of the water-filling property of optimal power allocation and can be found in [9]. The analysis through KKT conditions will be used extensively in Chapter 3.

2.4 System Level Considerations

In this section, we take a more detailed look at the optimization problem. As in most good engineering practise, a good problem definition is as good as or better than a solution to an ill defined problem. The purpose of this section is to put the problem into the proper perspective. Many comments stated in this section may seem to be common sense at first. However, these common sense observations will be frequently referred to as the problem statement is being developed. They are critical, not just for the initial problem statement, but also at a later stage when it may be necessary to simplify the problem or to make certain modeling assumptions.

For a wireless system to support the internet-like services as we know it today, the system under consideration must: 1. support multiple users, 2. be wireless, and 3. support heterogeneous traffic types, as in most modern cellular systems. Furthermore, a wireless system that is capable of supporting multiple users inevitably requires large total bandwidth. This leads to a system model with multiple parallel channels accessible to all users such as an orthogonal frequency division multiplexing (OFDM) or time division multiple access (TDMA) system. We will also assume that a centralized controller is available to perform the scheduling and that all CSI is known at this central location. Under these system level specifications, one can now consider the problem under the proper perspective.

Most importantly, the system must perform to meet the user demand. This is to say, the system should not schedule access to users based on signal strength as suggested by the capacity maximization formulation in [35] but schedule to provide the best service to all users according to the user demand. An extreme example would be the implementation of the 911 emergency calls in cell phones. In such situation, the signal is often weak with low SNR but here, it has an associated user requirement that it should have the highest priority over all other traffic types and should be

given access to the channel regardless of its signal strength. While it is possible to implement a scheduler to provide for this single exception when such signal occurs by interruption or reserving emergency bandwidth, it is often better to design a scheduler which is inherently quality of service (QoS) aware so that multiple level of priorities can be implemented.

QoS specifications forms an important part of any modern wireless communication standards. For example, the emerging mobile WiMAX standard specifies the following different traffic classes

UGS (Unsolicited Grant Service) real-time data streams comprising fixed-size data packets issued at periodic intervals.

ertPS (Extended Real-time Polling Service) Real-time service flows that generate variable size data packets at periodic intervals.

rtPS (Real-time Polling Service) real-time Data streams comprising variable size data packets that are issued at periodic intervals.

nrtPS (Non-real-time Polling Service) delay tolerant data streams comprising variable size packets for which minimum data rate is required.

BE (Best Effort) Data Stream for which no minimum service level is required and is handled on a space available basis.

From this specified set of QoS classes, it is interesting to note that packets are always issued at periodic intervals and may be at variable packet length for real-time traffics with no service supporting variable interarrival intervals. Furthermore, two real-time services, ertPS and rtPS are specified with very similar description. In fact, when one examines the standard more closely, it is found that the main difference is that ertPS has an additional jitter constraint while rtPS does not. This demonstrates a pragmatic engineering design choice common to most wireless communication system where a synchronous periodic transmission frame is used in the physical layer and

multiple variable rate data services are implemented in higher layers with variable length packets over these periodic transmission intervals. While a system conforming to the WiMAX standard must implement the five different QoS classes as specified, we are under no such constraint and will be working with a generalized traffic model. Specifically, the system model allows variable length packets arriving at variable intervals and it is possible for each packet to have a different constraint on allowable delay. Under this general model, a system serving a mixed class of traffic simultaneously can be modelled and we will not rely on any assumption on the regularity of the traffic for analysis purposes.

Consider a generic wireless communication system with n users communicating through a common trunk, consisting of two base stations with a broadband wireless link in between, to another n users, as shown in Figure 2.1. It is assumed that all traffic is packetized into variable length packets presented to the transmitter at irregular intervals. Furthermore, each packet has an associated deadline parameter to specify the time by which this packet must be transmitted to the other end of the wireless link. We are interested in the overall *packet schedule* defined as the transmission rate and power of each packet that is most energy efficient while meeting all deadline constraints. By allowing packets of variable length with random arrivals, a wide range of QoS requirements can be specified in this scheme. For example, the nrtPS QoS class for mobile WiMAX can be modelled as either periodic arrival of fixed length packets with an appropriate over-all rate for a single user, or as random arrival of variable length packets with delay constraint proportional to the packet length as depicted by the minimum rate.

Since we are interested in the power and traffic load aspect of the system, we will simply ignore the possibility of multistage routing in the trunk. There are three distinct scenarios to consider. The uplink stage, the trunk stage and the downlink stage.

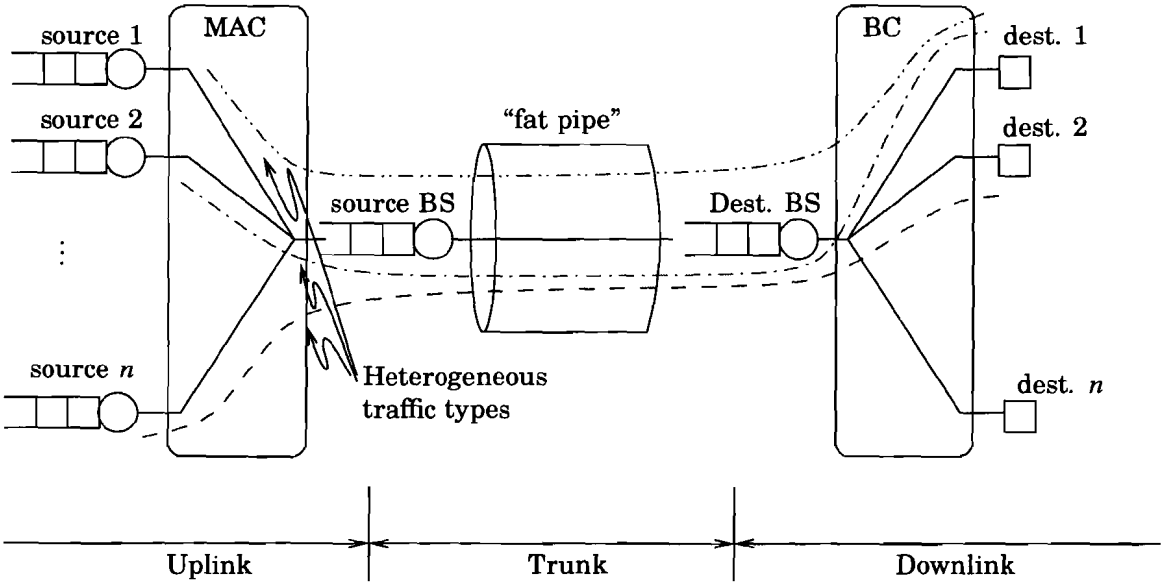

Figure 2.1: A simplified model of a trunked communication system.

During the trunk stage, all user traffic and system control traffic are firstly collected at one of the base station and transmitted through a broadband, (possible multi-carrier) wireless link in a managed manner. This is the simplest stage to consider as all traffic types and their associated QoS requirements are known at one location and can be scheduled optimally. Similarly, at the downlink stage, QoS requirements are known at the transmitter and the only different from scheduling in the trunk is that now there is a different set of channels for each user and one must also consider the effect of the broadcast channel (BC) model [36]. The difficulty for the design of the uplink scheduler lies in the fact that there is no central location for co-ordinating the transmission, not in the channel modeling. It is sufficient to say that if one wishes to implement uplink scheduling, a handshake protocol must be performed so that the uplinking base station has the channel state and traffic load information for scheduling purposes. One possible implementable protocol for uplinking is for all users to send a periodic request for resource allocation stating the packet length and deadline

requirements to the uplink base station. This base station will use the information provided in this request together with the channel estimates for all of the users to schedule the optimal time for transmission and inform each user of the schedule on the corresponding downlink. However, one must also account for the delay involved in this handshake.

2.5 Contribution

We consider the energy-optimal scheduling problem in two different settings. Firstly, under a deterministic setting where the problem is to find the optimal *schedule* for a given packet arrival and deadline constraints under a known time-varying channel. Under this deterministic setting, a novel *piecewise water-filling* solution is obtained with an efficient numerical method for computing this optimal schedule. The analysis of this deterministic system serves two purposes. Firstly, it aids in the understanding of the interaction of various traffic and channel parameters and secondly, the performance of the optimal solution serves as a bound on what is achievable for any real design. In addition to the deterministic formulation and the optimal prescient schedule, which we present in Chapter 3 and 4, we also propose an efficient and easy to implement online scheduler with performance close to that of the prescient optimal (Chapter 5). We also consider a general traffic and channel model where very few assumptions are required for analysis. Specifically, the traffic is modelled as discrete packets with individual deadline constraints and variable arrival time and packet length. We also consider a variety of time-varying channels ranging from single carrier flat fading channel to multi-user channels at various stages of a trunked communication system. Also, no specific statistical properties about the traffic or channel states are assumed in the deterministic formulation and while the online scheduler operates on the assumptions that the channel state is predictable, hence correlated

in time, it is not limited to be of any specific distribution.

As indicated earlier, the system under consideration must deal with not just a point-to-point multi-carrier channel but also the broadcast channel (BC) and multi-access channel (MAC) cases. Since the main difference between these is the channel models, we will delay the consideration for the multi-user channels till chapter 6 by concentrating on the point-to-point scenarios initially.

Chapter 3

Prescient Schedulers

In this chapter, we present the deterministic formulation of the scheduling problem for the single user channel. The deterministic view assumes that the present and future channel state information (CSI) and the offered traffic load are all known *a priori*. Despite being impossible to realize, the study of this prescient scheduler allows one to provide a performance bound on any other practical realization. That is, no other implementation can assume more knowledge than the prescient scheduler and hence can never outperform it.

For the rest of this chapter, we present the mathematical formulation of the problem and derive the *piecewise water-filling* property of the optimal prescient power profile which is essential for the derivation of efficient numerical methods presented in the next chapter.

3.1 A Toy Problem

We begin with a small toy problem that demonstrates the essential properties of the optimal packet schedule. Specifically, we consider a single user flat fading additive

white Gaussian noise (AWGN) channel with a first in first out (FIFO) input queue driving the input and the aim is to find the service rate profile that has the minimum energy usage. The problem definition and the associated solutions have also been presented by the author in [37]. To introduce the main concept, the *piecewise water-filling property* will be introduced through a numerical demonstration as a conjecture followed by a more rigorous proof of the observation. This *piecewise water-filling property* holds for the more general case with diversity channels, as well as multi-user systems, and the toy problem can be seen as a special case of these.

For the rest of this section, we start by introducing the cumulative arrivals and cumulative expiries as a way of determining the feasibility of any given transmission rate profile, from which the energy usage can be determined. A general representation of the channel as a time-varying function with little assumption of the channel model is then presented with its relation to the transmission rate and energy usage. With the traffic and channel properly specified, the scheduling problem is stated as a convex optimization problem from which efficient numerical methods can be used to find the solution [32] and a numerical example is used to demonstrate the *piecewise water-filling property*. Finally, we conclude this section with an analysis of the Karush-Kuhn-Tucker (KKT) conditions to derive the *piecewise water-filling property*.

3.1.1 System Model

The system model consists of a single flat fading wireless link with a FIFO queue at the input as shown in Figure 3.1. Each packet enters the queue at a random time and carries with it a quality of service (QoS) requirement stating its maximum allowable delay. The problem is to determine the optimal rate profile, the one that has the minimum transmission energy, assuming perfect knowledge of all past and future channel state information, packet arrivals and packet expiries.

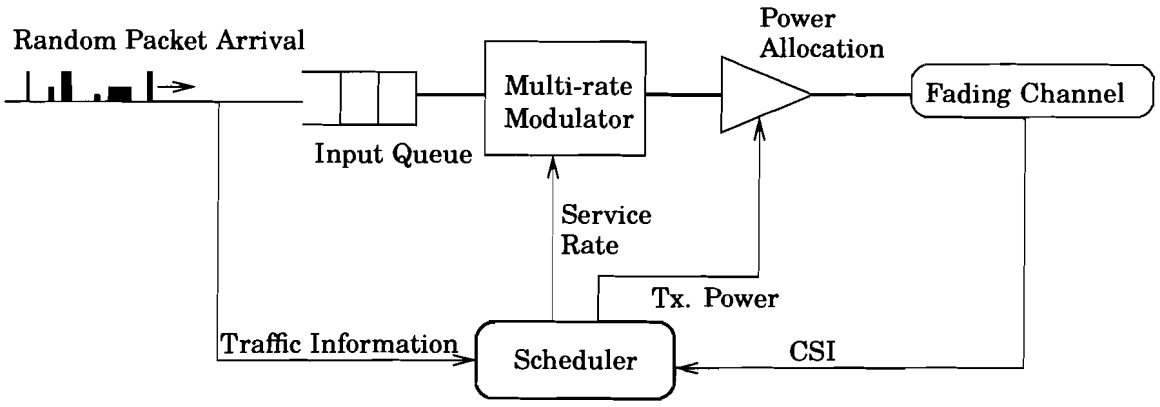

Figure 3.1: The system block diagram of the optimal cross-layer scheduler.

Let the random arrival time of packet i be denoted as $T_a(i)$, where the packet index i is assigned in the natural order of arrival such that $T_a(1) \leq T_a(2) \leq \ldots$. The corresponding packet expiry times $T_e(i) = T_a(i) + \tau(i)$ can be calculated by knowing the maximum allowable delay $\tau(i)$ of packet i. Furthermore, denote the amount of information contained in packet i as $D(i)$ *nats* (ln(2) times the number of bits) and observe that, in order to meet the transmission deadline of packet i, the system must transmit all information in packet i as well as all the packet preceding it. That is, the transmission rate $r(t)$ of the system must satisfy

$$\int_{t=0}^{T_e(i)} r(t)dt \geq \sum_{x=1}^{i} D(x) \quad \text{for all} \quad i \tag{3.1}$$

Similarly, the total amount of information serviced can never exceed that contained in the packets arrived. That is, the system cannot transmit more information than is available. Thus, we have a similar set of inequalities at each packet arrival instant:

$$\int_{t=0}^{T_a(i)} r(t)dt \leq \sum_{x=1}^{i-1} D(x) \quad \text{for all} \quad i \tag{3.2}$$

To allow for easier mathematical manipulation, the inequalities (3.1) and (3.2)

are reformulated as explicit function of time in terms of the cumulative arrivals and expiries. We will refer to the term on the left-hand-side of (3.1) and (3.2) as the cumulative rate function $R_{\text{cum}}(t) = \int_{x=0}^{t} r(x)\,\mathrm{d}x$ and the upper and the lower bounds of (3.1) and (3.2) can be rewritten as explicit function of time, the cumulative arrivals, $A_{\text{cum}}(t)$ and the Cumulative Expiries, $E_{\text{cum}}(t)$. The exact definition of these curves and its relation to the forms given previously is best demonstrated through a concrete example. Consider a sequence of eight packet arrivals as shown in Figure 3.2. The diagram shows the allowable maximum delay of each packet as the width of the box while the height of the box shows the information contained in each packet. With the boxes stacked in the order of arrival, it is easy to determine the exact form of the cumulative arrival, $A_{\text{cum}}(t)$ and the cumulative expiry, $E_{\text{cum}}(t)$ graphically such that

$$A_{\text{cum}}(t) \geq R_{\text{cum}}(t) \geq E_{\text{cum}}(t) \tag{3.3}$$

Since transmission rate is a non-negative quantity, it is also required that the cumulative rate be a non-decreasing function of time. In summary, we have the following requirement for the transmission rate profile of a variable rate max-delay constrained queueing system to be admissible.

Property 1 (Rate Bounding Property). *The transmission rate profile $r(t)$ of a variable service rate FIFO queue with per-packet maximum delay constraints is admissible if and only if it satisfies the following:*

1. *The cumulative information rate $R_{cum}(t)$ is upper bounded by the cumulative arrivals $A_{cum}(t)$.*
2. *The cumulative rate $R_{cum}(t)$ is lower bounded by the cumulative expiries $E_{cum}(t)$.*
3. *The transmission rate profile $r(t)$ is non-negative.*

Next, we present the formal definition of the convex optimization problem with the

objective being finding the admissible transmission rate profile that has the minimum energy usage.

3.1.2 Convex Formulation

Here, we consider the deterministic problem formulation where we assume that the traffic information is all known *a priori*. As for the cost of transmission, we assume a flat, time-varying, AWGN channel characterized by the time-varying power gain $g(t)$. Without loss of generality, we use a piecewise constant approximation of $g(t)$ which only allows the channel gain to change at discrete time instants. This block fading channel can be made arbitrarily close to a continuous channel by using a smaller time step.

Furthermore, due to the positivity constraint of the transmission rate, the upper and lower bound constraints only need to be tested at the arrival and expiry time of each packet (i.e. the corner point of the bunding staircase of Figure 3.2). This allows us to discretize the problem formulation into finite and discrete *epochs* for ease of computation. Following a very similar formulation to that used in [27], the time axis is divided into variable length segments known as epochs. An epoch is a continuous segment of time where the system parameters are constant. An epoch change takes place when the upper or lower bound on the cumulative rate changes, or when the channel state changes. That is, an epoch changes when a packet enters the queue, i.e. a change in upper bound, or at a constraining expiry constraint, i.e. a change in lower bound as depicted in Figure 3.2, or if the channel state changes which happens at discrete instants by the block fading approximation.

Let the *normalized duration* of epoch m, $x(m)$, be given by $x(m) = (t_{m+1} - t_m)W$, where W is the frequency bandwidth in Hz and t_m is the time at the beginning of epoch m. Since the channel gain, cumulative arrivals, cumulative expiries are con-

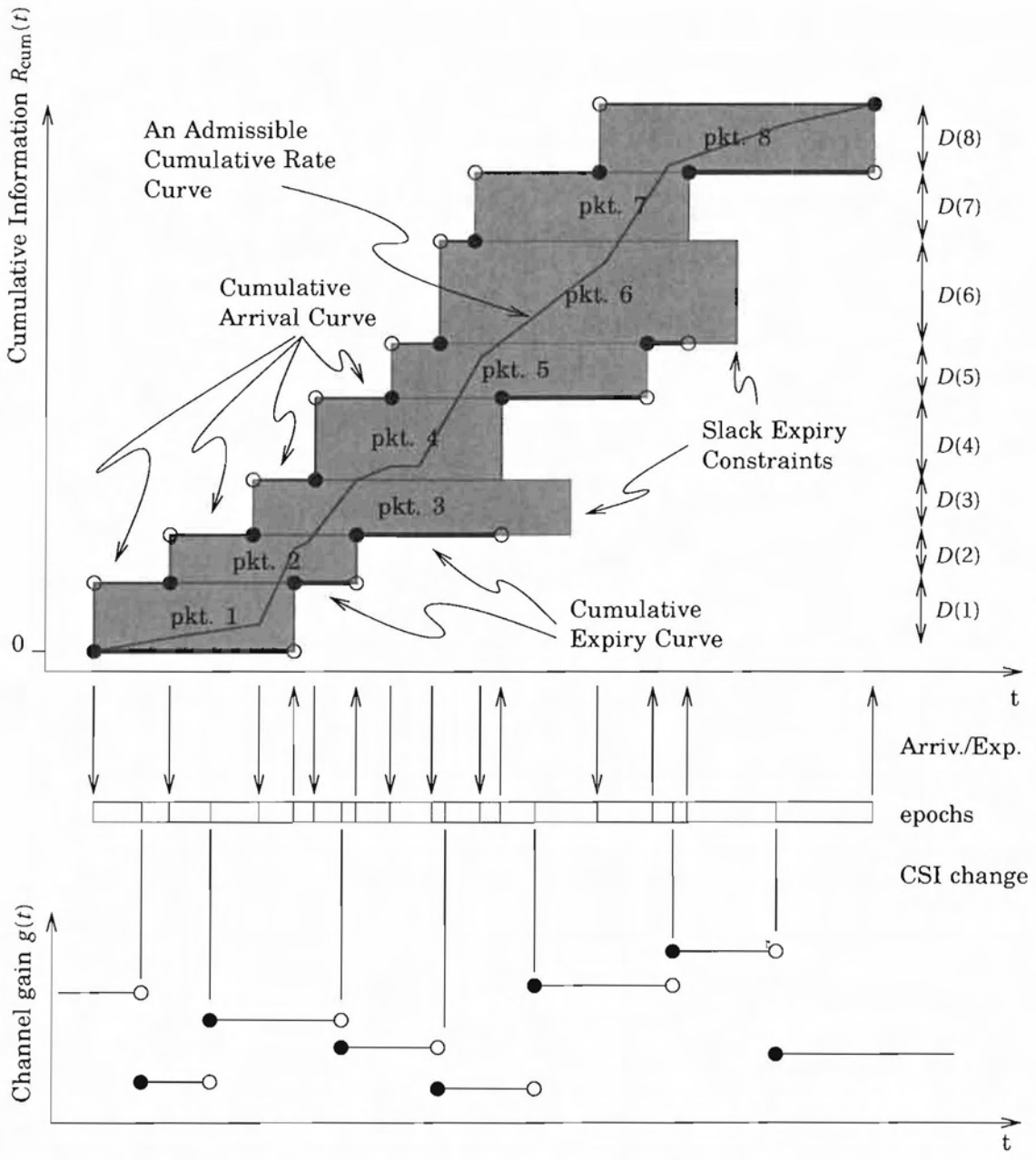

Figure 3.2: Diagram showing cumulative arrivals and expiries as the bound on the admissible cumulative rate profile. The solid dot (•) at the epoch boundaries shows the actual value taken for each of the parameters in discrete notation. In this diagram, the height of eack box corresponds to the size of the packet.

stant within each epoch, it follows that the optimal transmission rate must also be constant. For convenience, we will denote the discrete channel gain for epoch m as $g(m) = g(t_m)$, the discrete cumulative arrival and expiry as $A_{\text{cum}}(m) = A_{\text{cum}}(t_m)$ and $E_{\text{cum}}(m) = E_{\text{cum}}(t_m)$ respectively, and the transmission rate during epoch m as $r(m)$.

As noted previously, the cumulative rate curve is admissible if it meets the bounding constraints at arrival and expiry instants. The bounds specified in (3.3) can be stated in the discrete form as $A_{\text{cum}}(m) \geq \sum_{i=1}^{m} x(i)r(i) \geq E_{\text{cum}}(m)$. It together with $r(m) \geq 0$ forms the admissibility condition of the rate profile in the discrete form.

To formulate the objective of the optimization problem, let $f\big(r(m), g(m)\big)$ be the transmit signal to noise ratio (SNR) required to transmit at rate, $r(m)$, given a channel power gain of $g(m)$. As is common in adaptive transmission analysis, we consider the function to be increasing and strictly convex in the first argument, $r(m)$. Without loss of generality, also assume that no energy is used when the transmitter is idle, i.e., $f(0, g(m)) = 0$. The energy-optimal rate profile is therefore the solution to the following convex optimization problem, with infinite optimization horizon:

$$\text{Minimize:} \quad \sum_{m=1}^{\infty} x(m) f\big(r(m), g(m)\big) \tag{3.4a}$$

$$\text{Subject to:} \quad \sum_{i=1}^{m} x(i)r(i) \leq A_{\text{cum}}(m), \qquad \text{for} \quad m = 1 \ldots \infty \tag{3.4b}$$

$$\sum_{i=1}^{m} x(i)r(i) \geq E_{\text{cum}}(m), \qquad \text{for} \quad m = 1 \ldots \infty \tag{3.4c}$$

$$r(m) \geq 0 \qquad \text{for} \quad m = 1 \ldots \infty \tag{3.4d}$$

The cumulative arrivals and expiries form the upper and lower bounds of the cumulative rate profile in each epoch through the inequality constraints $A_{\text{cum}}(m) \geq \sum_{i=1}^{m} x(i)r(i) \geq E_{\text{cum}}(m)$ at each epoch boundary as shown in Figure 3.2.

While the power function $f(r, g)$ is assumed to be convex in the variable r with no other restriction for the problem to be solvable numerically, it is convenient to have an

actual expression to work in some places. For the toy problem, we assumed a specific form of $f(r,g)$ defined as

$$f(r,g) = \frac{e^r - 1}{g}; \qquad \frac{\partial}{\partial r} f(r,g) = \frac{e^r}{g} \tag{3.5}$$

following from Shannon's channel capacity [8] formulation.

3.1.3 A Numerical Example

Once a problem is formulated as a convex optimization problem, generic numerical methods can be used to derive the solutions efficiently. Here, we will give a numerical example of the optimization problem specified in the AMPL language [38] and solved using the freely available algencan solver from the TANGO project [39, 40] with the student edition of AMPL.

The AMPL equivalent of the problem definition (3.4) is given as follows:

```
param N;                                                    1
param g{t in 1..N};                                         2
param X{t in 1..N};                                         3
param A{t in 1..N};                                         4
param E{t in 1..N};                                         5
var r{t in 1..N};                                           6
                                                            7
minimize Energy:                                            8
        sum{t in 1..N} X[t]*(2^r[t]-1)/g[t];                9
subject to Arrival{t in 1..N}:                              10
        sum{i in 1..t}(X[i]*r[i]) <= A[t];                  11
subject to Expiry{t in 1..N}:                               12
```

```
        sum{i in 1..t}(X[i]*r[i]) >= E[t];                    13
subject to Positivity{t in 1..N}:                             14
        r[t] >= 0;                                            15
```

In the above model definition, we define N epochs with the associated system parameters g, X, A, and E in line 1–5, all as length N vectors. The desired variable r is declared in line 6. The rest of the model definition specifies the problem mathematically with the power function being specified by (3.5).

The actual numerical values of the various parameters are specified in a separate data file. To demonstrate the essential properties of *piecewise water-filling* we specify twelve epochs as follows:

```
param N := 12;                                                1
param:  g     X     A     E  :=                               2
    1   1.0   2     0     0                                   3
    2   0.7   2    10     0                                   4
    3   0.5   4    10     0                                   5
    4   0.1   2    25     0                                   6
    5   1     4    25    10                                   7
    6   0.5   2    25    10                                   8
    7   0.3   1    30    25                                   9
    8   0.5   2    30    25                                   10
    9   0.1   4    30    25                                   11
   10   0.5   5    40    25                                   12
   11   1     2    40    30                                   13
   12   0.7   4    40    40   ;                               14
```

The optimal rate profile can be calculated by running the two file listings through

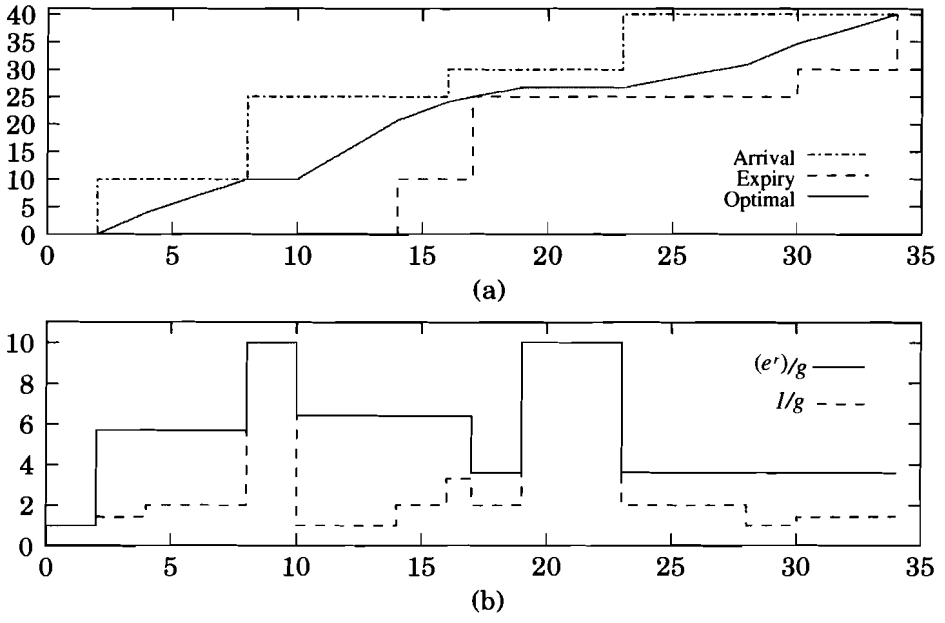

Figure 3.3: The optimal cumulative rate profile for the numerical example (a) and the corresponding water-filling diagram showing the piecewise water-filling property in (b). The optimal transmit power is the difference of the two curves shown in (b), $f(r,g) = (e^r - 1)/g$.

AMPL to obtain the optimal transmission rate for the twelve epochs. The cumulative rate profile with the cumulative arrival and cumulative expiry bounds are shown in Figure 3.3 (a). A more revealing graph is the plot of the equivalent noise power $1/g$ (floor to water-fill against) and the total noise and transmit power $(e^r - 1)/g + 1/g$ (the water level) in Figure 3.3 (b) which clearly shows the *piecewise water-filling* phenomena.

From the numerical result presented, we make several observations about the optimal rate profile for traffic with strict deadline constraints.

Observation 1. *The power allocation associated with the optimal rate profile follows a piecewise water-filling property where a constant waterlevel is maintained across multiple epochs and the waterlevel changes whenever the cumulative rate curve touches the arrival or expiry bounds.*

Furthermore, with several random trials, we also observe that

Observation 2. *The piecewise constant waterlevel can increase only if the cumulative rate touches the arrival bound. The converse also holds. That is, the waterlevel can decrease only if the cumulative rate touches the expiry bound.*

It should also be noted that waterlevel change does not occur when the floor to waterfill against, $1/g$ exceeds the waterlevel, i.e. when $e^r < 1$. This can be clear seen in Figure 3.3 where the first occurence of $e^r < 1$ has different waterlevel on either side while the waterlevel stays the same across the second occurence of $e^r < 1$. We will provide the formal proof of these observations next.

3.1.4 KKT Conditions and Their Interpretations

In the previous section, we have presented some observations without proof about the general structure of the optimal solution. In this section, we provide the mathematical proof that the fundamental structure of the solution is water-filling across the epochs between active arrival and expiry constraints (where the cumulative rate curve touches the arrivals or expiries staircases, as shown in Fig. 3.3 (a)). Consider the optimisation problem (3.4) and assign the non-negative Lagrangian multipliers $\{\mu_1 \ldots \mu_m \ldots\}$, $\{\nu_1 \ldots \nu_m \ldots\}$, and $\{\lambda_1 \ldots \lambda_m \ldots\}$ to constraints (3.4b), (3.4c) and (3.4d) respectively. The Lagrangian, $\mathscr{L}$ is

$$\mathscr{L} = \sum_{m=1}^{\infty} x(m) f\big(r(m), g(m)\big) - \sum_{m=1}^{\infty} \mu_m \left[A_{\text{cum}}(m) - \left(\sum_{i=1}^{m} x(i) r(i) \right) \right] - \sum_{m=1}^{\infty} \nu_m \left[\left(\sum_{i=1}^{m} x(i) r(i) \right) - E_{\text{cum}}(m) \right] - \sum_{m=1}^{\infty} \lambda_m r(m) \quad (3.6)$$

Next, we differentiate the Lagrangian with respect to $r(m)$ to obtain the set of

conditions

$$x(m)\frac{\partial}{\partial r(m)}f\big(r(m),g(m)\big)+x(m)\sum_{i=m}^{\infty}\mu_i-x(m)\sum_{i=m}^{\infty}\nu_i-\lambda_m=0. \tag{3.7}$$

Examination of (3.7) shows that through subtraction of (3.7) for m and $m-1$, the partial summation can be eliminated. An equivalent set of Lagrangian conditions can be stated as

$$\begin{aligned}\mu_m-\nu_m=&\frac{\partial}{\partial r}f\big(r(m),g(m)\big)-\frac{\lambda_m}{x(m)}\\&-\left[\frac{\partial}{\partial r}f\big(r(m-1),g(m-1)\big)-\frac{\lambda_{m-1}}{x(m-1)}\right]\quad\text{for all}\quad m=1\cdots\infty\end{aligned} \tag{3.8}$$

with the appropriate definition for $f(r(0),g(0))=0$ and $\lambda_0=0$ to take care of the initial condition of the difference relation.

The right hand side of (3.8) can be further simplified by noting that the complementary slackness conditions associated with μ_m and ν_m are

$$\sum_{i=1}^{m}x(i)r(i)=A_{\text{cum}}(m)\iff\mu_m>0 \tag{3.9}$$

$$\sum_{i=1}^{m}x(i)r(i)=E_{\text{cum}}(m)\iff\nu_m>0 \tag{3.10}$$

Thus, when the cumulative rate touches the arrivals or expiries staircases, the constraints are activated with $\mu_m>0$ or $\nu_m>0$, respectively. As a special case, epochs in which the two staircases are equal have an empty queue for any feasible solution, so the rate is zero. Such an epoch produces a finite optimization horizon. Consider a sequence of epochs with slack arrival and expiry constraints, i.e., $\mu_m=\nu_m=0$, and delimited by active arrival or expiry constraints. Within such a *water-fill segment* we

have

$$\begin{aligned}\omega &= \frac{\partial}{\partial r} f\big(r(m), g(m)\big) - \frac{\lambda_m}{x(m)} \\ &= \frac{\partial}{\partial r} f(r(m+1), g(m+1)) - \frac{\lambda_{m+1}}{x(m+1)}\end{aligned} \tag{3.11}$$

Here, a constant ω is used to simply denote the value of this chain of equality. Also recall the complementary slackness condition that $\lambda_m = 0$ whenever the rate allocation in epoch m is non-zero.

Lets first consider (3.11) for the epochs within the same water-fill segments. This has several implications. Firstly, two adjacent epochs must have identical power gradient in rate, i.e.

$$\frac{\partial}{\partial r} f\big(r(m), g(m)\big) = \frac{\partial}{\partial r} f\big(r(m+1), g(m+1)\big) \tag{3.12}$$

if r_m and r_{m+1} are both non-zero. Secondly, whether rates are zero or not, (3.11) can be chained in sequence to obtain

$$\begin{aligned}\frac{\partial}{\partial r} f\big(r(m), g(m)\big) - \frac{\lambda_m}{x(m)} &= \frac{\partial}{\partial r} f\big(r(m+1), g(m+1)\big) - \frac{\lambda_{m+1}}{x(m+1)} \\ &= \ldots \\ &= \frac{\partial}{\partial r} f\big(r(m+j), g(m+j)\big) - \frac{\lambda_{m+j}}{x(m+j)} = \omega\end{aligned} \tag{3.13}$$

for a sequence of j epochs within the same water-fill segment. The relation to water-filling is evident if the power function takes the form specified by the capacity equation. By substituting the derivative of (3.5) into (3.11), we obtain

$$\frac{e^{r(m)}}{g(m)} = f\big(r(m), g(m)\big) + \frac{1}{g(m)} = \omega + \frac{\lambda_m}{x(m)} \tag{3.14}$$

for all epochs connected by slack arrival and departure constraints. Making one final substitution for the complementary slackness condition for λ_m and using the fact that

$x(m) > 0$, we must have,

$$f\big(r(m), g(m)\big) = \begin{cases} \omega - \dfrac{1}{g(m)} & \text{for } \dfrac{1}{g(m)} \le \omega \quad (3.15a) \\ 0 & \text{for } \dfrac{1}{g(m)} > \omega \quad (3.15b) \end{cases}$$

which is the well known water-filling condition on power, and hence the piecewise water-filling property observed previously.

The water-filling in power as expressed in (3.15a) and (3.15b) can also be expressed as water-filling in rate. Solving (3.14) for rate, we obtain

$$r(m) = \begin{cases} \ln(\omega) + \ln(g(m)) & \text{for } \ln(\omega) \ge -\ln(g(m)) \quad (3.16a) \\ 0 & \text{for } \ln(\omega) < -\ln(g(m)) \quad (3.16b) \end{cases}$$

Since $1/g(m)$ is the normalized transmit noise power, we will refer to its logarithm as the *noise rate* and obtain the water-filling in rate where the optimal information rate is the difference between the rate water level $\ln(\omega)$ and the noise rate. Furthermore, we will refer to the epochs with a zero rate assignment in the optimal solution as the *idle epochs*. An idle epoch can be identified by $\ln(\omega) < -\ln(g(m))$.

Next, we turn our attention to the epoch boundaries that are not slack, i.e., either μ_m or ν_m is not zero. Provided that the queue is not in the empty state, the cumulative arrival bound can never equal the cumulative expiry bounds and without loss of generality, we assume that only one of arrival or expiry constraints can be active at any instant, hence only one of μ_m and ν_m is non-zero for any given m. Thus, the complementary slackness conditions combined with (3.8) can be re-written as

$$\mu_m > 0 \iff \frac{\partial}{\partial r} f\big(r(m), g(m)\big) - \frac{\lambda_m}{x(m)} > \frac{\partial}{\partial r} f\big(r(m-1), g(m-1)\big) - \frac{\lambda_{m-1}}{x(m-1)} \quad (3.17a)$$

$$\nu_m > 0 \iff \frac{\partial}{\partial r} f\big(r(m), g(m)\big) - \frac{\lambda_m}{x(m)} < \frac{\partial}{\partial r} f\big(r(m-1), g(m-1)\big) - \frac{\lambda_{m-1}}{x(m-1)} \quad (3.17b)$$

Since the inequality on the right hand side (RHS) of (3.17) simply compares the waterlevels for the previous and current waterfill segments by (3.13), an active arrival constraint, left hand side (LHS) of (3.17a), implies an increase in waterlevel, RHS of (3.17b) and an active expiry constraint implies an decrease in waterlevel, as observed numerically in the previous section.

The optimal solution can thus be interpreted as a set of different water-filling solutions applied to sequence of epochs delimited by active arrival or expiry constraints. These constraints can become active only at concave corners of the arrivals and expiries staircases, since the cumulative rate curve is non-decreasing. The curve can run along the "tread" of a staircase only in epochs with sub-threshold gain, where the rate is already zero.

In this section, we presented the formulation of a simple cross layer optimization problem that considers the dynamics of both channel and traffic load variations in a typical point-to-point communication system. By formulating the problem as a convex optimization problem, a numerical example was given which reveals an interesting property about the optimal solution, that the optimal rate profile can be described as *piecewise water-filling*, with changes in water level occurring only when the system meets active arrival or expiry constraints. In additional to examine the problem numerically, analytical proof of the *piecewise water-filling* property is also presented through the analysis of the KKT conditions.

3.2 Diversity Channels

Next, we consider a more general channel model where multiple parallel channels are available. Specifically, the following description uses orthogonal frequency division multiplexing (OFDM) as a specific example. However, the formulation that follows

can be applies to any single-user multi-carrier channels such as time division multiple access (TDMA) channels or even eigen-channels of multi-input multi-output (MIMO) systems.

With a multi-carrier channel, the general form of piecewise water-filling still holds and the optimal rate allocation is in the form of water-filling across time and subcarriers and the waterlevel is only allowed to change at time instants corresponding to active arrival or expiry constraints. The analysis of the multi-carrier system follows closely the analysis of the toy problem presented previously but differs in the following significant ways:

1. A multi-carrier channel is used.
2. A periodic symbol-spaced quantization of the time axis is used in place of the concept of epoch.
3. A more realistic power to rate formulation for m-ary quadrature amplitude modulation (mQAM) adaptive modulation [6] is used instead of ergodic capacity.
4. A generalized piecewise water-filling solution is derived.

3.2.1 System Model

The system under consideration is a mobile point-to-point wireless system utilizing multiple parallel channels for transmission. The user data arrives randomly into a queue and is serviced at a rate that is adjusted dynamically by the scheduler. The scheduler also varies the transmission power for all channels jointly with the transmission rate to minimize total energy usage. The system block diagram is shown in Figure 3.1.

Instead of dividing time into variable length epochs, the transmission of data is organized into constant duration time slots or conceptually, OFDM symbols, during which the channel is assumed constant. We also assume that the packet arrival time

and its associated deadline constraint are also quantized to the appropriate symbol boundary. That is, a packet arriving during the symbol m is available for transmission only in symbol $(m+1)$ or beyond and we quantize the arrival time to be at the end of symbol m. Similarly, a packet set to expire during symbol m must be serviced on or before symbol $(m-1)$ and we quantize the expiry time to be at the beginning of symbol m.

Notations similar to the toy problem will be used with an additional subscript to denote the subcarrier. Specifically, we will denote the subcarriers in the channel with a subscript n for an N channel system and we denote all quantities that are defined for each subcarrier individually with a small case letter and the corresponding capitalized version without the subscript n to denote quantities that are summed across all subcarriers, i.e. $r_n(m), p_n(m)$ denote the rate and power for symbol m in subcarrier n while $R(m), P(m)$ denote the total rate and power for symbol m in all the subcarriers. Furthermore, with the extensive use of cumulative quantities to specify the admissible rate profile, we will use the shorthand $R_{\text{cum}}(m) = \sum_{i=1}^{m} R(m)$ to denote the partial sum of the first m elements of the sequence $R(m)$, similar to the previous definition of A_{cum} and E_{cum}. Also note that a shorter notation $p_n(m)$ is used to denote power, and its dependence on the instantaneous transmission rate and channel gain should be noted implicitly.

Next, we turn our attention to the relationship between the required transmit power, $p_n(m)$, the desired modulation rate $r_n(m)$ (measured in nats per symbol per channel), the channel power gain, $g_n(m)$ and the bit error rate (BER) obtained with finite constellation.

Following the formulation given in [14], the achievable transmission rate is

$$r_n(m) = \ln\left(1 + K\frac{g_n(m)p_n(m)}{WN_0}\right) \tag{3.18}$$

where K is a constant related to the desired instantaneous target BER by

$$K \triangleq \frac{-1.5}{\ln(5\mathrm{BER}_{\mathrm{target}})} \tag{3.19}$$

for uncoded mQAM. In this formulation, we denote the receiver noise spect ral density as N_0 and the system bandwidth as W. (3.19) can be further modified with a coding gain multiplier for trellis coded mQAM [6] for finer rate and power steps and also for turbo coded systems [10]. Here, we make the same assumption as [14, 6], that a continuous set of rates is achievable. In practice, we approximate a continuous set of rates through various combinations of coding and modulation. This continuous approximation is necessary to provide a tractable convex optimization formulation of the problem. If $K = 1$, (3.18) simply reduces to the achievable rate expression specified by Shannon's capacity equation and is the form considered for the toy problem and as in [37].

3.2.2 Convex Optimization Formulation

The scheduling problem is equivalent to solving the convex optimization problem

$$\text{Minimize:} \quad \sum_{m=1}^{M} \sum_{n=1}^{N} p_n(m) \tag{3.20a}$$

$$\text{Subject to:} \quad \sum_{i=1}^{m} \sum_{n=1}^{N} r_n(i) \geq E_{\mathrm{cum}}(m) \quad \text{for} \quad m \in [1, M] \tag{3.20b}$$

$$\sum_{i=1}^{m} \sum_{n=1}^{N} r_n(i) \leq A_{\mathrm{cum}}(m) \quad \text{for} \quad m \in [1, M] \tag{3.20c}$$

$$r_n(m) \geq 0 \quad \text{for} \quad n \in [1, N];\ m \in [1, M] \tag{3.20d}$$

for the rate and power profile $\{r_n(m), p_n(m)\}$ over the OFDM tones (n) and over time (m) that minimizes the total energy (3.20a).

Without loss of generality, we assume that constraint (3.20b) and (3.20c) are both

satisfied with strict equality if and only if $m = M$. In other words, in the mathematical formulation, we disallow the cumulative arrival bounds to equal the cumulative expiry bounds in the interior of the optimization problem. This assumption does not compromise the validity of the model since any general problem can be separated into smaller subproblems at these instants and be solved independently from each other. This assumption also allows only one of the constraints to become active at any instant to provide an uncluttered view of the piecewise water-filling property.

3.2.3 KKT Conditions and Their Interpretations

By assigning the Lagrangian multipliers $\mu(m)$, $\nu(m)$ and $\lambda_n(m)$ to constraints (3.20b), (3.20c) and (3.20d) respectively, and differentiating with respect to $p_n(m)$, we obtain the set of equations

$$1 - \frac{\partial r_n(m)}{\partial p_n(m)}[\omega(m)] - \lambda_n(m) = 0 \qquad \text{for all} \quad n, m \tag{3.21}$$

where we define a new slack variable

$$\omega(m) = \sum_{i=m}^{M} \mu(i) - \sum_{i=m}^{M} \nu(i). \tag{3.22}$$

Note that $\omega(m)$ is constant over all tones in any given OFDM symbol m. By substituting the derivative of (3.18) into (3.21) and rearranging, we obtain

$$p_n(m) \;=\; \frac{\omega(m)}{1-\lambda_n(m)} - \frac{WN_0}{Kg_n(m)} \tag{3.23}$$

By the complementary slackness condition associated with constraint (3.20d),

$$\lambda_n(m) > 0 \iff p_n(m) = 0, \tag{3.24}$$

it follows that the power allocation within an OFDM symbol is in the form of water-filling in frequency (n), i.e., within any given OFDM symbol m,

$$p_n(m) = \begin{cases} \omega(m) - \frac{WN_0}{Kg_n(m)} & \text{for } \omega(m) \geq \frac{WN_0}{Kg_n(m)} \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

where $\omega(m)$ is the water-level and the term $WN_0/Kg_n(m)$ is the floor to water-fill against.

Next, we study the behaviour of the solution across the time dimension m. By examining the expression of the waterlevel ω for two consecutive OFDM symbols (3.22), we find that the difference in waterlevel $\Delta\omega(m) \triangleq \omega(m+1) - \omega(m)$ is

$$\Delta\omega(m) = \nu(m) - \mu(m) \qquad \text{for} \quad m = 1 \ldots M-1 \tag{3.26}$$

Note that the RHS of (3.26) are the slack variables associated with constraint (3.20b) and (3.20c), and since both of these constraints cannot be active simultaneously except when $m = M$, we have the following equivalent set of complementary slackness conditions for $\Delta\omega(m)$.

$$\Delta\omega(m) = \nu(m) > 0 \iff A_{\text{cum}}(m) = \sum_{i=1}^{m} R(i) > E_{\text{cum}}(m) \tag{3.27}$$

$$\Delta\omega(m) = 0 \iff A_{\text{cum}}(m) > \sum_{i=1}^{m} R(i) > E_{\text{cum}}(m) \tag{3.28}$$

$$\Delta\omega(m) = -\mu(m) < 0 \iff A_{\text{cum}}(m) > \sum_{i=1}^{m} R(i) = E_{\text{cum}}(m) \tag{3.29}$$

Thus, we have the following property,

Property 2 (Piecewise Water-filling Property). *The optimal power allocation $p_n(m)$ obeys the water-filling property across the tones of every OFDM symbol. Furthermore,*

when both constraints are slack, i.e., (3.28)*, the water-level for the next OFDM symbol* $\omega(m+1)$ *must equal the water-level of the current symbol* $\omega(m)$*. Thus, this waterlevel is piecewise constant and can change only when one of the traffic constraints* (3.20b), (3.20c) *is active. An active arrival constraint leads to an increase in waterlevel by* $\nu(m)$, (3.27)*, and an active expiry constraint leads to a decrease in waterlevel by* $\mu(m)$, (3.29).

For our purposes, rate is a more convenient quantity to consider since the traffic constraints are stated in terms of cumulative rates. By substituting (3.25) into (3.18) and rearranging, we obtain

$$r_n(m) = \begin{cases} \ln(\omega(m)) - \ln\left(\frac{WN_0}{Kg_n(m)}\right) & \text{for } \omega(m) > \frac{WN_0}{Kg_n(m)} \\ 0 & \text{otherwise} \end{cases} \tag{3.30}$$

That is, the optimal rate allocation is to allocate according to the water-filling principle with the floor $\ln(WN_0/Kg_n(m))$ and a waterlevel $\ln(\omega(m))$. Recall that a sequence of OFDM symbols with slack arrival and expiry constraints must share the same waterlevel $\omega(\cdot)$. Hence $\ln(\omega(m)) = \ln(\omega(m-1))$. Similarly, it follows that $\ln(\omega(m-1)) > \ln(\omega(m))$ if the expiry constraint is active and $\ln(\omega(m-1)) < \ln(\omega(m)$ if the arrival constraint is active. Thus the rate profile must follow the same piecewise water-filling property as the power profile. Thus the rate profile, like the power profile, must exhibit a piecewise water-filling property.

Property 3 (Piecewise Rate Water-filling Property)**.** *The optimal rate allocation* $r_n(m)$ *obeys the water-filling property across every OFDM tone. Furthermore, when both constraints are slack, i.e.,* (3.28)*, the water-level for the next OFDM symbol* $\ln(\omega(m+1))$ *must equal the water-level of the current symbol* $\ln(\omega(m))$*. Thus, this waterlevel is piecewise constant and can change only when one of the traffic constraints* (3.20b), (3.20c) *is active. An active arrival constraint leads to an increase in waterlevel by*

$\nu(m)$, (3.27), *and an active expiry constraint leads to a decrease in waterlevel by* $\mu(m)$, (3.29).

For the prescient scheduler, the instantaneous transmit power and rate are assigned jointly and are related by the convex power function $p(r,g)$. Whether rate or power are calculated from the optimization problem is not important as with perfect knowledge of the channel gain, the prescient scheduler can always convert from one to the other without ambiguity.

3.2.4 Numerical Result

In this section, we present a small time snapshot of a simulation to illustrate some of the main property of the prescient scheduler. An instance of a two-carrier system with independent Rayleigh fading and a randomly generated arrival and expiry bounds is shown in Figure 3.4. We present here only a snapshot of 15 packets to demonstrate the main properties of piecewise water-filling for multiple subcarriers. Simulation results showing the average energy usage over a large number of packets will be presented later in Chapter 5.

The graph presented in Figure 3.4(a) shows the optimal cumulative transmission rate and the cumulative arrival and expiries bounds. The more revealing graph is the rate water-filling graphs shown in Figure 3.4(b) and (c) for each of the two subcarriers. In these diagrams, the floor to waterfill against is the log of the equivalent noise power, $\ln(WN_0/Kg_n(m))$, and the transmission rate is indicated as the shaded region over and above this floor with the waterlevel $\ln(\omega(m))$ (see (3.30)). In addition to the piecewise waterfilling rate allocation for each subchannel, it can also be clearly seen that the waterlevel is shared across the two subchannels and the waterlevel changes occur simultaneously.

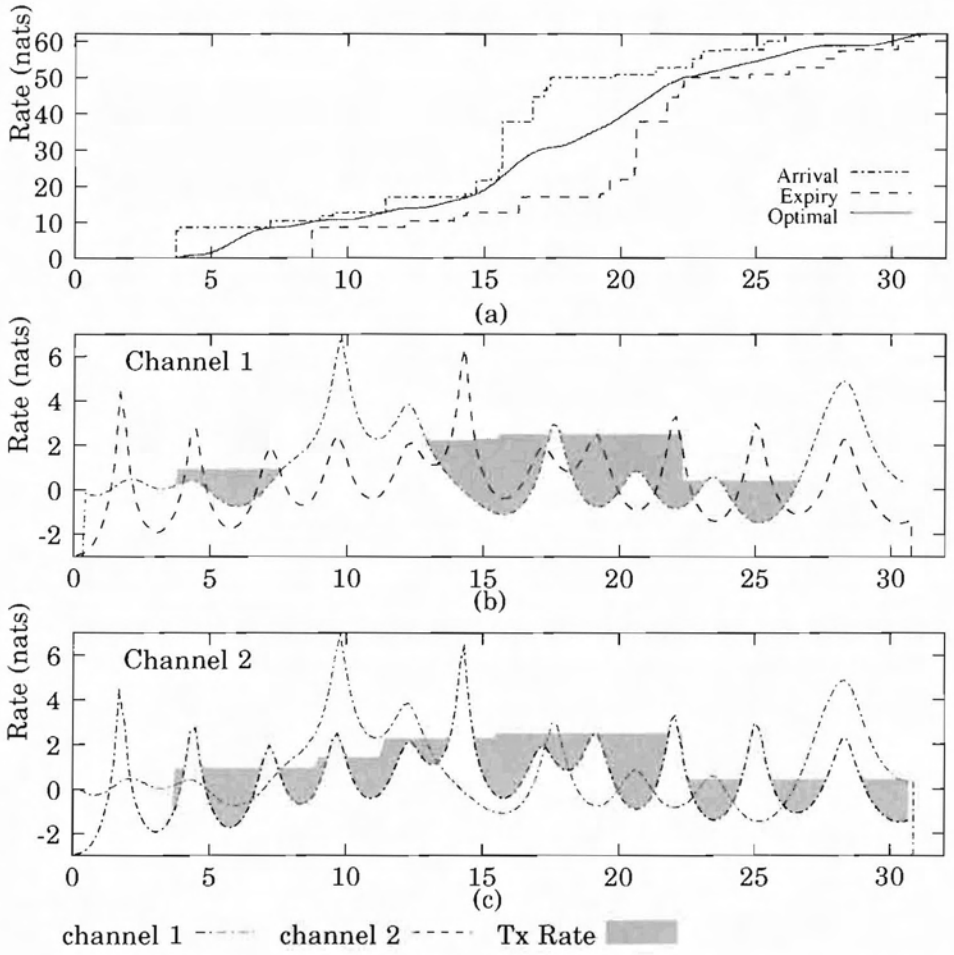

Figure 3.4: The optimal cumulative rate profile for a two-carrier system and the respective water-filling in rate diagram.

Chapter 4

Efficient Optimization Algorithms

In the previous chapter, the piecewise water-filling properties for both power and rate were demonstrated. With a convex optimization problem formulation, a solution can be easily obtained through generic convex optimization software such as the free algencan solver [39, 40]. However, the student edition is restricted to solving problems with 300 objectives plus constraints, which restricts the problem size that can be simulated to less than 100 epochs for the single-carrier system, and proportional reduction with more carriers. Furthermore, the computational complexity of these generic solvers grows as the cube of the dimensionality, that is $O(n^3)$, without exploiting the structure of the problem [32]. For our problem, $n = NM$, where M is the number of epochs and N is the number of subcarriers. This motivates the following development of an efficient method of finding the optimal solution by exploiting the piecewise water-filling property observed in Chapter 3.

The algorithm operates in a transformed space where the the optimal solution can be determined as a shortest path under cumulative arrival and expiry bounds through

geometric constructions. The finding of the proper transformed space is iterative while the process of finding the shortest path can be visualized as pulling a piece of string through a series of pegs, hence the name iterative string pulling (ISP) algorithm. A limited form of the shortest path property for a static single carrier channel has been observed previously in [28]. It can be seen as a special case, where the iterative transform is not required under this limiting channel condition.

While the ISP algorithm presented here has reduced computational complexity and has an interesting "string-pulling" interpretation, it is not realizable. This is because the channel state information and the traffic information at all times must be known *a priori* in order to specify the parameters of (3.20). Nevertheless, this prescient scheduler leads to some insights that are used in Chapter 5 to develop a practical causal scheduler.

For the rest of this chapter, we begin with a quick numerical demonstration of the shortest path property for the single carrier static channel followed by the algorithmic definition of the "string-pulling" algorithm, which is an $O(n^2)$ algorithm, for $n = NM$. The transformation required to take into account channel variations is given in Section 4.2, with the iterations presented in Section 4.2.2. Finally, we conclude this chapter with the complete algorithm in Section 4.3.

4.1 Shortest Path Property

In this section, we give a brief description of the shortest path property for a static channel simplification of the single-user, single-carrier toy problem presented earlier. By setting the channel gain $g(m) = 1$ for all m, the water-filling power allocation within each waterfill segment becomes a much simpler constant power allocation, as the floor to waterfill against is now a constant. Furthermore, the increase and decrease

of the waterlevel according to whether the cumulative arrival or expiry constraints are active can now be simplified to an increase or decrease of the actual transmit power. That is, when the floor to waterfill against is a constant, the behaviour of the waterlevel as observed in the previous chapter can be directly applied to the actual power allocation. Furthermore, the optimal transmission rate profile must also follow a similar piecewise constant property, as it is proportional to the power allocation. Thus, the power and rate are both piecewise constant, with changes only at active arrivl or expiry constraints.

Next, consider the bounding property for cumulative rate as shown in Figure 3.2. Since the cumulative rate is the time integral of the transmission rate profile $r(t)$, it follows that the cumulative rate curve must be continuous, non-decreasing, and piecewise linear. Furthermore, this piecewise linear cumulative rate curve can only bend at active constraint points where the cumulative rate profile touches the bounds. We further observe that by Property 2, the cumulative rate curve bends upwards when touching the upper bound and bends downwards when touching the lower bound. This is precisely the behaviour of a piece of taut string under bounding constraints as illustrated in Figure 4.1. The illustration also demonstrates a possible algorithm for finding the shortest path by considering the constraints one at a time.

Can it happen that there exists another piecewise linear curve within the bounding constraints that exhibits the same property without being the shortest? Visual inspection of Figure 4.1 shows that such curve is unique provided that both ends of the string are fixed.

4.1.1 String Pulling Algorithm

The string-pulling analogy also provides us with an efficient way of computing the optimal cumulative rate profile from the given constraint points (peg positions). The

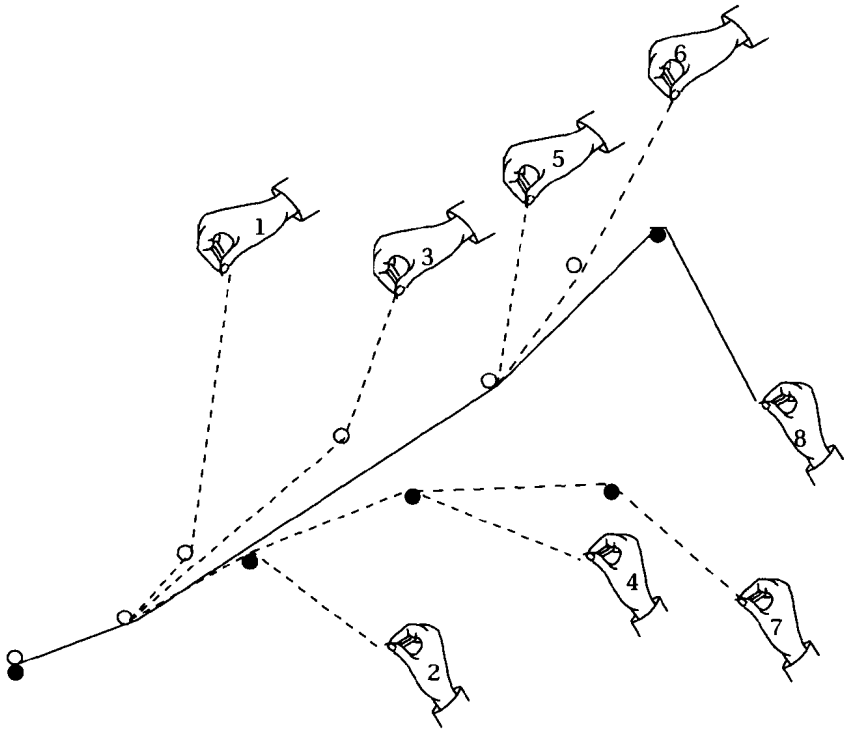

Figure 4.1: An illustration of the string-pulling process.

procedure provided here can compute the result in $O(n^2)$ where $n = 2NM$ is the number of constraint points. The process is illustrated graphically in Figure 4.1 and an algorithmic description is given below. The string-pulling algorithm takes a list of bounds B[•] as input, and as output, produces a list A[•] selected from the elements of B[•] which specifies the points through which the feasible shortest path must pass. The string-pulling procedure is most concisely described as

```
1: function StringPulling(B[])
2:     INPUT: B[]: List of Bounds
3:     OUTPUT: A[]: List of Bounds
4:     A[] ← B[1...2]
5:     end ← 2
6:     for i ← 3...K do
7:         Append B[i] to A[]
8:         end ← end + 1
9:         for j ← (i − 1) downto 2 do
10:            G1 ← Gradient (A[end − 1], A[end])
11:            G2 ← Gradient (A[end − 2], A[end − 1])
12:            if G1 < G2 and IsUpperBound(B[j]) then
13:                A[end − 1] ← B[j − 1]
14:            else
15:                Insert B[j − 1] before A[end − 1]
16:                end ← end + 1
17:            end if
18:            if G1 > G2 and IsLowerBound(B[j]) then
19:                A[end − 1] ← B[j − 1]
20:            else
```

```
21:              Insert B[j − 1] before A[end − 1]
22:              end ← end + 1
23:           end if
24:        end for
25:     end for
26:     return A
27: end function
```

In the algorithm specification, the input B[] is a list of co-ordinate points and can be implemented as a structure data type. By allowing each element of B[] to specify whether it is an upper or lower bound, a single list B[] is used to specify both the arrival and expiry bounds. The functions `IsUpperBound` and `IsLowerBound` are used to determine the type of bounds while the `Gradient` function simply computes the gradient between two bounding points as $(x_1 - x_2)/(y_1 - y_2)$. The overall operation of the algorithm is as follows. It constructs the shortest path by firstly assigning a temporary path following a straight line connecting the first two bounds (line 4). This temporary path is then extended with the next bound (7). The for-loop in lines 9 to 24 performs the checks of all the previous bounds to ensure that the shortest path is found. The two if-statements in line 12 and line 18 check if the shortest path is tight against the bounds and adds or subtracts the point from the returned list A[•] appropriately. Thus, for a problem setup with n bounds, there are $1 + 2 + \ldots + n = n^2/2$ checks in the form of lines 12 and 18 which are much simpler than the generic optimization solver solutions that require $O(n^3)$ computations of the gradients of the Lagrangians [32].

4.2 Information and Noise Rate

Next, we turn our attention back to the time-varying multi-carrier channel problem and show that through a simple transformation, we can account for the time varying channel and use the string-pulling algorithm iteratively to solve the full time-varying multi-carrier channel problem.

Recall that the optimal cumulative rate can be interpreted as the time integral of the rate profile that obeys the piecewise water-filling (PWF) property if the channel is static. However, with time varying channel gain that causes the cost of transmission at a certain rate to fluctuate, the optimal cumulative rate profile, as seen in the simple example of Figure 3.3, is no longer the shortest straight line. For the time varying channel, the trick is to work instead with the time integral of the constant waterlevel which has similar properties that lead to the shortest path interpretation.

A numerical example of a single-user-scheduling problem and its solution is shown in Figure 4.2 to aid the explanation of the ISP algorithm. The top and bottom graph show the cumulative rate and the water-filling plot of the solution. The middle plot shows the required transform in order to apply the shortest path interpretation. There are two important processes in finding the correct domain for performing string pulling. Firstly, the bounds are transformed by adding an appropriate offset (the noise rate) to the cumulative arrival and expiry bounds. Secondly, a process of identifying and removing idle epochs (shaded region in Figure 4.2), known as puncturing, is required since the shortest path property only applies to the set of non-idle epochs.

First, define the set of *active tones* for the m-th orthogonal frequency division multiplexing (OFDM) symbol as the set of tone indexes $n \in \Omega_m$ if and only if $\omega(m) > WN_0/(Kg_n(m))$. That is, Ω_m contains the indexes of the OFDM tones where the rate or power allocation is strictly positive as indicated by (3.30). We will refer to the com-

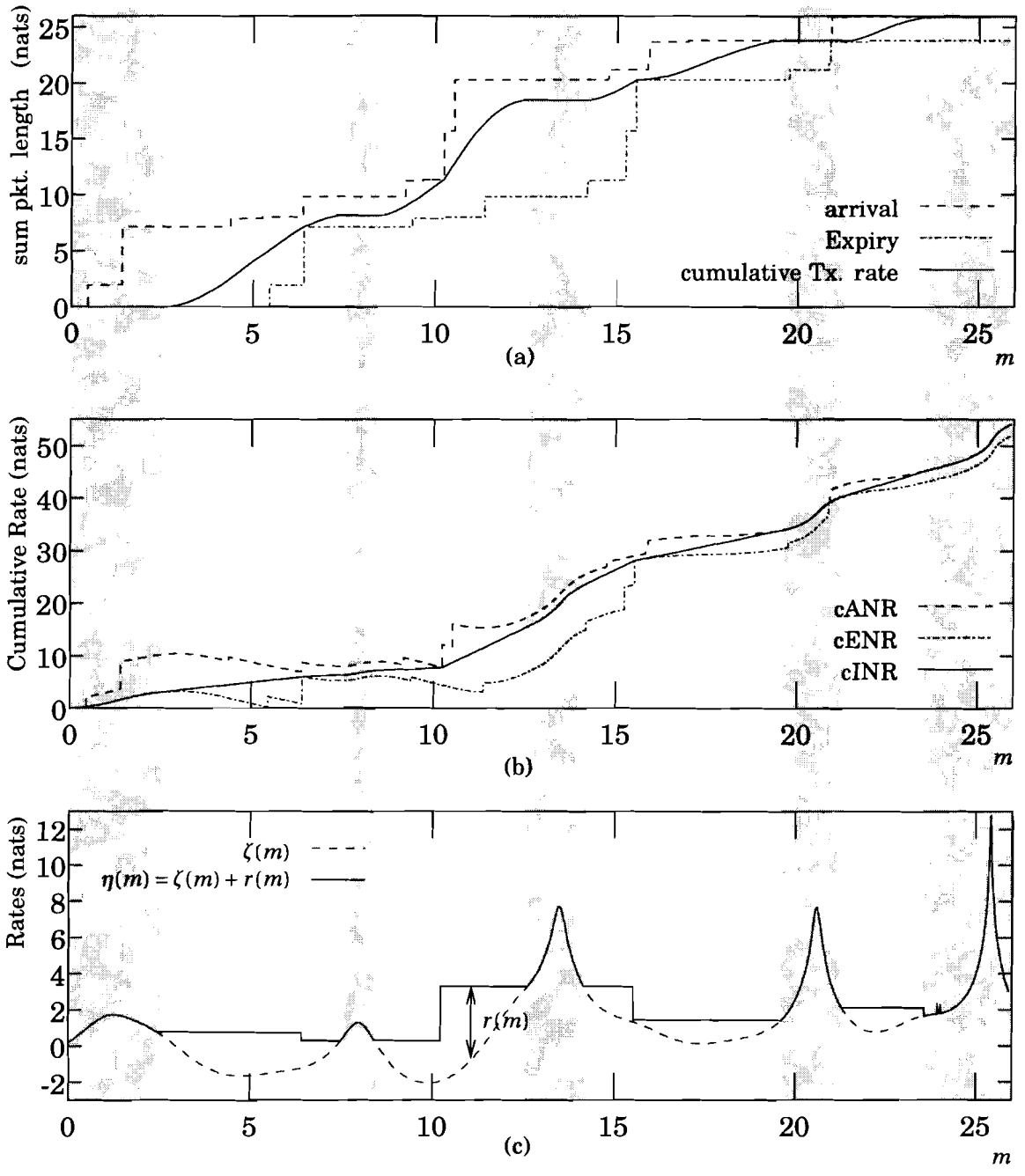

Figure 4.2: Diagram showing the relationship between the cumulative arrivals and expiries (a) and the water-filling diagram (c). The transformation required to show the shortest path property is shown in (b). Idle epochs are highlighted by the vertical strips. The cINR is piecewise linear outside of the idle epochs (shaded region).

plement as the set of the *idle tones* $\Omega_m^{\perp}$. Note that active tones actually correspond to slack non-negativity constraints and should not be confused with active constraints. Also, define the *noise rate* as $\eta_n(m) \triangleq \ln\left(WN_0/(Kg_n(m))\right)$ and refer to $r_n(m)$ as the *information rate*. Their sum, the information and noise rate (INR), $\zeta_n(m) \triangleq r_n(m) + \eta_n(m)$, allows us interpret the rate piecewise water-filling property simply as maintaining a constant $\zeta_n(m)$ over the active tones:

$$\zeta_n(m) = \begin{cases} \ln(\omega(m)) & \text{for } n \in \Omega_m \\ \ln\left(\frac{WN_0}{Kg_n(m)}\right) & \text{for } n \in \Omega_m^{\perp} \end{cases} \tag{4.1}$$

by noting that the optimal rate must obey the water-filling property of (3.30). Note that $\inf_n \zeta_n(m) = \ln(\omega(m))$ is achieved when $n \in \Omega_m$.

Recall that the goal of the offline prescient scheduler is to find the optimal rate profile that satisfies the Karush-Kuhn-Tucker (KKT) conditions, or equivalently, satisfies both the *rate bounding property* (Property 1), and the *piecewise rate water-filling property* (Property 3). The rate allocation over the idle tones $\Omega_m^{\perp}$ is zero as indicated by (3.30). For the rate allocation over the set of active tones Ω_m, it is more convenient to calculate the piecewise constant rate waterlevel $\ln(\omega(m))$ first, and use (4.1) to calculate the optimal rate $r_n(m)$.

4.2.1 Bounding the Information and Noise Rate

First, we denote the *average Information and Noise Rate* over an arbitrary subset of OFDM tones, Ψ_m as

$$Z^{\Psi_m}(m) \triangleq \frac{1}{|\Psi_m|} \sum_{n \in \Psi_m} \zeta_n(m) \tag{4.2}$$

and the cINR averaged over any arbitrary collection of OFDM tones for all symbols in the problem $\mathbf{\Psi} = \{\Psi_1 \ldots \Psi_M\}$ as

$$Z^{\mathbf{\Psi}}_{\text{cum}}(m) \triangleq \sum_{i=1}^{m} Z^{\Psi_i}(i) \tag{4.3}$$

where $|\Psi_m| = \text{size}(\Psi_m)$ is the number of non-idle channels in the set and the bold face $\mathbf{\Psi}$ denotes a sequence of sets over several epochs.

Recall that the traffic constraints are most conveniently represented as upper and lower bounds on the cumulative rates $R_{\text{cum}}(m)$ defined in Property 1. The quantity $Z^{\mathbf{\Psi}}_{\text{cum}}(m)$ is bounded in a similar way:

$$A_{\text{cum}}(m) + \sum_{i=1}^{m} \left(\frac{1}{|\Psi_m|} \sum_{n \in \Psi_m} \eta_n(i) \right) \leq Z^{\mathbf{\Psi}}_{\text{cum}}(m) \leq E_{\text{cum}}(m) + \sum_{i=1}^{m} \left(\frac{1}{|\Psi_m|} \sum_{n \in \Psi_m} \eta_n(i) \right) \tag{4.4}$$

for all $m = 1 \ldots M$. That is, the cINR is bounded above by the cumulative expiries and noise rate (cENR) and bounded below by the cumulative arrivals and noise rate (cANR) over any arbitrary subset of OFDM tones. Geometrically speaking, this bound combines the staircase bounds as shown in Figure 3.2 with an extra time-varying offset equal to the amount of *cumulative noise rate* added (see Figure 4.2 (b)).

Finally, we present an interesting property of the cINR averaged over the set of active tones, $\mathbf{\Psi} = \mathbf{\Omega} = \{\Omega_1 \ldots \Omega_M\}$. When we limit our consideration to the set of active tones, the average INR over the set of active tones is simply the rate waterlevel.

$$Z^{\mathbf{\Omega}}(m) = \ln(\omega(m)) \tag{4.5}$$

and the cINR over the set of active tones is

$$Z^{\mathbf{\Omega}}_{\text{cum}}(m) = \sum_{i=1}^{m} \ln(\omega(i)) \tag{4.6}$$

It should also be noted that (4.6) can be interpreted as the time integral of the positive rate water-level $\ln(\omega)$ so it is an increasing continuous, piecewise linear curve. This property can be formally stated as the shortest path property of the optimal cINR curve and it allows an efficient "string-pulling" algorithm for calculating this optimal *cINR*. Hence, by limiting our attention to only the set of active tones, we have the following shortest path property for the cINR curve.

Property 4 (Shortest Path Property). *The x-y plot of the cINR over the active tones, $Z_{cum}^{\Omega}(m)$ against m, must fall on the shortest path within the cANR and cENR bounding constraints as specified by* (4.4).

Proof. From (4.6), we note that $Z_{\text{cum}}^{\mathbf{\Omega}}(m)$ must be piecewise linear with the slope equal to the rate waterlevel $\ln(\omega(m))$. Furthermore, from the piecewise water-filling property, this piecewise linear curve bends upward (increases in slope) only if it touches the upper bound and bends downward (decreases in slope) only if it touches the lower bound. This precisely describes the geometric property of a taut piece of string pulled tight against pointwise bounding constraints and is known to form the shortest path. □

This shortest property for the example given in Figure 4.2 can be best visualized by physically cutting out the shaded region of the graph to remove idle epochs and pasting together the remaining pieces to form a continuous piecewise linear cINR curve as seen in Figure 4.3. Thus, the "string-pulling" algorithm [37], which finds the shortest constrained path by simulating pulling a piece of string through a set of pegs representing the constraints, can be used to compute the optimal rate waterlevel $\ln(\omega(m))$ by first finding $Z_{\text{cum}}^{\mathbf{\Omega}}(m)$ and then solving (4.6) for the optimal $\ln(\omega(m))$ *provided that the active tones Ω_m are known.* Next, we will provide a method for finding Ω_m through a method of elimination.

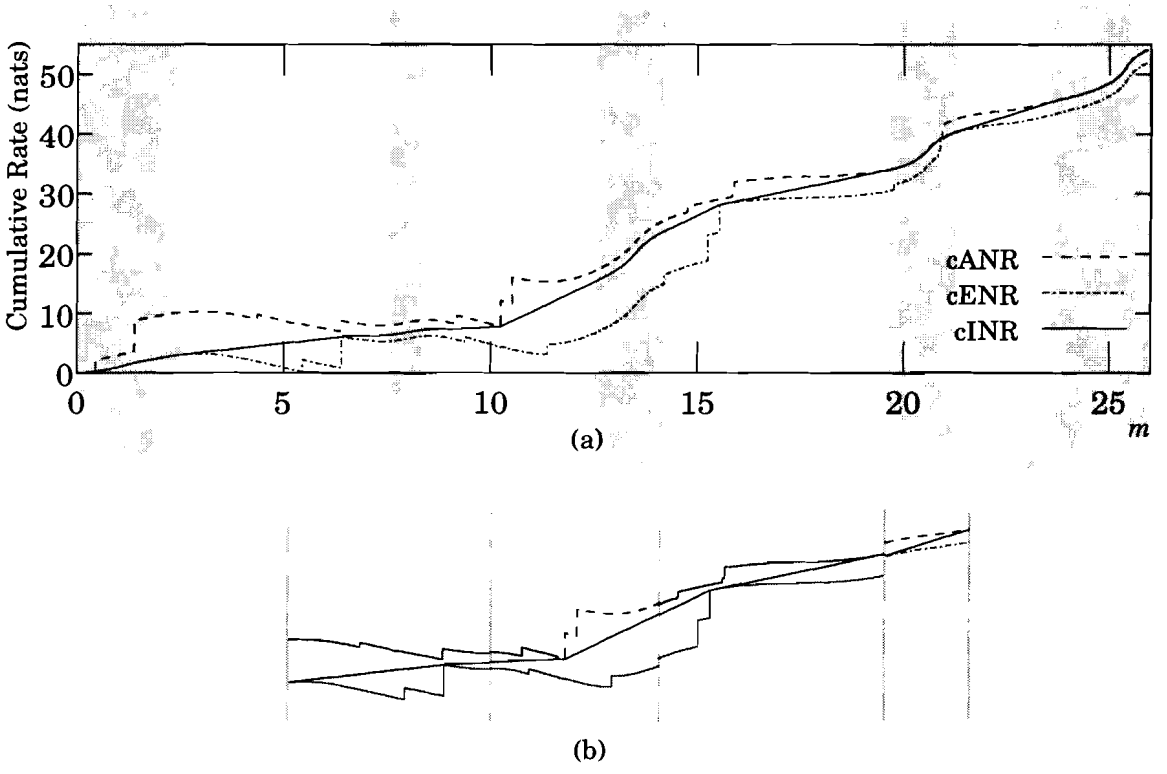

Figure 4.3: Graphs showing the cANR, cENR and cINR curves in (a). The shortest path property can be best visualized by piecing together consecutive segments of active tones as shown in (b).

4.2.2 Iteratively Finding the Active Tones

Firstly, we note that the "string-pulling" algorithm is a general method to minimize the geometric length of the $Z_{\text{cum}}^{\mathbf{\Omega}}(m)$ trajectory in Cartesian coordinates under constraint (4.4) over some arbitrarily selected sequence $\mathbf{\Psi}$ of tone sets. However, if idle tones are included, this solution will make their rates negative or zero. Since we do not know *a priori* which tones should be idle, we can remove them over a few iterations, a process referred to as puncturing. The iterations can be described with the following steps. In the first iteration, we assume that all tones are active and initialize $\mathbf{\Psi}$ to contain all tones at all times. Equivalent, we write $\mathbf{\Psi} = \mathbf{\Omega} \cup \mathbf{\Omega}^{\perp}$; that is, the initial guess at the idle set is empty. Next, apply the string-pulling algorithm to find the shortest path and hence $Z^{\mathbf{\Psi}^{[1]}}(m)$. From this $Z^{\mathbf{\Psi}^{[1]}}(m)$, we have the INRs $\zeta_n(m)$ in the slopes of $Z^{\mathbf{\Psi}^{[1]}}(m) = Z_{\text{cum}}^{\mathbf{\Psi}^{[1]}}(m) - Z_{\text{cum}}^{\mathbf{\Psi}^{[1]}}(m-1)$. Finally, subtract the noise rates, $\eta_n(m)$ from $\zeta_n(m)$, to obtain the information rate, $r_n(m)$, which concludes the first iteratio n. If all rates are positive, we have the optimal solution and the algorithm terminates. Otherwise, we assign tones with non-positive rates to the idle sets, $\Omega^{\perp[2]}(m)$ in iteration two, leaving a reduced $\mathbf{\Psi}^{[2]}(m) = \mathbf{\Psi}^{[1]}(m) \setminus \mathbf{\Omega}^{\perp[2]}(m)$. The shortest path calculation and idle tone removal is repeated until there are no non-positive tones.

To prove that the iterations will converge to the optimal solution, we must show firstly, that tones removed from consideration because they violate the non-negativity constraint are indeed idle in the final optimal solution, i.e, $\eta_n(m) \geq Z^{[i]}(m) \Rightarrow n \in \Omega_m^{\perp}$, and secondly, that the iteration will terminate within finite number of iterations.

Proof. Firstly, note that all interim waterlevel, $Z^{[i]}(m)$, satisfies the rate bounding constraints, since the cumulative rates fall between the cumulative arrival and expiry constraints. This interim waterlevel is infeasible only if one or more rate allocations are non positive. In the next iteration, the algorithm compensates by removing all

tones with non-positive rates and moving them to the idle set. This process effectively reduces the overall waterlevel since the rate allocation in other symbols can be lower, since they no longer need to compensate for the negative rate. Thus, the waterlevel profile calculated in earlier iterations can never be lower than ones calculated in later iterations.

Thus, the interim waterlevel, $Z^{[i]}(m)$, is a non-increasing sequence; that is $Z^{[1]}(m) \geq Z^{[2]}(m) \geq \cdots \geq Z^{\Omega}(m)$. As a consequence, if for a given iteration i, we have the n-th tone identified as being idle, i.e., $Z^{[i]} \leq \eta_n(m)$, it follows that $\eta_n(m) \geq Z^{\Omega}(m)$ and the tone is idle in all subsequent iterations, including the final one, which is the optimum solution.

Furthermore, since the algorithm only repeats when there is at least one negative-rate channel remaining, and since the total number of channels is finite, the algorithm terminates within a finite number of iterations. □

4.3 Simulation Result

We conclude the description of the ISP algorithm with a step by step walkthrough with a single channel example. System simulation results will be presented at the end of Chapter 5 after the discussion on the implementation of causal schedulers. The traffic constraints and the channel gains for the example are shown in Figure 4.4 (a) and (b). Initially, we do not know any idle epochs, thus the string-pulling procedure is performed with the cANR and cENR constraints calculated from the all epochs (Figure 4.4 (c)). The interim calculation of the cINR indicated by the piecewise straight line in Figure 4.4 (c) can be transformed back to the cumulative rate domain by subtracting the noise rate offset. However, since the idle epochs are not yet properly removed, the resulting rate profile is not strictly positive and an idle epoch is identified.

For the example given, the negative rate allocation is clearly visible from the water-filling diagram, Figure 4.4 (b).

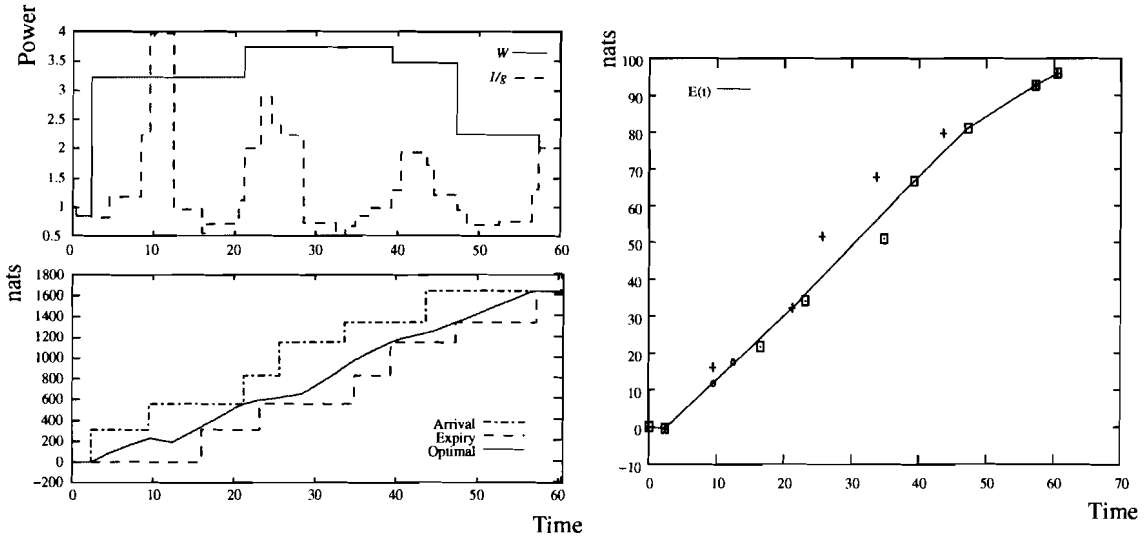

Figure 4.4: The water-filling diagram and the cINR plot for the first iteration of the ISP algorithm.

To prepare for the second iteration, the idle epochs are removed and a new set of cANR and cENR constraints are formed as shown in Figure 4.5. This can be seems as physically cutting out the unwanted epochs and pasting the remaining pieces together. The same transform → string-pulling → inverse transform steps are performed with these newly punctured set of constraints to find a better approximation to the optimal solution. At the end of this second iteration, there are no more idle epochs being identified for the example given, thus the algorithm terminates and the resulting rate profile is the optimal rate allocation for the active epochs while the identified idle epochs must take the rate of zero.

Although the number of iterations can potentially be as many as the number of epochs, making the upper bound on the overall computational complexity $O(n^3)$, that is n iterations of the string-pulling procedure at $O(n^2)$ complexity, it was observed

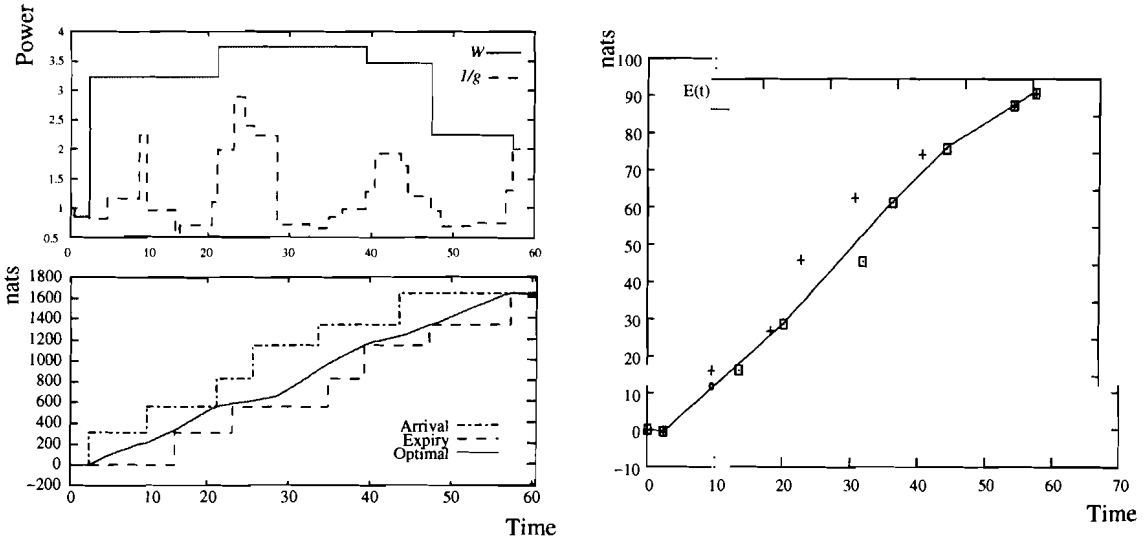

Figure 4.5: The water-filling diagram and the cINR plot for the second (final) iteration of the ISP algorithm.

through simulation that the actual number of iterations required to find all idle epochs remains relatively constant irrespective of the number of epochs in any given simulation. This is due to the fact that, statistically speaking, epochs with a low inverse channel gain (floor of water-filling) relative to the actual waterlevel have a higher probability of being identified with a small number of iterations, while an epoch with inverse channel gain close to the actual waterlevel may require a few iterations before it "rises from the water". Thus the number of iterations required is actually dependent on the statistical distribution of the channel gains and is independent of the total number of epochs being simulated. Hence, the computational complexity of the proposed algorithm is $O(n^2)$.

Chapter 5

Single User Online Scheduling Algorithm

Any real world scheduler must operate without the perfect knowledge of future arrivals and future channel conditions enjoyed by the prescient scheduler studied in Chapter 3 and 4. In this chapter, we present several causal modifications to the optimization problem to obtain practically implementable schedulers and compare their performances. Some of these causal adaptations have been previously published by the author in [37, 41].

The emphasis on the design of the causal scheduler is to provide optimization formulations that are solvable using only causal knowledge and provides comparable performances to the prescient formulation. This chapter is organized as follows, In Section 5.1, we present a sequence of causal adaptations ranging from the most computationally intensive to ones that are practical and easily implementable while remaining optimal in some restricted sense. There are several ways of dealing with uncertainty in the future and we have chosen a relaxed robust formulation that guarantees the feasibility of the rate allocation while not being too restrictive as be un-

solvable. More details of this formulation can be found in Section 5.1.1. All of the schedulers presented here can be solved using the iterative string pulling (ISP) algorithm used for the prescient scheduler and, in certain special cases, only one iteration of the inner string-pulling procedure is required.

In Section 5.3, the performance of some of these online schedulers are compared with that of the prescient scheduler and it was shown that under correlated Rayleigh fading conditions, the channel prediction scheduler can perform to within 3 dB of the prescient scheduler in single carrier channels and even better with diversity in multi-carrier channels.

5.1 Optimal Online Scheduler

In this section, we introduce several scheduler formulations that provide optimal solutions in some statistical sense. Firstly, we describe the common requirement of causality and the limitation it placed on the optimization problem formulation. We then present three variations on the problem formulations, two of which will be developed into practical algorithms for further simulation later in this chapter.

Firstly, a scheduler is practically implementable if the optimal transmission rates $\{r_n(m) \ \forall \ n\}$ for epoch m, can be computed at the beginning of epoch m using channel and traffic information known to the transmitter at the time. This is the optimal time for making the decision, for any earlier, we risk throwing away valuable channel and traffic information for making the rate assignment while any later violates the causality requirement. Obviously, the performance of any causal scheduler is upper bounded by that of the prescient scheduler. This bound is tight in the sense that under some specific circumstances, the causal scheduler can attain the performance of the prescient version. The performance of the prescient scheduler is only attainable

if 1. the packet arrival and expiry time as well as packet length are deterministic, and 2. the channel gains for the future are perfectly predictable. Allowing either traffic or future channel states to be random causes the performance of any causal scheduler to be worse then that of the prescient scheduler.

Thus, for all the causal schedulers considered in this section, we will concentrate on describing the problem of finding an *optimal* rate $r_n(m)$ for epoch m only. A similar problem with updated channel and traffic information will be used for the next epoch to calculate $r_n(m+1)$. We also assume that the channel gains for all past and present epochs are known perfectly. That is, $g_n(1) \ldots g_n(m)$ are known perfectly at the beginning of epoch m. By assuming that the current channel gain is known, it is possible to convert rate allocation to power allocation without considering the probability of outage. In practice, only noisy channel estimates can be obtained and one would use a slight underestimate of the channel gain to ensure that the signal strength is high enough to ensure an acceptable probability of outage. We will assume that these values are known perfectly, as the main emphasis here is to provide a causal formulation and the inclusion of these other practical concerns will only distract us from the main issue at hand.

Next, we turn our attention to the traffic constraints. It is easier to impose causality on the packet arrival process and then use it to determine which portion of the traffic constraints is known. Specifically, consider the arrival and expiry of the packets as shown in Figure 5.1, taken from the example used in Chapter 3. At the beginning of epoch 11, triggered by a channel state change not shown here, there have been six packet arrivals with two past expiries. Hence, there are potentially four known expiries that are in the future, and part of the cumulative expiry bounds can be determined. To distinguish this partial cumulative expiries bound from the actual cumulative expiries bound, we denote it as $\tilde{E}_{\text{cum}}^{[m]}(\kappa)$ where the superscript $[m]$ indicates

that it is computed from causal information at the beginning of epoch m and we will use κ to denote the time index relative to the current time. It should also be noted that this partially known constraint boundary is only a lower bound of the actual constraint for the prescient case, where the difference is due to the unknown arrivals and expiries due to packets arriving in the future. In addition to these past observations, $\boldsymbol{g}_{\text{past}}^{[m]} = g_n(1)\ldots g_n(m)$ and $\tilde{E}_{\text{cum}}^{[m]}(\kappa)$, we also need to specify the current arrival constraints which is simply a constant equal to the current queue length and is also equal to the last entry of $\tilde{E}_{\text{cum}}^{[m]}(\kappa)$. Hence, the past is fully specified by $\left\{\boldsymbol{g}_{\text{past}}^{[m]}, \tilde{E}_{\text{cum}}^{[m]}(\kappa)\right\}$. Also, for convenience, we denote the current queue length as $Q^{[m]} = \sum_{i=1}^{m} D(i) - \sum_{i=1}^{m-1} R(i)$ assuming all unity bandwidth epochs, $x_i = 1$.

Next, we provide a description of the notation required to specify the future. Lets denote the uncertain future channel gains as a random vector $\boldsymbol{g}_{\text{future}}^{[m]} = g_n(m+1)\ldots g_n(m+k)$ where k is chosen to be large enough to cover symbols up to the expiry time instant of the last packet in the queue. Also, denote the future portion of the cumulative arrivals and expiries, i.e. the part contributed by packet not yet arrived at symbol m, as the random vector $A_{\text{cum}}^{[m]}(\kappa)$ and $E_{\text{cum}}^{[m]}(\kappa)$ for $\kappa = 1\ldots k$. That is, a rate allocation is admissible given this future state if the cumulative rate profile satisfies

$$Q^{[m]} + A_{\text{cum}}^{[m]}(\kappa) \leq \sum_{\kappa=1}^{k} r(\kappa) \leq \tilde{E}_{\text{cum}}^{[m]}(\kappa) + E_{\text{cum}}^{[m]}(\kappa) \tag{5.1}$$

A graphical illustration of the relationship of these various quantities is shown in Figure 5.2.

5.1.1 Optimal Casual Scheduler with Conditional Future Averages

In this section, we discuss the meaning of optimality under uncertainty of system parameters. Here, we consider the evolution of the system states for the next k epochs

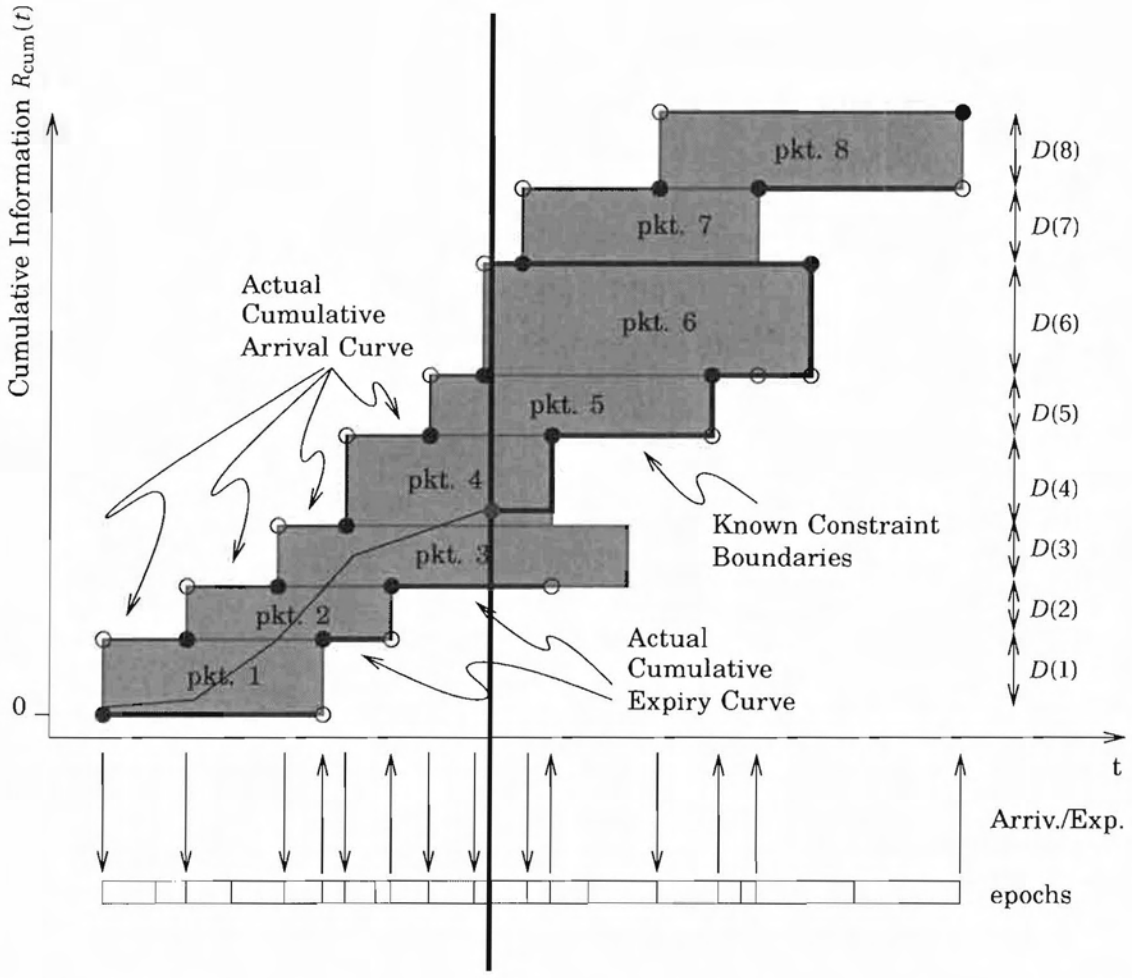

Figure 5.1: Diagram showing the traffic constraints known to the causal schedulers at the time instant as indicated by the vertical cursor.

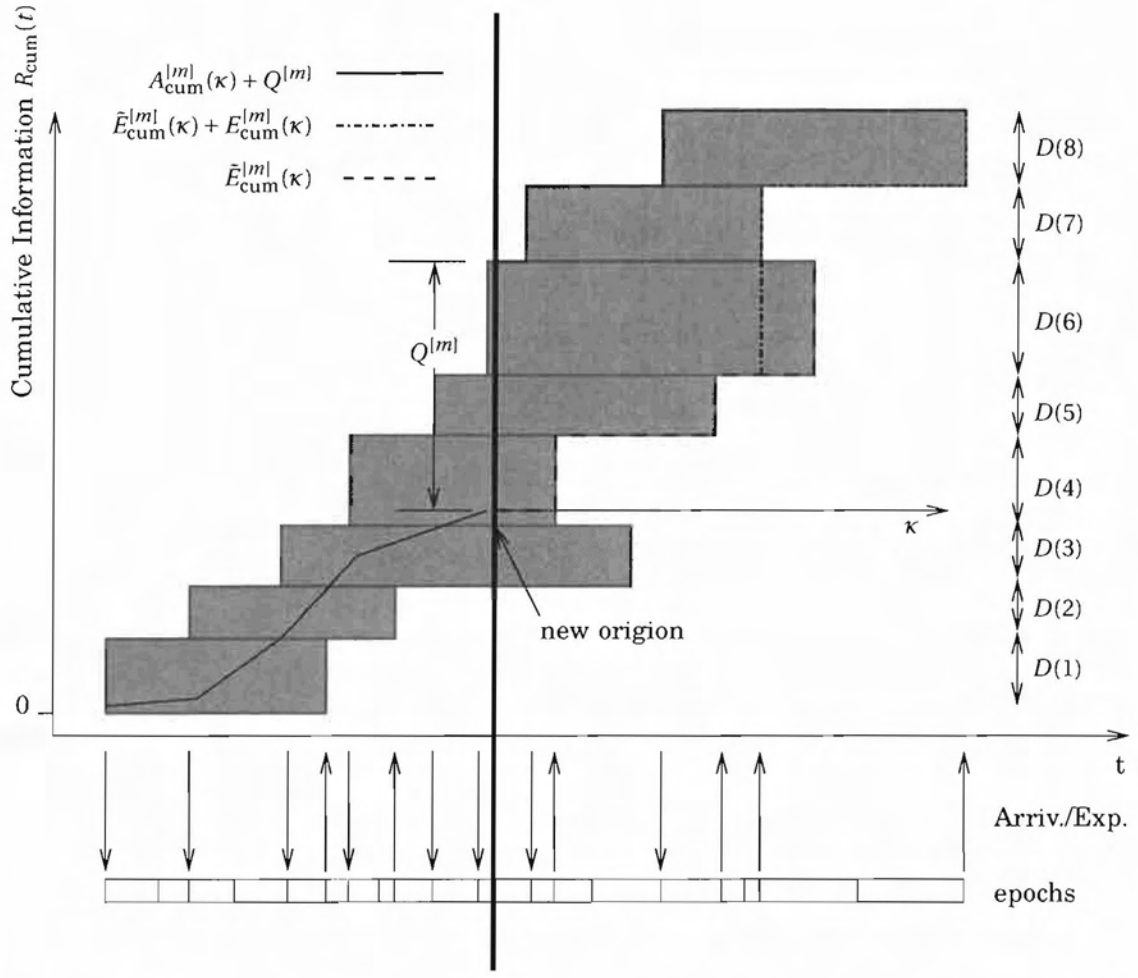

Figure 5.2: Diagram showing the definition of the actual and causal expiry and arrival curves with reference to the current time indicated by the vertical cursor.

and all quantities are indexed relative to the current time, m, with the dummy variable κ.

Denote the ensemble of all possible futures as $\left\{\boldsymbol{g}_{\text{future}}^{[m]}, A_{\text{cum}}^{[m]}(\kappa), E_{\text{cum}}^{[m]}(\kappa)\right\} \in \mathscr{F}$. Firstly, we need an expression denoting the conditional probability of any specific future occurring given an observed past. Since the channel gain evolves independently of the packet arrival process, we denote the probability that the channel evolves according to a given trajectory $\boldsymbol{g}_{\text{future}}^{[m]}$ conditional on the past observations, $\boldsymbol{g}_{\text{past}}^{[m]}$, as $\Pr\left(\boldsymbol{g}_{\text{future}}^{[m]} \middle| \boldsymbol{g}_{\text{past}}^{[m]}\right)$. Also, by assuming memoryless arrival and expiries, the probability that a given traffic constraints $T^{[m]} = \left[A_{\text{cum}}^{[m]}(\kappa), E_{\text{cum}}^{[m]}(\kappa)\right]$ occurring due to new packet arrivals is assumed to be independent of the past and can be expressed as $\Pr\left(T^{[m]}\right)$. It is also possible to formulate the causal scheduling problem with more general arrival and expiry distributions by replacing the unconditional probability of future traffic with an conditional one. However, the memoryless assumption is often used in places where the traffic model is not known, as it represents the worst case scenario.

Given the above notation, we can now describe the process in obtaining a statistically optimal rate allocation. For demonstration purposes only, consider the case where there is only a discrete number of possible futures. Specifically, assume that the evolution of future channel gains are only allowed to follow one of three possible trajectories $\boldsymbol{g}_1$, $\boldsymbol{g}_2$, $\boldsymbol{g}_3$ with probability $\Pr(\boldsymbol{g}_1)$, $\Pr(\boldsymbol{g}_2)$ and $\Pr(\boldsymbol{g}_3)$. We further assume that there are two possible future traffic states, $\boldsymbol{T}_1, \boldsymbol{T}_2$ with probability of occurring being $\Pr(\boldsymbol{T}_1)$ and $\Pr(\boldsymbol{T}_2)$ respectively. One obvious but suboptimal brute force approach to deal with uncertainty would be to calculate the optimal rate allocation using the prescient formulation for all six possible combinations and assign the actual transmission rate as a weighted average (weighted by the respective probability) of the calculated rates. However, this is not necessarily optimal due to the fact that the average of a function is not the same as the function of the average unless the function

is linear. And in our case, the functional map from the parameter space of channel gains and traffic constraints to the optimal rate is not linear.

The correct approach to the probabilistic optimization problem is to minimize the expected value of the objective over the future ensemble, while attempting to meet all of the possible constraints. That is, using the example with three channel states and two traffic states, the optimization problem can be formulated as

$$\text{Minimize:} \quad \sum_{n=1}^{3} f(\boldsymbol{r}, \boldsymbol{g}_n) \Pr(\boldsymbol{g}_n) \tag{5.2a}$$

$$\text{Subject to:} \quad \sum_{\kappa=1}^{k} r(\kappa) \geq Q^{[m]} + A_{\text{cum}}^{[m]}(\kappa) \qquad \text{for} \quad \left\{A_{\text{cum}}^{[m]}(\kappa), E_{\text{cum}}^{[m]}(\kappa)\right\} \in \{\boldsymbol{T}_1, \boldsymbol{T}_2\} \tag{5.2b}$$

$$\sum_{\kappa=1}^{k} r(\kappa) \leq \tilde{E}_{\text{cum}}^{[m]}(\kappa) + E_{\text{cum}}^{[m]}(\kappa) \qquad \text{for} \quad \left\{A_{\text{cum}}^{[m]}(\kappa), E_{\text{cum}}^{[m]}(\kappa)\right\} \in \{\boldsymbol{T}_1, \boldsymbol{T}_2\} \tag{5.2c}$$

$$\boldsymbol{r} \geq \boldsymbol{0} \tag{5.2d}$$

assuming non-zero probability probability for each state. Note that this formulation is not dependent on the probability distribution of the future traffic states $Pr(\boldsymbol{T}_i)$. In order for the solution to be robust, it must be admissible under *all possible* future traffic states irrespective of how often each traffic states is likely to occur.

We note that this problem remains a convex optimization problem with the objective dependent only on the statistics of the channel gain while the constraints are deterministic and covers all *possible* future traffic states. The fact that the channel random process is independent from the traffic random process also helps in the reduction of the problem by allowing the objective and the constraints to be considered separately. Furthermore, the only unknown to be determined in this formulation is the joint conditional probability distribution of the channel gain vectors.

This formulation is theoretically appealing in that it optimizes for all possible eventualities and the weighted objective can be seen as minimizing the mean error in energy cost due to uncertain future, although there are practical difficulties in

obtaining and working with the actual joint probability distribution of the channel gain vectors. This line of development was not pursued. Instead, as it can be seen in the following sections, that much less computationally intensive methods that approaches the performance of the prescient scheduler to within 3-dB in single-carrier systems and less than one dB for multi-carriers were obtained. This leaves very little room for improvement by taking this more complex approach.

5.1.2 Optimal Causal Scheduler with MMSE Channel Prediction

Here we present a practical implementation of the causal scheduler based on minimizing the energy usage assuming the most likely channel gain trajectory. Under this formulation, we have removed the need to optimize a weighted sum of energy usages by considering only the most likely path. Next, we present the exact formulation of the objective to minimize followed by the simplification of the constraints.

The prescient objective of minimizing total energy consumption, (3.20a), is dependent on the future channel gains, which are partially unknown at the beginning of symbol m, so it is not well defined for optimization. Hence, we use a statistical approximation $\hat{g}_n^{[m]}(i) \simeq g_n(m+i)$ at symbol m of the unknown future channel gains. That is, instead of an objective that averages over several scenarios, we will use only one single representative scenario. Since the power and rate allocation for the past can no longer be altered, we minimize the *future* energy usage under estimated channels for symbols m up to $m+k$ only. While this allows us to calculate the rate allocation for the next k epochs, we are only interested in the first one. Denote the transmission rate for symbol $m+\kappa$ calculated at symbol time m as $r_n^{[m]}(\kappa)$, and the *future* energy usage can be expressed as

$$\sum_{\kappa=0}^{k} \sum_{n=1}^{N} \frac{WN_0\left(e^{r_n^{[m]}(\kappa)}-1\right)}{K\hat{g}_n^{[m]}(\kappa)} \tag{5.3}$$

where $\hat{g}_n^{[m]}(0) = g_n(m)$ is the channel gain of the current symbol, which is assumed to be known perfectly. By assuming that the channel gain of the *current* symbol is known, we can assign its rates and powers without incurring the possibility of channel outage. Note that outage was not a problem in the prescient formulation since the channel is assumed to be known perfectly for all symbols *a priori*.

To provide a concrete example for channel prediction, we assume that the channels undergo correlated Rayleigh fading with Jakes Doppler spectrum. In these circumstances, the maximum likelihood channel power gain predictor is based on a simple linear L-point minimum mean square error (MMSE) predictor for the complex channel gains. It is used to estimate the future power gains, $\hat{g}_n^{[m]}(\kappa)$, as the conditional means of $g_n(m+\kappa) = |h_n(m+\kappa)|^2$ given the past observation of complex channel gains, $h_n(m)$, $h_n(m-\delta)$, ... , $h_n(m-L\delta+\delta)$, where δ is selected to provide Nyquist sampling of the Doppler spectrum. The details of the MMSE predictor are given in Appendix A. For those readers interested in the accuracy of various channel prediction methods, see also [42].

Next, we find a deterministic set of constraints that will produce a robust solution. Then we can constrain the cumulative rate $\sum_{\kappa=0}^{k}\sum_{n=1}^{N} r_n^{[m]}(\kappa)$ in a similar way to the prescient problem by the cumulative arrivals and expiries. Assuming for the moment that there is no new packet arrival within the next k symbols, we can formulate a deterministic convex minimization problem as follows:

$$\text{Minimize:} \quad \sum_{\kappa=0}^{k}\sum_{n=1}^{N} \frac{WN_0(e^{r_n^{[m]}(\kappa)}-1)}{K\hat{g}_n^{[m]}(\kappa)} \tag{5.4a}$$

$$\text{Subject to:} \quad \sum_{\kappa=0}^{k}\sum_{n=1}^{N} r_n^{[m]}(\kappa) \le Q^{[m]} \tag{5.4b}$$

$$\sum_{i=0}^{\kappa}\sum_{n=1}^{N} r_n^{[m]}(i) \ge \sum_{i=0}^{\kappa} \tilde{E}^{[m]}(i) \qquad \text{for} \quad \kappa \in [0,k] \tag{5.4c}$$

$$r_n^{[m]}(k) \ge 0 \qquad \text{for} \quad n \in [1,N], \kappa \in [0,k] \tag{5.4d}$$

where the cumulative expiries can be derived from the expiries of the packets already in the queue and the cumulative arrivals is given by the current queue length $Q^{[m]} = \sum_{i=1}^{m} A(i) - \sum_{i=1}^{m-1} R(i)$ in nats. This problem closely resembles the prescient optimization (3.20), with termination time equal to the expiry time κ of the last packet in the queue, and it can be solved in the same way. In fact, in this formulation, the arrival constraint is always slack except at the end and can be used to simplify the ISP algorithm, as there is only one type of constraints to test against.

Also, note that the scheduler uses only the rate allocations $r_n^{[m]}(0)$ for the current symbol, and discards allocations for future symbols, since a new optimization is performed at every symbol time. Thus we can relax the robustness condition by requiring *only* $r_n^{[m]}(0)$ to meet all possible constraints rather than requiring *all* $r_n^{[m]}(\kappa)$ to meet all constraints for all $\kappa = 1 \ldots k$. Next, we show that $r_n^{[m]}(0)$ is robust by noting that no packet arriving on or after symbol m can expire at the end of symbol m. Thus $r_n^{[m]}(0)$ satisfies the expiry constraints of all possible scenarios. Furthermore, future arrivals can only relax the arrival constraint (5.4b) and thus any $r_n^{[m]}(0)$ satisfying (5.4b) must satisfy the constraints of all other possible arrival scenarios.

In summary, the robust scheduler must perform an optimization at the beginning of every symbol by constructing and solving (5.4). This system with only k symbols and a simplified set of constraints can be quickly solved using the ISP algorithm presented for the prescient scheduler and will result in an robustly optimal causal scheduler. Furthermore, a recalculation of current transmission rate, $r_n^{[m]}(0)$, only needs to occur when the channel state changes, or when there is a new packet arrival, instead of at every orthogonal frequency division multiplexing (OFDM) symbol. This represents a considerable reduction in processing requirement in a system with a slowly changing channel and long packet inter-arrival time.

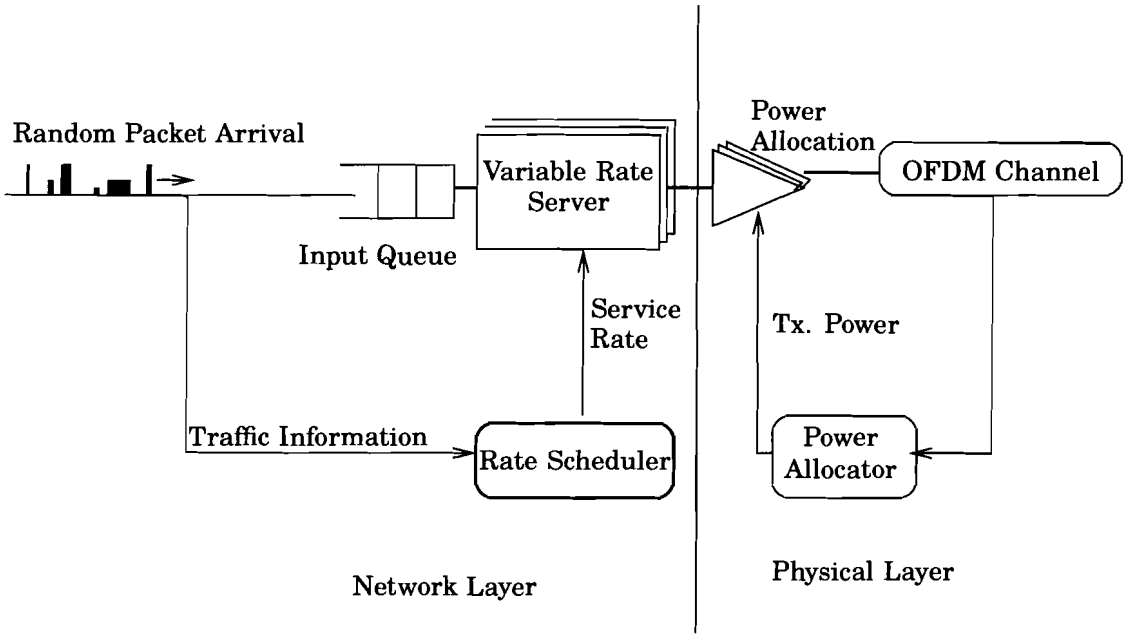

Figure 5.3: The system diagram for the layered scheduler. Note that the channel state information is not used in determining the service rate.

5.1.3 The OSI Layered Scheduler

In this section, we consider a different causal scheduler where we assume that the only knowledge about future channel gain is the ensemble average. That is, we assume $g_n^{[m]}(\kappa) = 1$. This allows two important simplifications to the online scheduling algorithm compared to the channel prediction scheduler presented earlier. Firstly, there is no need to perform channel predictions. And secondly, by assuming that the channel gains are constant, the ISP algorithm can be greatly simplified resulting in a linear complexity scheduler. It also represents the optimal scheduler under the layering constraint imposed by the open system interconnection (OSI) reference model. That is, the channel gain information is confined to be only known in the physical layer and is not available for making rate assignment decisions which is a function of higher layers. For the system diagram, see Figure 5.3.

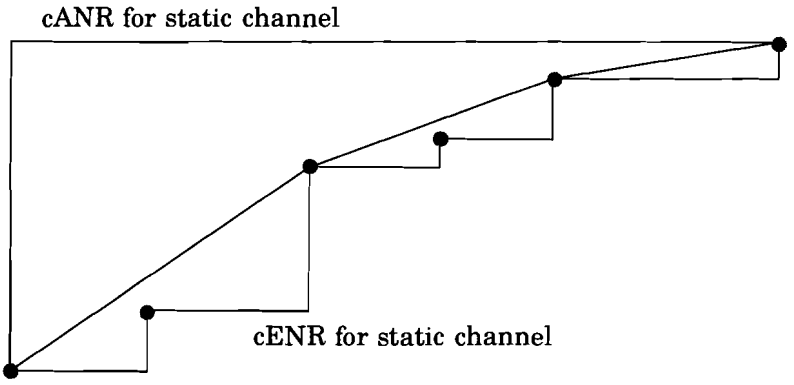

Figure 5.4: The shortest path under only cumulative expiry bounds must form the convex hull over the constraint point as shown in the diagram.

The formulation for this scheduler is

$$\text{Minimize:} \quad \sum_{k=0}^{\kappa} \sum_{n=1}^{N} (e^{r_n^{[m]}(k)} - 1) \tag{5.5a}$$

$$\text{Subject to:} \quad \sum_{i=0}^{\kappa} \sum_{n=1}^{N} r_n^{[m]}(i) \leq Q_m \tag{5.5b}$$

$$\sum_{i=0}^{k} \sum_{n=1}^{N} r_n^{[m]}(i) \geq \sum_{i=0}^{k} \hat{E}^{[m]}(i) \quad \text{for} \quad k \in [0, \kappa] \tag{5.5c}$$

$$r_n^{[m]}(k) \geq 0 \quad \text{for} \quad n \in [1, N], k \in [0, \kappa] \tag{5.5d}$$

where the objective is no longer a function of the time-varying channel gain and the robust traffic constraint remain the same as that of (5.4) and is indicated by the up-side-down staircase shaped region shown in Figure 5.1.

We also note that in this formulation, the ISP algorithm can be applied directly without the adding of the $\ln(g)$ offset as $g = 1$ in this particular formulation. We further observe that the taut string can only be constrained by the expiry staircase and forms the convex hull of the expiry bounds, as demonstrated by Figure 5.4, and the initial slope of this "taut string" is determined by the maximum of the slopes of

all line segments connecting the origin to all the expiry constraints

$$r_n(m) = \max_{\kappa} \left(\frac{E_{\text{cum}}^{\lfloor m \rfloor}(\kappa)}{\sum_{i=1}^{\kappa} x_i} \right) \tag{5.6}$$

for the variable length epoch formulation.

In this simplified model where the channel gain is assume constant, the string-pulling algorithm can be further simplified by checking less number of constraint points in line 7. This results in an $O(n)$ complexity as oppose to $O(n^2)$ for the variable channel case.

5.2 Qualitative Comparison of Online Schedulers

In this section, we provide some qualitative discussion on the performance of various online schedulers proposed. For comparison purposes, the system parameters used in the example given in Chapter 3.2.4 – Figure 3.4 is used.

Figure 5.5 shows the cumulative rate plot and the rate water-filling diagram of the simplest, layered scheduler described in Section 5.1.3. The diagram shows the cumulative rate and the cumulative arrival and expiry bounds (a) as well as the rate allocations for each of the two subcarriers as the shaded region. This scheduler calculates the optimal cumulative curve based only on the partially known expiry constraints and uses no channel state information in computing the rate profile. Thus, the cumulative rate curve as seen in Figure 5.5 (a) approximates a straight line. Furthermore, since the rate allocation only adapts to past traffic, it can be seen from Figure 5.5 (b) and (c) that a constant rate is transmitted in both good and bad channel states. Thus, we could predict that the power usage of this layered scheduler resembles that of a channel inversion power control system.

Next, we consider the optimal scheduler assuming a one tap channel predictor.

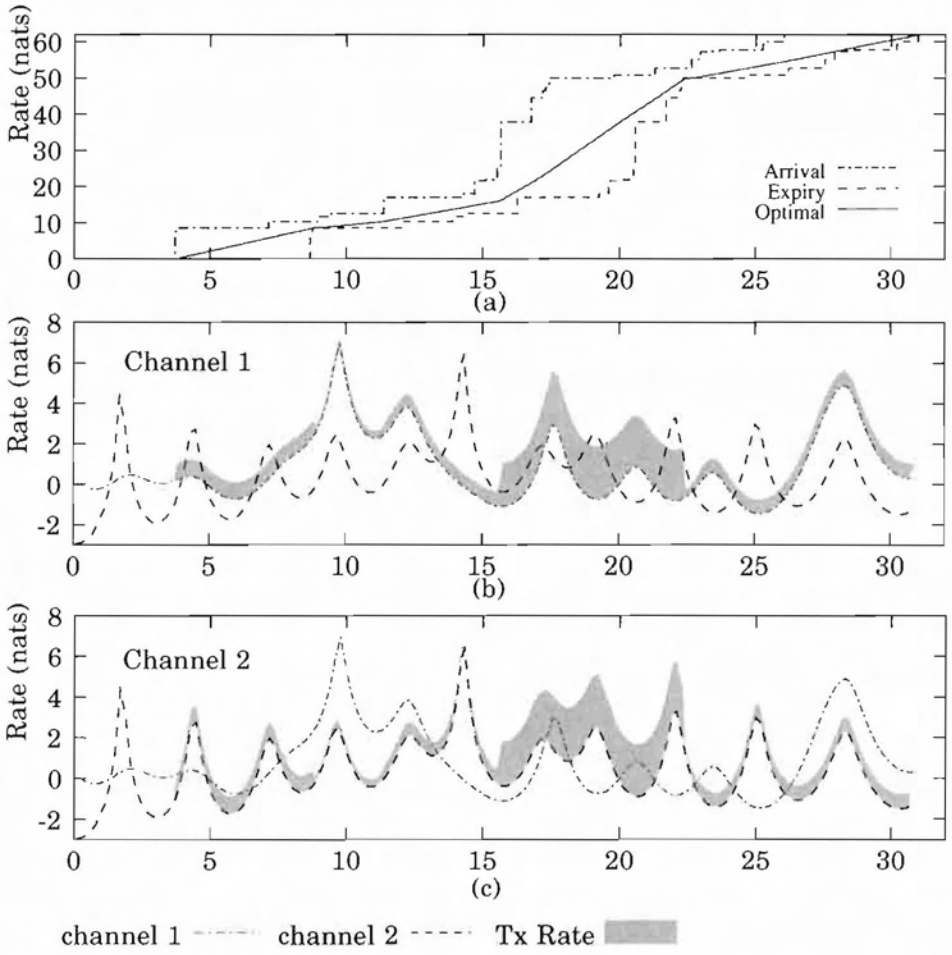

Figure 5.5: The cumulative rate plot and the rate water-filling diagram of the layered scheduler.

From the channel gain trajectories shown in Figure A.1 from Appendix A, it can be seen that the one tap channel predictor assumes a return to mean channel prediction and can be easily calculated from the current channel gains. The cumulative rate plot and the rate water-filling diagram for the one tap prediction scheduler are shown in Figure 5.6. Compared with the result shown in Figure 5.5, it is easy to see that the inclusion of even a very crude channel predictor in the rate allocation consideration allows the scheduler to avoid allocating rate in places where the channel is bad. This can be easily seen by comparing Figure 5.5 and Figure 5.6 in the region around $t = 10$ and 27. With knowledge of the channel in each of the subcarriers, the scheduler can now perform water-filling between these two subcarriers and it is easy to see that the channel prediction scheduler simply avoids using the channel with a bad channel gain.

Finally, Figure 5.7 shows the cumulative rate plot and the rate water-filling diagrams for the channel prediction scheduler with an 8-tap channel predictor. In the figure, it can be clearly seen from the water-filling diagrams that the rate allocation now closely resembles the one for the prescient scheduler as seen in Figure 3.4.

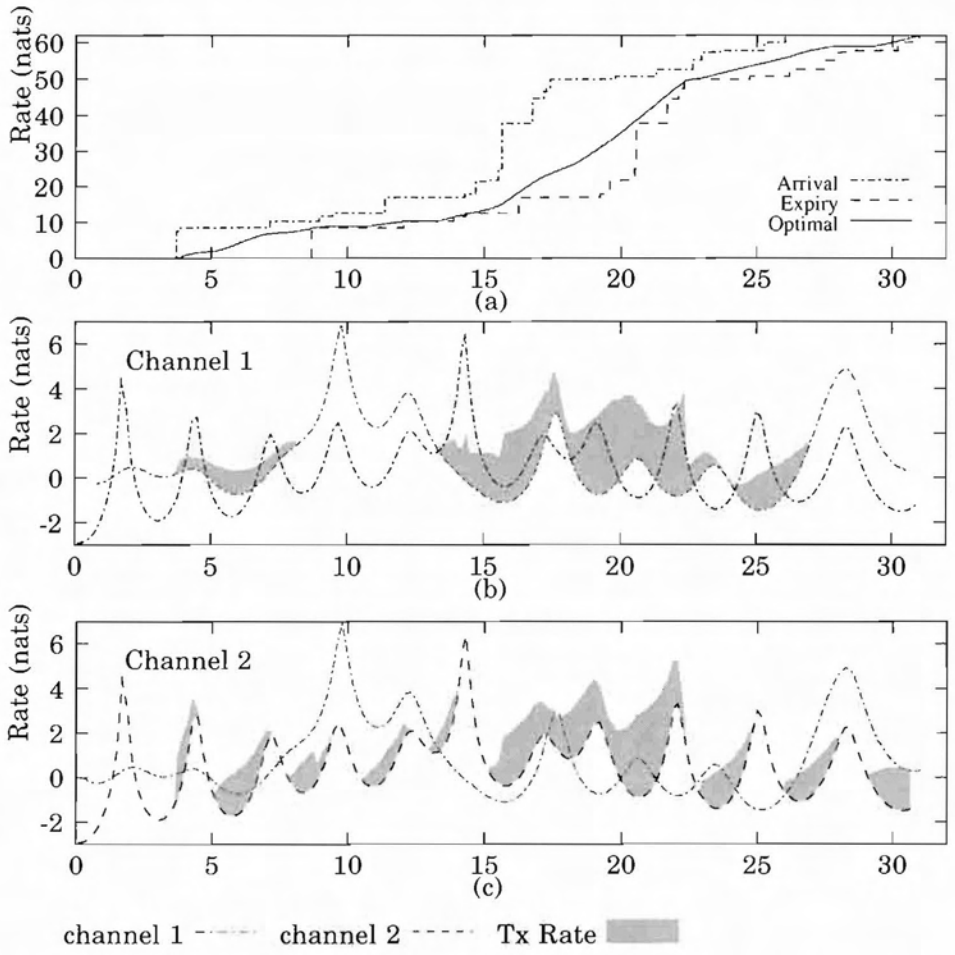

Figure 5.6: The cumulative rate plot and the rate water-filling diagram for the one tap prediction scheduler.

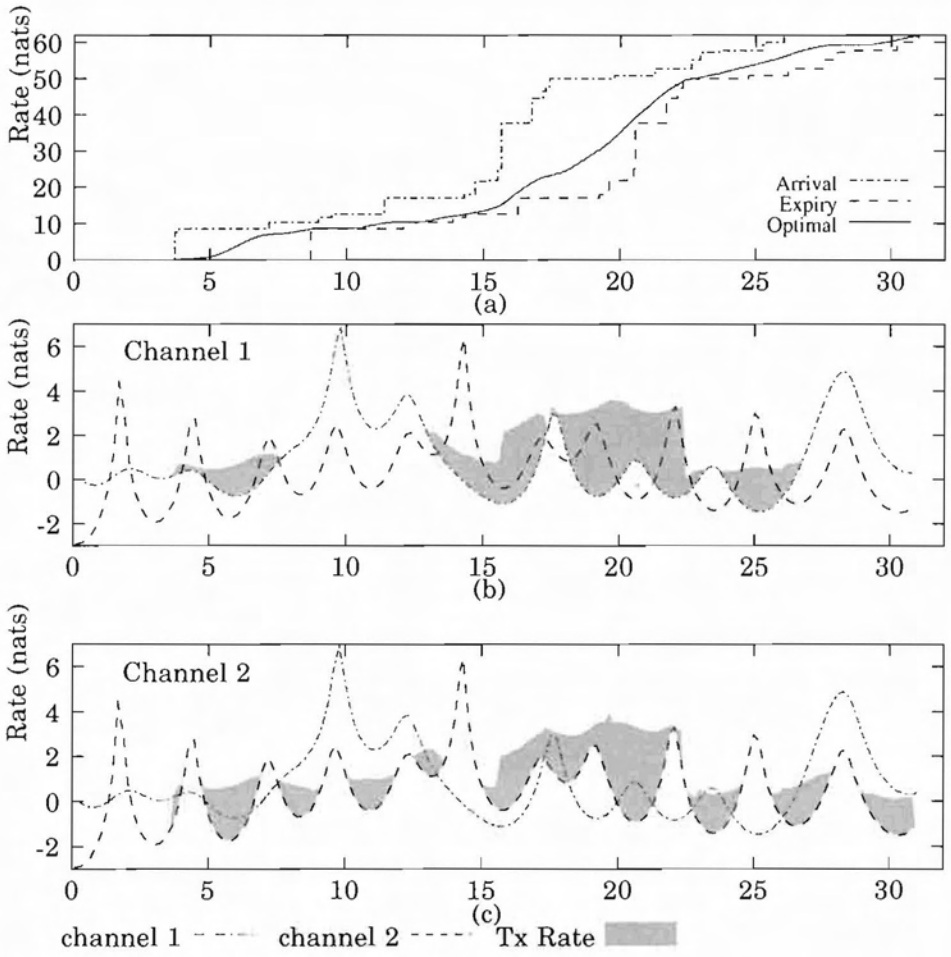

Figure 5.7: The cumulative rate plot and the rate water-filling diagram for the 8 tap prediction scheduler.

5.3 Performance Comparison

We have addressed minimum-energy scheduling for the multi-carrier channel with per-packet delay constraints, and have obtained efficient means of computing the optimal rate and power allocation in the form of the optimal prescient scheduler and the robustly optimal causal scheduler. How well do they work? In this section we compare the energy usage of these schedulers under Rayleigh fading channel and various traffic patterns and show that the performance of the causal scheduler approaches that of the prescient scheduler, especially as the number of carriers or the order of the predictor increase.

To demonstrate the value of joint traffic and channel optimization, we also simulate the layered scheduler, comprising a channel-unaware queue server and a traffic-unaware power control unit. The queue server calculates the rates $r_n^{[m]}(0)$ by optimizing (5.4) with the unconditional expected gains $g_n^{[m]}(0) = 1$, and the power control unit simply meets those rates by inverting (3.18), assuming knowledge of the current channel gains. With this layered structure, the resulting power profile contains large power spikes during deep fades, resulting in an near-infinite average power. This is analogous to the infinite average power of a channel inversion power allocation scheme in Rayleigh fading channels [43]. In order to compare it with the scheduler above, a truncated average power measure is used [43]. Specifically, we omit the 0.1% of bits with the greatest energy allocation from the average [41]. Note that it is the comparison scheduler that is being presented in a somewhat optimistic fashion, not the schedulers derived above. For the prescient and causal schedulers, there is no problem with these infinite power spikes during deep fades and the E_b usage is calculated over all symbols.

To make meaningful comparison of the above-mentioned schedulers, we need to compare their energy usages under identical traffic and channel conditions. In the

next section, we will present the traffic and channel model from which the traffic parameters $A(m)$, $E(m)$, $\hat{E}^{[m]}(\kappa)$, and the channel parameters $g_n(m)$ can be derived.

5.3.1 Normalized Traffic and Channel Model

For modeling the traffic, let the packet arrival be modelled as a Poisson process with exponentially distributed inter-arrival times t_{IA} with mean $\bar{t}_{\mathrm{IA}}$ (seconds), and the packet length θ be exponentially distributed with mean $\bar{\theta}$ (bits). Thus, the average throughput of the system is $\bar{\theta}/\bar{t}_{\mathrm{IA}}$ bits per second. For a total system bandwidth of NW Hz, the spectral efficiency ρ is

$$\rho = \frac{\bar{\theta}}{NW\bar{t}_{\mathrm{IA}}} \quad \text{bits/sec/Hz} \tag{5.7}$$

We also denote the maximum allowable packet delay as t_e.

For modeling the fading channel, the complex channel gains $h_n(m)$ are generated as correlated complex Gaussian processes that are correlated in time (m) and independent in frequency (n). For a typical OFDM system, independence in frequency implies significant separation of subcarriers, which is a somewhat unrealistic assumption. However, we are more interested in the diversity effect of multiple parallel channels than in the detailed implementation of an OFDM system, so in our simulation, we use only a small number of independent channels ($N = 1, 2$ and 8). As for the temporal correlation, we assume a Jakes spectrum [44] with a maximum Doppler frequency f_d Hz.

5.3.2 Defining the Simulation Scenarios

When comparing the performances of the schedulers, it is not possible to simulate for all combinations of the parameters $\{\bar{t}_{\mathrm{IA}}, W, \rho, N, f_d, t_e, N_0, K\}$ defined above. How-

ever, their number can be reduced by normalizing them in meaningful dimensionless groups. Firstly, instead of comparing the total energy usage of various schedulers, we will compare the average SNR per bit $\bar{\gamma}_b = \sum_{n,m} p_n(m)/(\rho W N_0 N M)$. Without loss of generality, we will also set $K = 1$ in the rate-power conversion (3.18), corresponding to the Shannon capacity. For systems with other values of K, we note that the required power usage is simply K^{-1} times more and does not affect the comparison result in dB. Finally, we normalize the remaining time-dependent quantities by the average interarrival time, leaving us with four independent parameters, $\{N, \rho, f_d \bar{t}_{\mathrm{IA}}, t_e/\bar{t}_{\mathrm{IA}}\}$.

To begin, we can distinguish two limiting cases in which the effects of channel variation and traffic demands become decoupled, and there is little benefit in performing joint optimization. As t_e becomes large, i.e., the maximum allowable delay for each packet becomes infinite, the optimal rate and power allocation must approach that of water-filling with a constant water level over time and frequency. On the other hand, when the maximum allowable delay becomes smaller than the interarrival time, the system will spent most of its time oscillating between idling and servicing only one packet. In this case, the performance of the system is dominated by the mark-to-space ratio, $t_e/(\bar{t}_{\mathrm{IA}} - t_e)$, of the queue state, and under a block fading assumption that the channel state remain roughly constant during the transmission of a packet, the power consumption must follow that of channel inversion.

For the rest of this section, we will present the performance comparisons with $t_e/\bar{t}_{IA}i$ in the range of 1/4 to 8 where there is a rich variability in the traffic constraints and the interaction with the Doppler frequency is important. As for the value of Doppler frequency, we will simulate at a slow Doppler, $f_d = 0$ and at a fast Doppler, $f_d = 0.05\bar{t}_{\mathrm{IA}}^{-1}$ which is chosen to be the fastest Doppler frequency under which a reasonable accuracy of channel prediction can still be obtained up to κ symbols into the future.

Next, we will compare the average signal to noise ratio (SNR), $\bar{\gamma}_b$, of the prescient

scheduler and the causal scheduler with various accuracies of channel prediction, and the truncated average SNR, $\bar{\gamma}_b$, for the layered scheduler under the two extreme ends of Doppler frequency ($f_d = 0$ and $f_d = 0.05\bar{t}_{\mathrm{IA}}^{-1}$), at $N = 1, 2$, and 8. For the causal scheduler with channel predictions, we use a channel sample spacing of $\delta = 0.3/f_d$. The comparison is initially performed at a spectral efficiency $\rho = 1$ which corresponds to the operating point of 0-dB SNR under static additive white Gaussian noise (AWGN) channel. Results for a single subcarrier at higher spectral efficiencies are also presented at the end of this chapter.

5.3.3 Performance in Static Flat Channel

Firstly, we consider the performance under the simplest, static flat fading channel where the only performance difference of the schedulers must be due to the different way of formulating the traffic constraints. The average SNR per bit for the various schedulers under static channel conditions ($f_d = 0$) are plotted against $t_e/\bar{t}_{\mathrm{IA}}$ in Figure 5.8. The performances of all the robust causal schedulers and the layered scheduler are identical for all values of N. This is because the channel is constant in time, so that the $\bar{\gamma}_b$ performance measure is independent of the channel prediction accuracy.

Also note that as $t_e/\bar{t}_{\mathrm{IA}}$ increases, the gap between the upper and lower staircase bounds on the cumulative rate widens and the occurance of active constraints became less frequent. As a consequence, the system performance approaches that of an unconstrained system. That is, at a spectral efficiency of $\rho = 1$, the required SNR per bit is 0 dB, which is the asymptotic value shown in Figure 5.8.

Despite the fact that the performance curves for all causal schedulers collapse into one, we see that there is only a very small, constant gap (0.5 dB) between the causal schedulers and the prescient scheduler. Since all schedulers can make perfect

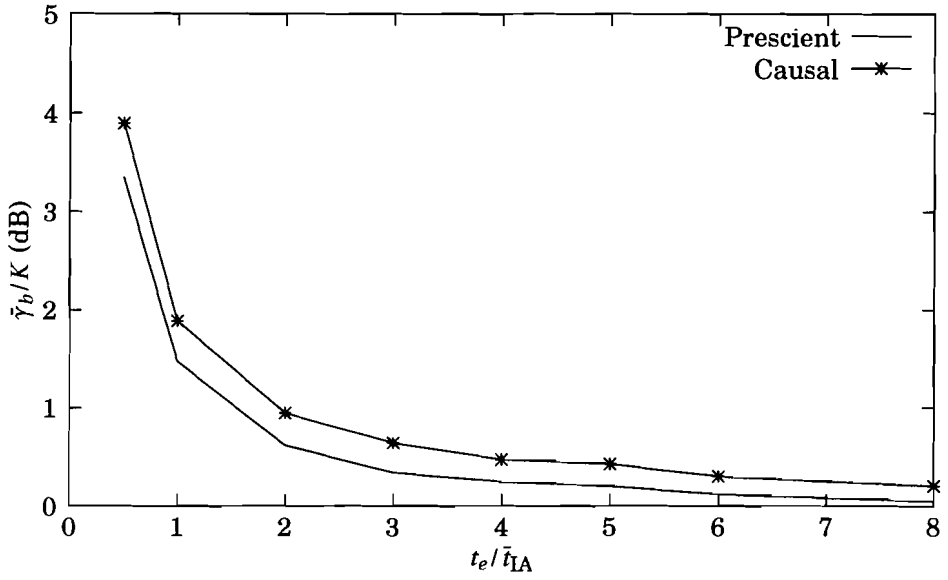

Figure 5.8: Comparison of the average SNR per bit ($\bar{\gamma}_b$) for the causal and prescient schedulers under static channel conditions.

prediction under this static condition, we can attribute this half dB difference to the "no future arrivals" assumption in the robust formulation of the causal schedulers, which make their rate allocation necessarily small at times.

5.3.4 Performance in Fading Channels

Figure 5.9 – 5.11 shows the performance comparison of the various schedulers under $N = 1, 2$, and 8 parallel channels respectively. The traffic load is normalized such that a system with higher total bandwidth is carrying proportionally more bits so that all systems maintain an average spectral efficiency of one bit per second per Hz. Firstly, we note that in the single channel case (Figure 5.9), as we increase the number of past channel samples used in the prediction, the performance of the causal schedulers improves and with a reasonably easy-to-implement 8-point predictor, we can achieve performance that is only 3 dB away from that of the prescient optimal.

Furthermore, as we increase the diversity by increasing the number (N) of sub-

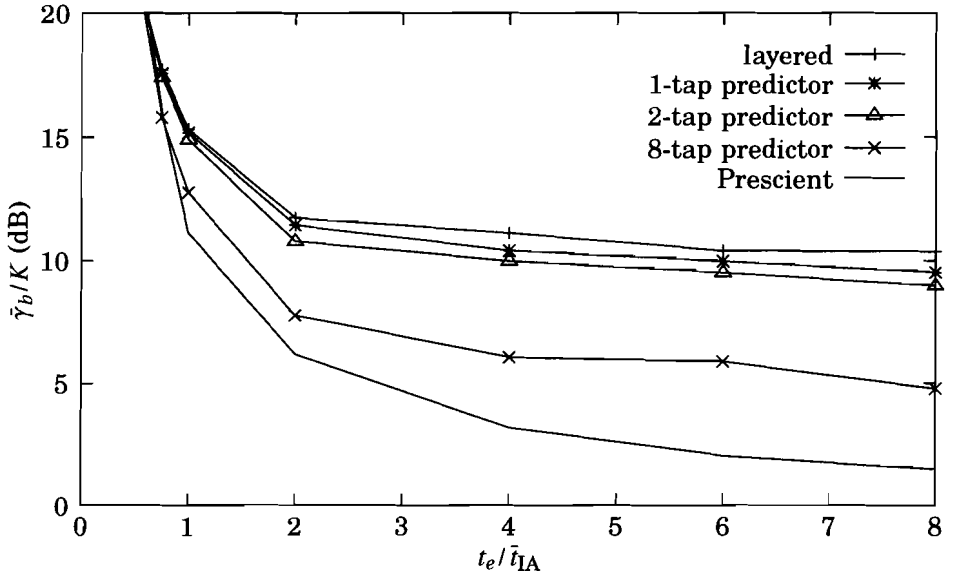

Figure 5.9: Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 1$ subcarriers at spectral efficiency $\rho = 1$ bits per second per Hz.

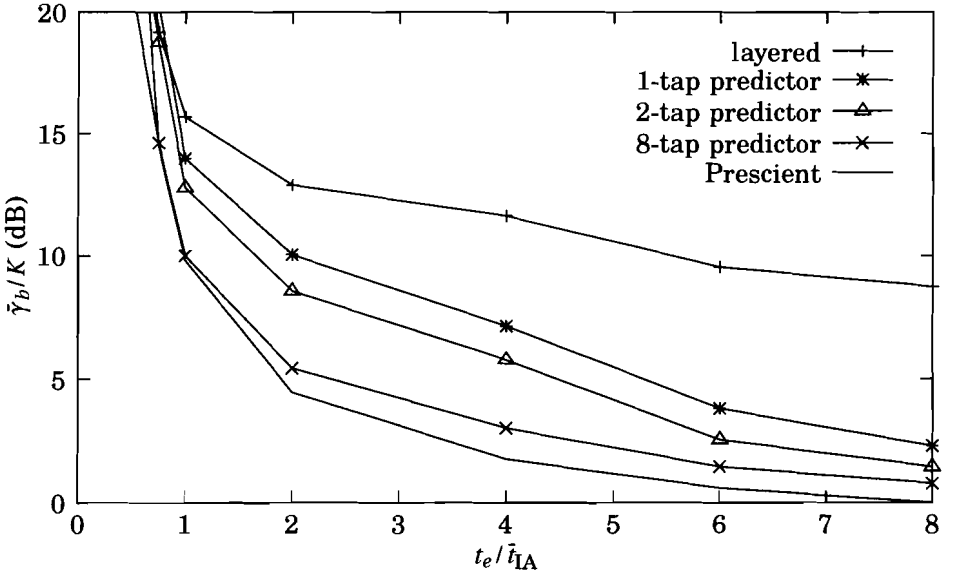

Figure 5.10: Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 2$ subcarriers at spectral efficiency $\rho = 1$ bits per second per Hz.

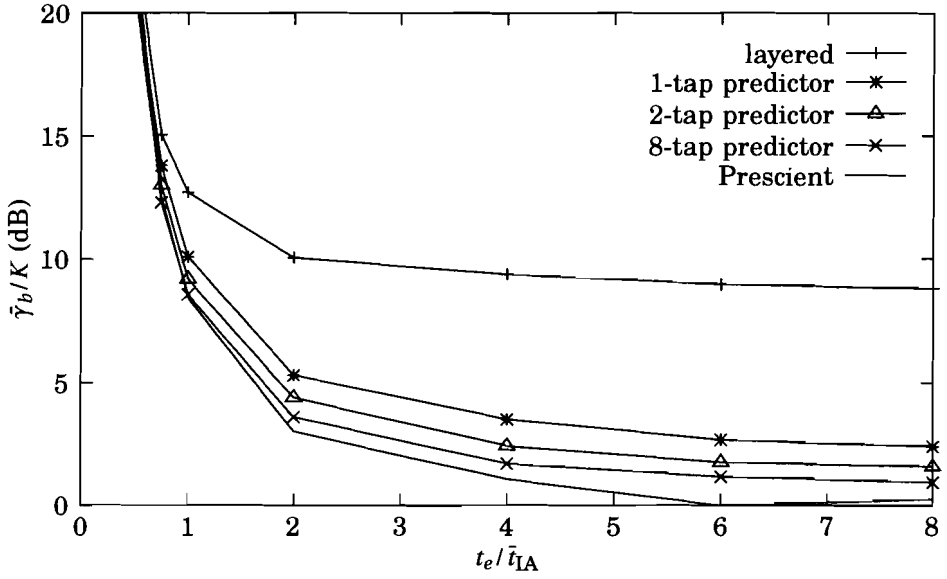

Figure 5.11: Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 8$ subcarriers at spectral efficiency $\rho = 1$ bits per second per Hz.

carriers, the performance of the causal scheduler with lower order channel predictors improves and starts to approach that of the prescient optimal. This performance gain can be explained in terms of channel hardening; as the number of parallel channel increases, the coefficient of variation of capacity per symbol decreases, and the accuracy of channel prediction becomes less important. The improvement is less dramatic but also significant for the 8-point prediction scheduler, which reduces from 3 dB away from the prescient optimal (Figure 5.9) for the single channel case, to just a little more than 1 dB for the case with eight-fold diversity (Figure 5.11).

On the other end of the spectrum, the layered scheduler that assigns transmission rate without considering the channel gain, is consistently the worst of the group. This illustrates the value of cross-layer optimization. It is interesting that, as $t_e/\bar{t}_{\mathrm{IA}}$ approaches 0, that is, as the packets are set to be transmitted with a very short delay constraint, all optimizing schedulers, prescient and causal, approach the performance of this layered scheduler, as predicted previously.

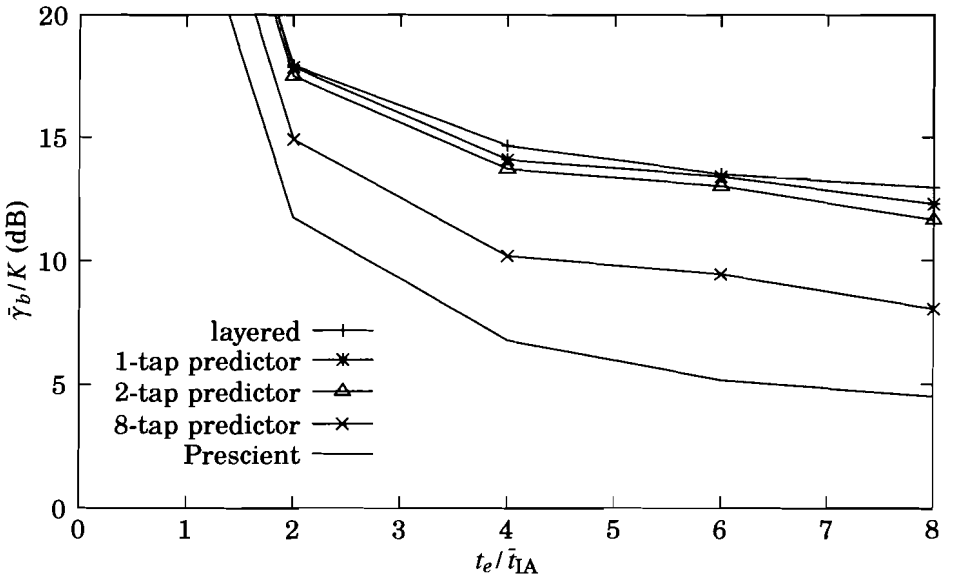

Figure 5.12: Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 1$ subcarriers at spectral efficiency $\rho = 2$ bits per second per Hz.

Finally, we present the comparison of the required average SNR per bit under a spectral efficiency $\rho = 2$ bits per second per Hz and $N = 1$ subcarrier. Comparing with the simulation for $\rho = 1$ bits per second per Hz with one subcarrier shown in Figure 5.9, we note that the relative positions of the different schedulers remain the same, but they all shift 3 dB upwards. To explain this 3 dB offset, we note that, as $t_e/\bar{t}_{IA}$ becomes large, the prescient scheduler performance approaches that of a water-filling-in-time power control scheme with no traffic constraints. The required SNR of a water-filling-in-time power control scheme to achieve any given spectral efficiency has been calculated previously by [43] and the values of interest are 0 dB SNR at one bits per second per Hz and 6 dB at two bits per second per Hz [43, Figure. 9]. Converting these SNR values to SNR per bit by dividing by ρ, we obtain the asymptotic lower bound of 3 dB which is observed in Figure 5.12. A similar simulation at a spectral efficiency $\rho = 3$ bits per second per Hz and $N = 1$ subcarrier is shown in Figure 5.13.

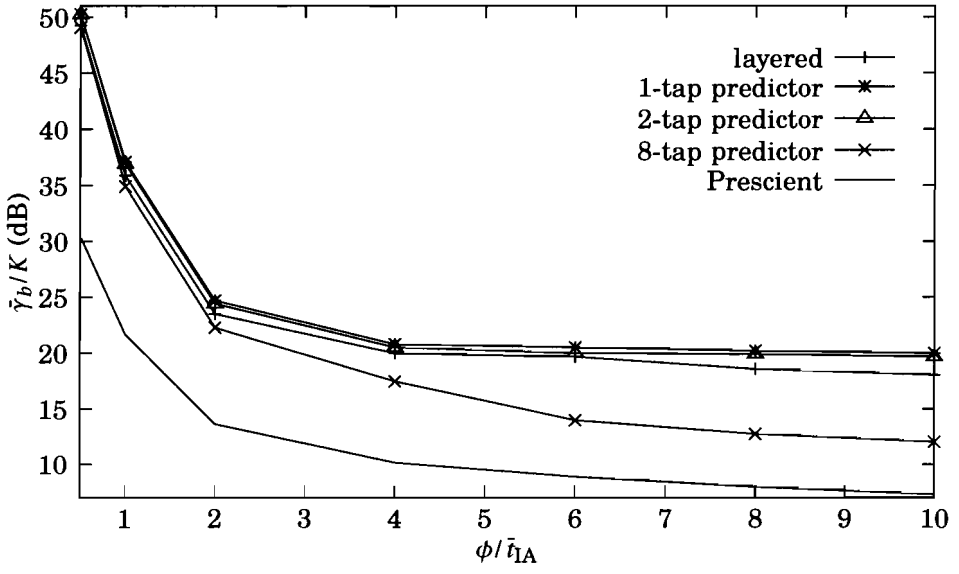

Figure 5.13: Comparison of the average SNR per bit ($\bar{\gamma}_b$) for $N = 1$ subcarriers at spectral efficiency $\rho = 2$ bits per second per Hz.

Chapter 6

Effect of Queueing Disciplines

So far, we have limited our attention to systems with first in first out (FIFO) queues. In this chapter, we discuss two earliest deadline first (EDF) queueing disciplines and provide a description of the modification required to the iterative string pulling (ISP) algorithm necessary for obtaining the optimal solutions under these queueing disciplines. Specifically, we consider both non-preemptive priority queue and preemptive priority queues. Note that from the arrival and expiry diagram shown in Figure 3.2, it is clear that the arrival of a packet with a short deadline would require all previously arrived packet to be cleared out of the queue even if they have a later deadline in the case of a simple FIFO queueing discipline. For example, see packet 7 in Figure 3.2. This requires a higher transmission rate than if packets can be re-arranged so that the packet with the earliest deadline is serviced first.

Intuitively, being able to rearrange the queue allows a late-arriving packet with an earlier deadline to be transmitted first without the need to service all of the packets in the queue, and it follows that some performance improvement can be obtained. However, the expected performance gain is highly dependent on the type and mix of quality of service (QoS) classes present in the system. For example, the constant

delay cases used in providing the performance result of Chapter 5 requires no queue rearrangement, since late arriving packets always have a later deadline than those already in the queue, due to the constant delay.

Since the performance improvement obtained is highly dependent on the type of traffic, we concentrate instead on describing how one can obtain the *prescient* performance bounds for these queueing disciplines in Section 6.1 and Section 6.2, instead of providing simulation results for an arbitrarily chosen traffic model. Simulation based on the methods described in the rest of this chapter can be performed to determine if there is any need for implementing the more sophisticated queueing disciplines once the exact traffic model is determined.

As for incorporating these priority queues into the online schedulers, recall that the online schedulers described in Chapter 5 perform optimization at the beginning of every epoch based on traffic constraints calculated from the "current" queue state. Since the expiry constraints (5.4c) are updated at the beginning of every epoch, rearranging the queue in the EDF order can be performed prior to computing the "current" expiry constraints for optimization. This extra step requires little computation and may result in reducing the total transmission energy. The question remaining is how to obtain the performance bound of the *prescient* scheduler with these queueing disciplines.

6.1 Non Preemptive Priority Queue

In this section, we consider EDF queue rearrangement with a non-preemptive priority queue. A non-preemptive priority queue allows higher priority packets to be placed at the head of the queue to be transmitted right after the current packet is transmitted. That is, with a non-preemptive queue server, the system must complete the trans-

mission of the packet currently in transmission before servicing this "high priority" packet. With this mode of transmission, the receiver will receive a complete packet at a time and requires little extra overhead in constructing the communication protocol over the top of a FIFO queue.

As an example, consider the packet arrivals and expiries as shown in Figure 3.2. At the arrival instant of Packet 7, the queue can be in one of five states. It may be servicing any of the packets 3 – 6, or empty as shown in Figure 6.1. With a non preemptive priority queueing discipline using the EDF ordering, there is an option of changing the order of servicing for packet 6 and 7 if the queue state is such that packet 6 is not already in service. The alternative expiry curve with reordering is shown as the dotted line in Figure 6.1 and we observe that at this instant in time, if the queue is in state 1 – 3, then the relaxed expiry constraints should be used as the cumulative expiries curve instead.

Next, we observe that the alternative expiry curve is always a less restrictive constraint for the optimization problem. This observation allows us a quick method for determining the optimal rate schedule under this new queueing discipline as described below.

Consider the time period immediately surrounding the arrival instant of packet 7 shown in larger magnification in Figure 6.1. Assume first that the system following the optimal transmission rate profile of the FIFO scheme up to the time instant just prior to the arrival of packet 7, that is, this is the first time queue rearrangement took place after an empty queue state. Depending on the actual system parameters, the queue could be in one of the four states shown in the figure. Note that EDF queue reordering can only take place only if the optimal transmission rate profile calculated under the FIFO queueing discipline place the system in state 1 – 3 at this time instant. Note that packet 6 is in service if the system is in state 4 and the EDF reordering can take

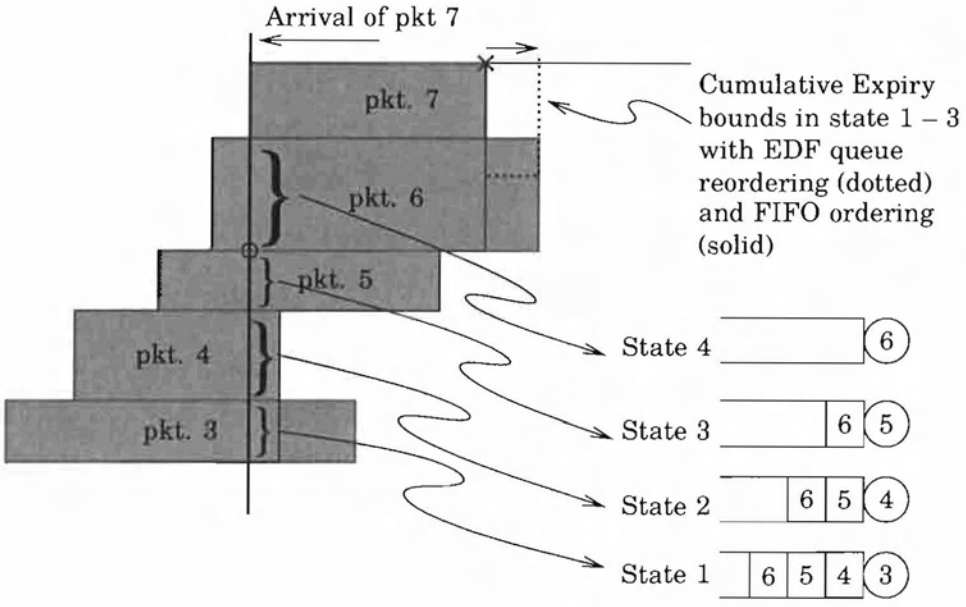

Figure 6.1: A close up view of the cumulative arrivals and expiries showing the relaxation of the expiry constraint when queue reordering takes place. Each of the queue diagram shows the state of the system assuming that the actual cumulative rate curve intersects the vertical axis at the point range shown.

place with a non-preemptive priority queue. Furthermore, relaxation of constraints in a convex optimization problem can provide a better objective only if the constraint being relaxed is an active constraint (indicated by the cross in the diagram). That is, an EDF non preemptive priority queueing discipline can have a lower energy usage if and only if the optimal transmission profile for the FIFO queue under the identical traffic and channel conditions satisfies the following:

1. Queue reordering can take place, i.e. cumulative rate curve intersect the vertical arrival indicator line below the point indicated by the circle in Figure 6.1.
2. The expiry constraint at the end of the overtaking packet is active, i.e. cumulative rate curve touches the expiry bound at the point indicated by the cross in Figure 6.1.

If reordering cannot take place or if the constraint to be relaxed is not active, then the transmission profile for the FIFO discipline is also optimal under the non preemptive EDF queueing scheme and we simply use the solution obtained from solving the FIFO system. Otherwise, we use the relaxed expiry constraints after reordering to compute the optimal rate profile. Thus, to obtain the *prescient* optimal transmission profile under the non-preemptive EDF queueing discipline, one would firstly obtain the optimal rate profile assuming a FIFO queueing discipline. Next, the active expiry constraints of the FIFO solution shall be considered for possible relaxation and the relaxed cumulative expiries curve be used to compute the updated transmission rate profile. The relaxation step shall be repeated until no queue reordering need to take place.

Finally, it shall be noted that if in an earlier iteration, the queue is in a relaxable state, e.g. when queue re-ordering can take place as in state 1–3 in example given in Figure 6.1, then relaxing the constraint can only reduce the slope of the cumulative rate profile and the queue will remain within a relaxable state.

A corresponding online scheduler using the EDF non-preemptive priority queue can be easily derived following the a similar process shown in Chapter 5 with an additional queue reordering step prior to the computation of the partial expiry curves.

6.2 Preemptive EDF Queue

Next, we consider the case with a preemptive EDF priority queue. Conceptually, allowing an urgent packet to be transmitted immediately by interrupting a current packet transmission should result in a better performance than that of a non-preemptive system. However, in order for a system to operate in this mode, the system must be able to deal with interruption and resumption of packets and would require extra overhead in the transmission protocol. Despite this undesirable implementation overhead, we will show the steps necessary in analyzing the performance of this preemptive EDF queue for completeness.

The optimal solution for the preemptive case is similar to the non-preemptive case in that the solution for the FIFO equivalent is the optimal solution provided that the constraint of the preempting packet is not active. However, that is where the similarity ends. Since a preemptive priority queue allow packet currently in service to be interrupted and resumed once the preempting packet exits the queue, there is a continuous set of possible relaxations of expiry constraints, depending on when the interruption is initiated, instead of a finite number defined by the possible ordering of packets and the method of examining all possible relaxations of Section 6.1 cannot be used.

For the analysis of this preemptive EDF system, a layered view of the interrupt and resume processes is required. Consider firstly, some background on the common method of analyzing a preemptive priority queue. Assume first, a priority queue with

two priority classes. On the arrival of a high priority packet, it enters the queue with all existing high priority packets placed ahead of it and overtakes all low priority packets. Hence, for all packets in the high priority class, it sees a FIFO queue with only packets in the highest priority class. For the lower priority class, it sees a queue with a server that is only available if it is not servicing a higher priority packet. Hence, we can model the system as two FIFO queues with two servers, one for each priority class. For the high priority queue, it operates as a normal FIFO queue and for the lower priority queue, the server is only available when the higher priority queue is in the empty state.

With this model of preemptive priority queue modelled as layers of queues in mind, we now examine a similarly layered view of the water-filling process. The description given here is based on the geometric interpretation of a concept known as *marginal utility* and a rigorous mathematical treatment can be found in [45]. Consider for now, the transmission of a single packet of length L to be transmitted within T seconds over a time-varying channel $g(t)$ as shown in Figure 6.2. It can be easily shown that the energy-minimal way of transmitting this packet is water-filling over the finite period of this T seconds. Now consider the addition of a higher priority packet as shown with length L_2 bits arriving at time T_a and expiring at time T_e. The addition of this packet has the following two effects:

1. The total number of bits transmitted in the period $[0, T]$ increases by L_2 bits.
2. The system has an additional constraint that the bits transmitted during the period $[T_a, T_e]$ must be equal to or higher than L_2 bits.

To obtain the optimal power allocation that satisfies the above two constraints, one can firstly perform water-filling in the period $[T_a, T_e]$ so that there is enough power to transmit L_2 bits. Next, we "freeze" this column of water in place to ensure that condition 2 is always satisfied and then perform water-filling again over the period

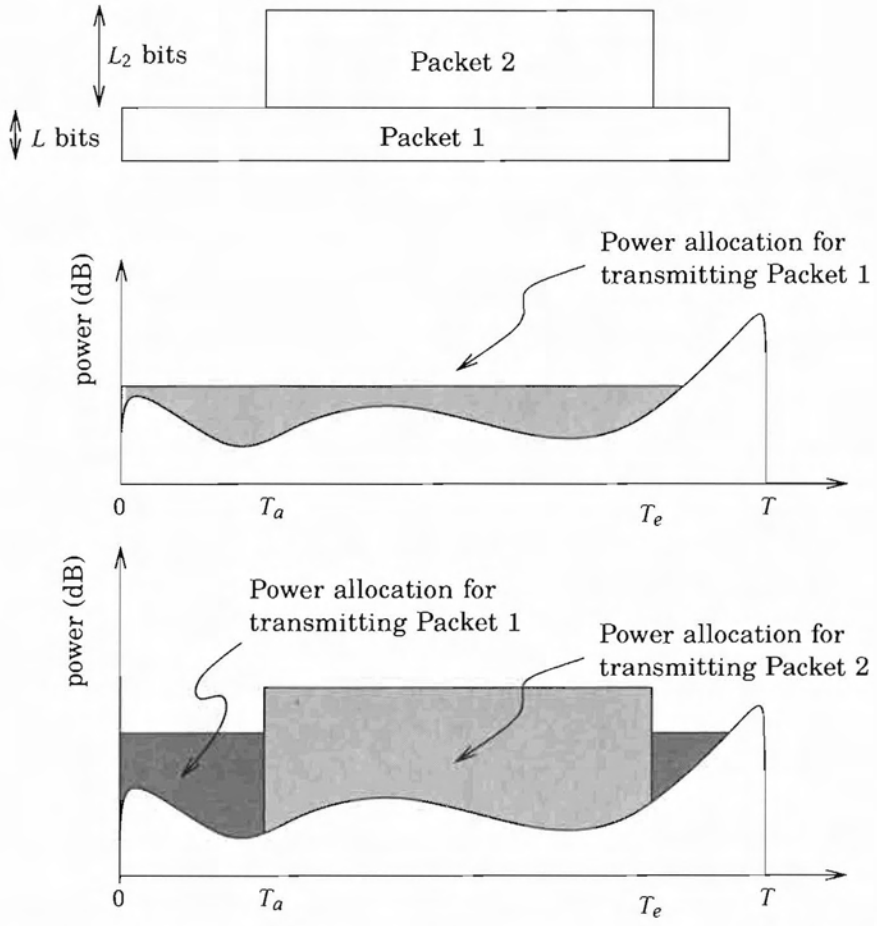

Figure 6.2: Diagram showing optimal power allocation for a preemptive priority queueing scheme.

$[0, T]$ with the remaining L bits. With L_2 bits frozen in place, this second iteration of water-filling would flow around and produce the optimal power profile of transmitting these lower priority bits. Next, we extend the discussion to the general case where packets are arriving and expiring randomly.

To compute the optimal rate profile for this preemptive priority queue, the packets need to be classified into different priority levels based on the priority level of the packets it "overtakes" when it arrives in the queue. Qualitatively speaking, these priority levels are assigned so that the packets with the highest priority see only themselves while packets with lower priorities see all packets with equal or higher priority than it. That is, an arriving packet only "overtake" packets with priority lower than it. Given these priority levels, the optimal rate profile can be calculated by firstly compute the optimal transmission rate with only the highest priority packets since these packets only sees the effect of the time-varying channel and themselves. The next highest priority packets are considered next. To transmit these packets, power needs to be allocated over and above what is already being transmitted and the extra power allocation can be computed as another layer of piecewise water-filling against the effective transmit noise to signal ratio $(1/g)$ and the power to transmit the higher priority packets optimally. Using the water-filling analogy, the process is similar to performing piecewise water-filling based on the highest priority packets, and freezing the water before another layer of piecewise water-filling is performed.

Next, we examine more closely the details of assigning the priority levels to the packets based on their deadline constraints. Firstly, we note that by definition, a packet that overtakes another must have priority higher than the packet it overtook. Secondly, a packet can only overtake packets that are still in the queue. Hence, the priority level of a sequence of packets with arrival time $T_a(i)$ and expiry time $T_e(i)$ for packet i can be assigned firstly by letting the first packet having the lowest priority of

1. Then, consider the packets in the sequence they arrive and form a queue. If packet 2 has an expiry time earlier than that of packet 1, it is assigned a priority level of 2, or else it is assigned a priority of 1. In general, we compare the expiry time of packet n with the expiry time of all the packet preceding it, and assign an appropriate priority level so that is has priority level at least one higher then the priority level of all the packets it overtook during queue rearrangement.

6.3 Conclusion

In this chapter, we discuss the method for obtaining the performance of the *prescient* scheduler under both preemptive and nonpreemptive EDF priority queue. Firstly, we observe that by allowing packets with earlier deadlines to have higher priority in the queue, the associated cumulative expiry constraint is relaxed. These results in potential improvement of performance. However, the amount of improvement is highly dependent on the type of traffic present in the system. With no specific traffic model in mind, we present only the methods for obtaining the *prescient* performance bound under these EDF dicipline and leave numerical simulation as future work when more indepth knowledge of the traffic model for such wireless internet applications is available.

The procedures for obtaining the *prescient* performance bound for the nonpreemptive and preemptive queues are very different. For the nonpreemptive EDF queues, an iterative relaxation of expiry constraints is applied to the packets as reordering takes place. Due to the nonpreemptive nature of the queue, there is only a finite number of possible packet orders and an iterative procedure is presented for obtaining the *prescient* performance bound under the nonpreemptive queueing dicipline.

As for the preemptive queues, a high priority packet can interrupt any low priority

packet currently in transmission. This results in an infinite number of possible expiry curves and the iterative procedure for analyzing the nonpreemptive system cannot be used here. Instead, we resort to the *marginal utility* description of the water-filling process and by assigning priority levels to packets based on the packets they overtook during preemptive reordering, a multilevel waterfilling solution was obtained. The method for obtaining the solution can be described by repeatedly performing piecewise water-filling and "freezing" for packets from the highest priority class to the lowest.

Finally, it should be noted that EDF queue reordering can be easily implemented for the causal schedulers presented in Chapter 5 by adding an extra queue reordering step just prior to calculating the causal expiry constraints for each symbol or epoch.

Chapter 7

Multi-user Schedulers

In this chapter, we consider the packet scheduling problem for multi-user scenarios, specifically the uplinking and downlinking stages of a trunked communication system (see Figure 2.1). In these multi-user scenarios, the cross layer schedulers must also consider the issue of user access control in addition to the time variation in channel states and the random arrival and expiry of packets.

To see how this additional consideration of user access control can affect the performance of the scheduler, first consider the following simple scenario. Let us begin by consider a layered design where packet scheduling is performed independently of the user access control. Specifically, consider a two-user system with a simple round robin access control scheme that is not aware of channel and traffic information. That is, user A is allocated to transmit in all the odd time devision multiplexing (TDM) time slots while user B is allocated the even slots. Since the two users are allocated orthogonal channels each with half of the system bandwidth, each user can perform optimal packet scheduling independently without knowledge of the other users channel state information (CSI). Thus, the average energy per bit usage over these two users is simply the average of the average energy per bit for each user with half of

the total system bandwidth. The causal schedulers can also be used with this type of round robin access control and the performance of such system can be inferred from the simulation results presented previously. The big question here is this: is there any energy saving that can be gained by jointly considering user access control as part of the scheduler? The answer is yes, and we will see how this is possible next. Now, consider the same two-user system described previously and note that each user independently performs optimal scheduling within its allocated time slots. Under the assumption that the channel gains are time-varying, some of these time slots will be idle due to a combination of bad channel gain and low urgency of the traffic in these slots for this given user. Since this user is not utilizing all of its time slots, these idle slots can be allocated to the other user for transmission without affecting the performance of this user while potentially decreasing the overall energy usage of the other user. Hence, there is a diversity gain to be achieved by this simple reallocation of time slot without even resorting to the more complicated multi-user coding schemes.

In the rest of this chapter, we provide a more detailed discussion of the design of the multi-user cross-layer scheduler. Following a similar progression as the development of the single user schedulers, we begin with a brief overview and literature survey of the user access control problem by highlighting some of the important results in Section 7.1. We then introduce the information theoretical formulation of the energy minimizing problem for both the multi-access channel (MAC) and broadcast channel (BC) channels in Section 7.2 and show that the problem is convex, hence can be solved efficiently. Instead of concerning ourselves with the details of the Karush-Kuhn-Tucker (KKT) conditions and the exact derivation of the result, we provide a more qualitative discussion on the general behaviour of the multi-user piecewise water-filling solution and the required steps to derive an equivalent set of online schedulers for the multi-user setting. Finally, we consider a more practical and prag-

matic scheduling approach by concluding the treatment of multi-user channels in Section 7.4 by revisiting the round robin access control scheme and providing a pragmatic scheduler design that does not require special multi-user coding.

7.1 Background on User Admission Control

The user admission control problem for maximizing the sum capacity of multi-user channels is widely studied. In [35], the problem of maximizing the sum capacity of a MAC channel without diversity is studied. It was found that the optimal transmission scheme is to allow only one user to access the channel at any instant to avoid performance degradation due to interference. However, by maximizing the total throughput of the system, the users with a better average channel gains are given access more frequently than the weaker users. This can be seen as a direct consequence of formulating the problem to maximize total throughput. With no other constraints, the best channel is utilized to provide the maximum throughput irrespective of whether there is any actual data to transmit. To provide a *fair* share of the resources, often a fairness parameter is introduced which results in various proportionally fair scheduling algorithms such as those proposed in [46]. The introduction of this fairness parameter allows weaker user to gain access to the channel while preserving the desired one-user-at-a-time transmission structure but the solution no longer achieves capacity. By characterizing the complete capacity region of the MAC with inter symbol interference (ISI) (frequency-varying channels), [47] presents the optimal transmission power profiles to achieve any rate combinations on the capacity boundary. From this solution, one can easily find the capacity achieving transmission scheme along the direction of any rate vector as desired to achieve fairness.

Similar results also apply to the BC and can be found in [48]. In this formulation, a weighted sum of transmission rates is maximized and these weights indicate

the relative *priority* of each user. With these multi-user waterfilling solutions, time sharing ceases to be optimal. In these cases, simultaneous transmission of the user signals with interference cancelling receiver MAC or superposition coding (BC) is required to achieve capacity. Most importantly, the optimal solution can be interpreted with a geometric interpretation of multi-user water-filling (see [48, Figures 4–8] and [47, Figures 12–13]). We will not dwelling too much on the exact derivation of the multi-user water-filling solution. Instead, interested readers are referred to [47].

While these researches are for capacity maximization, the geometric interpretation of multi-user water-filling provided by [47] and [48] can be used to provide some insight into the energy minimization problem for the multi-user system. That is, the optimal transmission rate profile for the multi-user channel must take the form of piecewise multi-user water-filling where within each water-filling segments delimited by active arrival and expiry constraints, the optimal power allocation must take on the form of multi-user water-filling as found in [48, 47], and is only allowed to change at epoch boundaries where one of the constraints is active.

To conclude the background section, we discuss briefly the differences and similarities between capacity maximization and energy minimisation formulations. In the single user case, there is vary little difference as discussed in Chapter 2. Both formulations results in the waterfilling solution and differs only in the actual waterlevel to be used. Specifically, consider the Lagrangian for these two problem formulations. The Lagrangian for the capacity maximization is in the form of $\texttt{(capacity)} - \mu(\texttt{energy} - \texttt{const2})$ where `const` is the average power constraint. On the other hand, the Lagrangian for the energy minimisation problem is in the form of $\texttt{energy} - \nu(\texttt{capacity} - \texttt{const1})$. With some simple mathematical manipulations, it is easy to see that these two Lagrangian is equivelent with $\mu = \nu^{-1}$ and both results in the waterfilling solution with the actual waterlevel depending in the value of `const1` and `const2`. The case for

the multi-user case is, however, different as will be presented next.

For the multi-user system, we no longer have the simple inverse relationship between the Lagrangian multipliers. This is due to the fact that there are now N constraints with one objective for an N user system and there is no longer a one to one match between the variables of these two different formulation. For the purpose of this thesis, we will concerntrate on minimizing the total energy usage with rate (capacity) constraints for each user.

7.2 Prescient Problem Formulation

In this section, we present the mathematical formulation of the prescient schedulers for the MAC and BC. Just like the single user case, the optimal rate obtained through optimizing the prescient formulations gives us a tight performance bound for any practical, causal scheduler variations we wish to consider. However, the actual numerical simulation and comparison are not included in this thesis as the emphasis of this thesis is on the theoretical study of the schedulers rather then system modeling issues.

7.2.1 Multi-access Channel

The multi-access channel (MAC) is a multipoint-to-point communication system. Packets from different sources arrive at a single base station and can potentially interfare with each other. For simplicity, we consider only the case where each user channel is a single-carrier flat fading channel and denote the N user MAC channel power gain by the length-N vector of individual channel power gains $\boldsymbol{g}(m) = \left(g^{[1]}(m), \cdots, g^{[N]}(m)\right)$. As for the traffic constraints, the cumulative arrival and expiry curves must also be specified for each user individually, i.e. $\boldsymbol{A}_{\text{cum}}(m) = \left\{A^{[1]}_{\text{cum}}(m), \cdots, A^{[N]}_{\text{cum}}(m)\right\}$ and $\boldsymbol{E}_{\text{cum}}(m) = \left\{E^{[1]}_{\text{cum}}(m), \cdots, E^{[N]}_{\text{cum}}(m)\right\}$. The joint transmission rate profiles $\boldsymbol{r}(m) = \left\{r^{[1]}(m), \cdots, r^{[N]}(m)\right\}$

are admissible if the rate profile for each user is admissible as defined for the point-to-point system. That is, in vector notation, $\boldsymbol{r}(m)$ is admissible if and only if the individual rate profile meets the bounding and non-negativity constraint, i.e.

$$\boldsymbol{r}(m) \succeq \boldsymbol{0} \tag{7.1}$$

$$\boldsymbol{A}_{\text{cum}}(m) \succeq \boldsymbol{r}(m) \succeq \boldsymbol{E}_{\text{cum}}(m) \tag{7.2}$$

where $\succeq$ denotes elementwise inequality.

Next, we turn our attention to the formulation of the objective. The energy minimal way to transmit in a MAC without traffic constraints is to allow all users to transmit simultaneously and use successive interference cancellation in the receiver [9]. Consider a two-user MAC with static channel gains g_1 and g_2 and the system is to transmit at r_1 and r_2 nats of information over unit bandwidth. Without loss of generality, we assume $g_1 > g_2$. The signal to noise ratio (SNR) requirement is

$$\frac{e^{2r_1}-1}{g_1} + \frac{e^{2r_2}-1}{g_2} \tag{7.3}$$

if the two users were to transmit at 50% duty cycle each.

With successive interference cancellation, the rate pair (r_1, r_2) can be achieved with transmit power $P_{\text{total}} = P_1 + P_2$ if the following set of inequalities holds.

$$r_1 \leq \ln\left(1 + \frac{g_1 P_1}{W N_0}\right) \tag{7.4}$$

$$r_2 \leq \ln\left(1 + \frac{g_2 P_2}{W g_1 P_1 + W N_0}\right) \tag{7.5}$$

The optimal energy usage is achieved when each user only sees weaker users as interference and signal from the stronger users are assumed to be detected and cancelled perfectly before detection. It can be shown (see for example, [9]) that the minimum

power usage for any given rate pair can be expressed as

$$P_1 + P_2 = WN_0\left(\frac{e^{r_2}(e^{r_1}-1)}{g_1} + \frac{e^{r_2}-1}{g_2}\right) \tag{7.6}$$

Expanding and removing the constant terms, it is easy to see that the total power usage is simply a sum of exponential of the rates and is convex in these variables.

Next, we introduce the mathematical form for the time varying channel. Again, we assume that the channel is block fading and the deterministic problem is divided into m symbols such that the channel state vector $\boldsymbol{g}(m)$ has constant entries. For convenience of notation, we define the order function $H_{12}(m)$ and $H_{21}(m)$ as

$$H_{12}(m) = \begin{cases} 1 & g_1(m) \geq g_2(m) \\ 0 & g_1(m) < g_2(m) \end{cases} \quad \text{and} \quad H_{21}(m) = \begin{cases} 0 & g_1(m) \geq g_2(m) \\ 1 & g_1(m) < g_2(m) \end{cases} \tag{7.7}$$

for the two-user case. In general, let $H_{\text{abc}}(m) = 1$ if and only if the channel gains is in the order $g_a \geq g_b \geq g_c$. Note that H_{12} and H_{21} can be determined from knowledge of the channel gains and are assumed to be known *a priori* for the deterministic formulation. The time-varying MAC optimization problem for two users can thus be stated as

$$\begin{aligned} \text{Minimize: } & \sum_{m=1}^{M} H_{12}(m)\left(\frac{e^{r^{[2]}(m)}(e^{r^{[1]}(m)}-1)}{g^{[1]}} + \frac{e^{r^{2}(m)}-1}{g^{[2]}(m)}\right) \\ & + \sum_{m=1}^{M} H_{21}(m)\left(\frac{e^{r^{[1]}(m)}(e^{r^{[2]}(m)}-1)}{g^{[2]}(m)} + \frac{e^{r^{[1]}(m)}-1}{g^{[2]}(m)}\right) \end{aligned} \tag{7.8a}$$

$$\text{Subject to: } \sum_{i=1}^{m} r^{[n]}(i) \geq E^{[n]}_{\text{cum}}(m) \quad \text{for} \quad m \in [1,M],\ n \in \{1,2\} \tag{7.8b}$$

$$\sum_{i=1}^{m} r^{[n]}(i) \leq A^{[n]}_{\text{cum}}(m) \quad \text{for} \quad m \in [1,M],\ n = \{1,2\} \tag{7.8c}$$

$$r^{[n]}(m) \geq 0 \quad \text{for} \quad n \in \{1,2\},\ m \in [1,M] \tag{7.8d}$$

Similarly a general expression for the objective for N users can be derived as

$$P_{\text{total}} = N_0 \sum_{n=1}^{N} \left(\frac{(e^{r_n} - 1) \exp\left(\sum_{i=n+1}^{N} r_i\right)}{g_n} \right) \tag{7.9}$$

assuming that the channel gains are ordered such that $g_1 > g_2 > \cdots g_N$. For those interested in the derivation of (7.9), see Appendix B.

While this problem formulation is very cumbersome for analysis purposes, it is easy to specify numerically and can be easily solved using AMPL following similar steps as specified in Section 3.1.

7.2.2 Broadcast Channel

The scheduling problem for the BC can be formulated in a similar way. In fact, it can be shown that the optimal power and rate allocations for the BC can be found by solving the MAC problem with the same channel gain vectors. Specifically, we use the duality between MAC and BC. Simply, to quote [49] on duality between multi-input multi-output (MIMO) MAC and MIMO BC:

> We establish this duality by showing that all rates achievable in the dual MIMO MAC with power constraints whose sum equals the BC power constraint are also achievable in the MIMO BC, and vice versa.

While this quote is in reference to the more general dual relationship between MIMO BC and MIMO MAC, the result also applies to the more restricted case of single input single output (SISO) channels. In the context of energy minimization, the energy required to transmit at any given set of rates in BC is equal to the total energy required to achieve the same rate vector under the dual MAC, which is exactly the objective that we are minimizing in the previous section. In other words, under identical traffic constraints, the rate profile that achieves minimal energy usage in a MAC must be

also achievable in the dual BC. To show that these rate profiles also achieve minimal energy usage in the dual MAC we perform proof by contradiction. Assume that there exist rate profiles with lower energy usage in the dual BC, then by the reverse statement of duality, these new rates are also achievable in the MAC and have lower energy usage than the minimal rate profiles which is a contradiction. Thus, there is only one problem to solve for both the MAC and BC channels.

7.3 Multi-User Piecewise Waterfilling

In this section, we briefly discuss the general form of the solution to problem (7.8) and show that in some special cases, the iterative string pulling (ISP) algorithm can be used to obtain the optimal solution.

Note firstly, that by forming the Lagrangian and differentiating with respect to the rate for each user, $r^{[n]}(m)$, the differentiation operation sifts out only the constraints associated with user n. Thus, the set of KKT conditions for the single user problem presented in Section 3.1.4 holds for the multi-user case with the exception of the definition of the energy objective $f(r(m), g(m))$, which should be replaced with the multi-user version of the energy objective as shown in (7.8a). Next, we generalize the definition of a *water-filling segment* for each user to be a sequence of consecutive epochs with slack arrival and expiry constraints for the specific user, in which the ordering of the channel gains remain the same across these epochs. Thus, we must have

$$\frac{\partial}{\partial r^{[n]}} f(\boldsymbol{r}(m), \boldsymbol{g}(m)) = \frac{\partial}{\partial r^{[n]}} f(\boldsymbol{r}(m+1), \boldsymbol{g}(m+1)) = \omega^{[i]} \tag{7.10}$$

if the traffic constraints for epoch m is slack, i.e.,

$$\sum_{i=1}^{m} r^{[n]}(i) > E_{\text{cum}}^{[n]}(m) \tag{7.11}$$

$$\sum_{i=1}^{m} r^{[n]}(i) < A_{\text{cum}}^{[n]}(m) \tag{7.12}$$

By letting $f(\boldsymbol{r}(m), \boldsymbol{g}(m))$ to be the energy usage implied by the multi-user capacity formulation, it follows that the optimal power allocation across epochs with slack traffic constraints must be in the form of multi-user water-filling. Furthermore, by noting that this rate vector is on the capacity boundary and is only achievable through the use of successive interference cancellation, it follows that the each user must allocate power against equivalent receiver noise and interferences from weaker users. Thus, the multi-user water-filling solution can be interpreted as the following. Firstly, the optimal power allocation for the weakest user must follow the form of single user water-filling, as it sees no interference. The power allocation for the next weakest user can be interpreted as water-filling against a floor of receiver noise and interference from the weakest user.

In the special case where the order of the channel gains is unchanged over a complete simulation run, the optimal rate profiles can be obtained by performing the ISP algorithm for the traffic and channel for the weakest user, as this user sees no interference. Once the optimal power and rate allocation are determined for this user, the interference level seen by the next weakest user can be determined and the optimal transmission power and rate can be determined for the second weakest user. This process is to be repeated till the rate and power allocations for all users are determined.

For the prescient problem in general, the order of the power gains can change and the problem may be solved using general convex optimization algorithms such as the

simplex methods or the interior point methods. However, if one were to consider the causal problem formulation for the multi-user case, the assumption that the ordering of channel gains is unchanged may be true over the short optimization horizon and the successive application of the ISP algorithm can be used to implement the causal scheduler efficiently.

7.4 Practical Considerations

There are several practical issues associated with implementing an optimal scheduler that takes advantage of multi-user diversity. Firstly, the prescient multi-user scheduling problem (7.8), cannot be solved using the ISP algorithm proposed in Chapter 4 in general. Furthermore, the power function used to formulate the objective is based on ergodic capacity and the successive interference cancellation required for decoding the signals has a problem with error propagation in practice particularly for uncoded signals.

In this section, we turn our attention to a more pragmatic approach where the channel is shared among the users by allowing only one user to transmit at a time. Again, consider the round robin user admission scheme for a two-user system described at the beginning of this chapter. In the previous discussion related to this scheduler, we noticed that it is possible to reallocate the idle channels for user A to user B and vise versa. We will now describe a design based on this observation.

Initially, the transmission time is divided between two users equally and each user performs optimal scheduling within its own allocated time slots and meets its own packet expiry constraints. This results in an admissible, but suboptimal multi-user transmission profile. We have noted that by reallocating the idle slots of user A to user B, it is possible for user B to obtain a schedule with lower total energy usage

with these extra time slots. In fact, user B would calculate a new rate allocation profile with these extra channels added.

Furthermore, we note that obtaining more slots from other user might reduce some of the waterlevels but it can never increase the waterlevel. Hence, it is possible to iteratively identify more idle slots to be given to the other user. In general, we propose that for the N user system, the idle slots shall be reallocated based on the channel gain and shall be allocated to the non-idle user with the worst channel gain. That is, if the time slot is idle for user n, it is reallocated to user n' such that user n' is non-idle and it has the worst channel gain amoung all non-idle users. This rule is based on the premise that an extra time slot provide more energy saving in low channel gain situations then that of higher channel gains. This heuristic is plausible, but has not yet been tested. Testing and comparison with the optimal solution obtained using a generic convex optimization package is planned for future work.

7.5 Conclusion

In this chapter, multi-user considerations are discussed. The discussion begins with the convex formulation of the *prescient* scheduler for multiuser channels by a simple modification to the energy objective function. An interesting side effect of formulating the problem as an energy minimizing problem is that, through duality, the MAC and BC channels have identical optimal rate profiles and the achievable total energy usage is also the same. We also propose a more pragmatic approach to multi-user scheduling by modifying the round robin user access scheme with slot reassignment to obtain better channel usage.

Chapter 8

Conclusions and Future Directions

8.1 Summary

This thesis presents the work on obtaining optimal packet schedulers under per-packet delay constraints. The problem is considered from both the theoretical and practical perspectives.

Through the formulation of the *prescient* schedulers under various channel models, it was established that the solution has the form of piecewise water-filling. Specifically, for the single user systems, the solution can be constructed by water-filling within each *waterfilling segment* delimited by active traffic constraint points over all parallel channels. We also consider the modification required for multi-user channels and the effect of queueing disciplines. On the more practical side, we present the details of the formulation of several causal schedulers based on minimizing energy usage of the expected future channel gain under various assumptions.

Several novel and interesting results were obtained in this investigation. Firstly,

and most importantly, it was established that the optimal power and rate allocation follows the form of piecewise water-filling. Simply speaking, the transmission profile can be divided into smaller time segments within which the water-filling phenomena applies. Furthermore, these time segments are delimited by active traffic constraints corresponding to either an empty queue state or just meeting a deadline constraint. From the perspective of packets in the queue, this phenomenon can be summarized with the following rule:

> Maintain a constant waterlevel unless there is a near miss of a packet deadline or there is nothing to transmit.

On the practical perspective, an algorithm based on finding the shortest path between two points under boundary constraints was proposed that solves the scheduling problem efficiently. This algorithm was developed firstly to provide a quick way of obtaining the *prescient* result but it was also useful in obtaining the result, for the more practically oriented causal schedulers. It is the author's opinion that using the iterative string pulling (ISP) algorithm, the proposed online schedulers can be implemented in real-time with currently available micro processors for deployment in a real communication system.

While the system under consideration for most part of this thesis is a single user multi-carrier system, most of the results can be extended to the multi-user system by a simple substitution of the objective function with the multi-user equivalent and the general form of multi-user piecewise water-filling solution can be inferred.

8.2 Future Directions

While this thesis covers the broad area of energy optimal scheduling under a variety of channels and causality assumptions, there are many detailed investigations that

can be carried out for the piecewise water-filling solution. For example, simulation and analytical results on the statistics of the duration of each water-filling segment as a function of fade rate and traffic parameters would further the understanding of the interaction between the traffic and the channel.

It should also be noted that the main emphasis of this thesis is to provide a theoretical analysis of the optimal scheduler under strict deadline constraints in a general setting. That is, we do not assume any specific statistical characteristics of the arriving traffic or channel gain variations. Despite obtaining a performance close to the *prescient* scheduler for the 8-tap channel prediction causal scheduler in the specific example provided, it is not representative of any real world systems. Further simulation and modeling is required to obtain realistic performance measures for real world systems. It would also be very interesting to derive an analytical expression of the performance gap between the prescient scheduler and the many causal variants.

Appendix A

Channel Power Gain Prediction

A Rayleigh fading channel with Doppler is often modelled as a complex gain that is a Gaussian distributed random process with a Jakes spectrum. With this model, a Wiener prediction filter can be used to provide an unbiased estimate of a future complex gain, $\hat{h} \simeq h$, with a well defined estimation error variance σ_h^2 [50].

Consider the L-point MMSE linear predictor $\boldsymbol{w}_k$ written as a column vector of filter weights satisfying the equation

$$\boldsymbol{w}_k^H \boldsymbol{h}^{[m]} = \hat{h}_k^{[m]} \tag{A.1}$$

where $\boldsymbol{h}^{[m]} = [h(m)\ h(m-\delta)\ \ldots\ h(m-L\delta+\delta)]^T$ contains L present and past channel gain samples at Nyquist spacing δ and $\hat{h}_k^{[m]}$ is the estimated channel gain for symbol $m+k$ at the beginning of symbol m. This prediction filter $\boldsymbol{w}_k$ is independent of m and can be obtained from solving the normal equation

$$\boldsymbol{R}_h \boldsymbol{w}_k = \boldsymbol{p}_k \tag{A.2}$$

where $\boldsymbol{R}_h$ is the covariance matrix of past samples and $\boldsymbol{p}_k$ is the vector of cross corre-

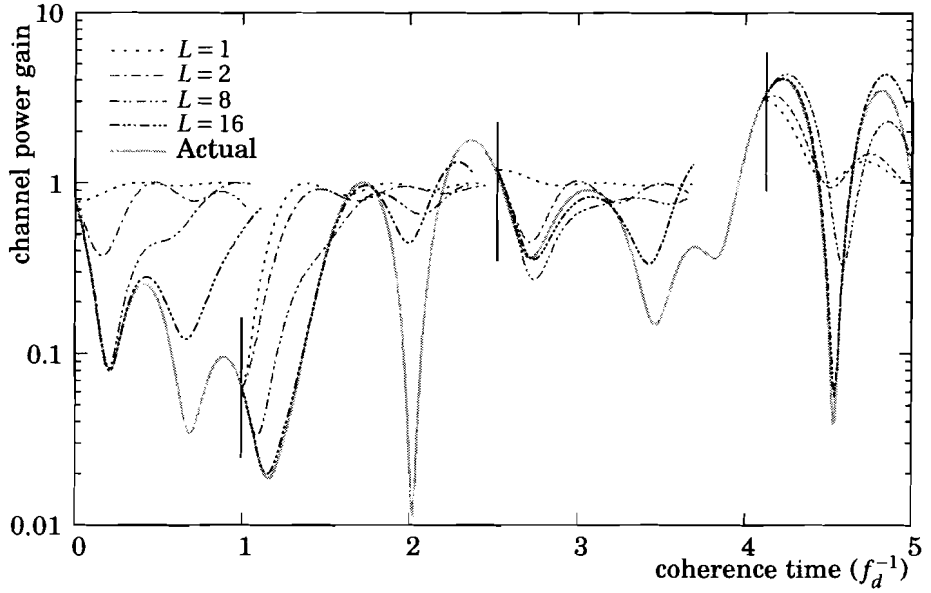

Figure A.1: Plot of the power gain of a Rayleigh fading channel and various MMSE predictor outputs for 1, 2, 8, and 16 taps.

lation of the past observations with the desired prediction value.

The prediction error variance for k symbol into the future can also be expressed simply in terms of the auto-correlation function as

$$\sigma_k^2 = 1 - \boldsymbol{p}_k^T \boldsymbol{R}_h^{-1} \boldsymbol{p}_k. \tag{A.3}$$

again, independent of m.

Given that the prediction $\hat{h}_k^{[m]}$ and the prediction error are uncorrelated, the conditional mean of the power gain at symbol $m+k$ is

$$\hat{g}_k^{[m]} = E\left[|h(m+k)|^2\right] = \left|\hat{h}_k^{[m]}\right|^2 + \sigma_h^2 \tag{A.4}$$

which is used to provide the channel power gain estimate for the causal scheduler. Figure A.1 shows four specific instances of the channel power gain predictors used

in the simulation with 1,2,8, and 16 taps. From this figure, we note that a one tap predictor simply decays from the last known channel gain back to the expected value while the higher order predictors can approximate the occurrence and duration of the first fade.

Appendix B

The Power Formula for the Multi-access Channel

The achievable rates for a general N user multi-access channel (MAC) with channel power gains $g_1 \cdots g_N$ with transmit power $P_1 \cdots P_n$ assuming $g_1 \geq g_2 \geq \cdots \geq g_N$ is at the corner of the polymatroid specified by

$$r_1 = \ln\left(1 + \frac{g_1 P_1}{N_0}\right) \tag{B.1}$$

$$r_2 = \ln\left(1 + \frac{g_2 P_2}{N_0 + g_1 P_1}\right) \tag{B.2}$$

$$r_3 = \ln\left(1 + \frac{g_3 P_3}{N_0 + g_1 P_1 + g_2 P_2}\right) \tag{B.3}$$

$$\vdots$$

$$r_N = \ln\left(1 + \frac{g_N P_N}{N_0 + \sum_{i=1}^{N-1} g_i P_i}\right) \tag{B.4}$$

Firstly, we rearrange (B.1) to obtain

$$g_1 P_1 = N_0\left(e^{r_1} - 1\right) \tag{B.5}$$

and substituting the left hand side (LHS) of (B.5) for $g_1 P_1$ in (B.2) and (B.3), we obtain

$$r_2 = \ln\left(1 + \frac{g_2 P_2}{N_0(e^{r_1})}\right) \tag{B.6}$$

$$r_3 = \ln\left(1 + \frac{g_3 P_3}{N_0(e^{r_1}) + g_2 P_2}\right) \tag{B.7}$$

Again, we can manipulate (B.6) to obtain an expression for $g_2 P_2$ as

$$g_2 P_2 = N_o e^{r_1}\left(e^{r_2} - 1\right) \tag{B.8}$$

and substituting into (B.7) to obtain

$$r_3 = \ln\left(1 + \frac{g_3 r_3}{N_0 e^{r_1 + r_2}}\right) \tag{B.9}$$

which in turn, can be rearranged to give $g_3 P_3$ as

$$g_3 P_3 = N_0 e^{r_1 + r_2}\left(e^{r_3} - 1\right) \tag{B.10}$$

and it is easy to see that in general

$$g_n P_n = N_0 e^{\sum_{i=1}^{n-1} r_i}\left(e^{r_n} - 1\right) \tag{B.11}$$

and the total power usage is

$$P_{\text{total}} = \sum_{n=1}^{N} P_n = N_0 \sum_{n=1}^{N} \frac{e^{\sum_{i=1}^{n-1} r_i}\,(e^{r_n} - 1)}{g_n} \tag{B.12}$$

Bibliography

[1] R. Berry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Process. Mag.*, vol. 21, no. 5, pp. 59–68, Sept. 2004.

[2] S. Nanda, K. Balachandra, and S. Kumar, "Adaption techniques in wireless packet data services," *IEEE Commun. Mag.*, vol. 38, pp. 54 – 64, 2000.

[3] A. Goldsmith and P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[4] J. Hayes, "Adaptive feedback communications," *IEEE Trans. Commun.*, vol. 16, no. 1, pp. 29–34, Feb 1968.

[5] J. K. Cavers, "Variable-rate transmission for Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 20, no. 1, pp. 15–22, Feb 1972.

[6] S.-G. Chua and A. J. Goldsmith, "Adaptive coded modulation for fading channels," in *Proc. ICC 97*, vol. 3, 8-12 June 1997, pp. 1488–1492.

[7] A. Goldsmith and S.-G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 595–602, May 1998.

[8] C. E. Shannon, "A mathematical theory of communications," *Bell System Technical Journal*, vol. 27, pp. 379–424 and 623–656, July and Oct. 1948.

[9] T. M. Cover and J. A. Thomas, *Elements of information theory*. Wiley & Son, 1991.

[10] S. Vishwanath and A. Goldsmith, "Exploring adaptive turbo coded modulation for flat fading channels," in *Proc. 52nd Vehicular Technology Conference IEEE VTS-Fall VTC 2000*, vol. 4, 2000, pp. 1778–1783 vol.4.

[11] S. T. Chung and A. Goldsmith, "Degrees of freedom in adaptive modulation: a unified view," *IEEE Trans. Commun.*, vol. 49, no. 9, pp. 1561–1571, 2001.

[12] W.-J. Choi, K.-W. Cheong, and J. Cioffi, "Adaptive modulation with limited peak power for fading channels," in *Proc. IEEE 51st VTC 2000-Spring Tokyo Vehicular Technology*, vol. 3, 2000, pp. 2568–2572 vol.3.

[13] W. Yu and J. Cioffi, "On constant power waterfilling," in *Proc. IEEE International Conference on Communications ICC 2001*, vol. 6, 2001, pp. 1665–1669 vol.6.

[14] A. J. Goldsmith and S.-G. Chua, "Variable-rate variable power MQAM for fading channels," *IEEE Trans. Commun.*, vol. 45, no. 10, pp. 1218 – 1230, 1997.

[15] V. Lau and M. Macleod, "Variable rate adaptive trellis coded qam for high bandwidth efficiency applications in rayleigh fading channels," in *Proc. 48th IEEE Vehicular Technology Conference VTC 98*, vol. 1, 1998, pp. 348–352 vol.1.

[16] D. Goeckel, "Strongly robust adaptive signaling for time-varying channels," in *Conference Record.1998 IEEE International Conference on Communications ICC 98*, vol. 1, 1998, pp. 454–458 vol.1.

[17] ——, "Adaptive coding for time-varying channels using outdated fading estimates," *IEEE Trans. Commun.*, vol. 47, no. 6, pp. 844–855, 1999.

[18] M. van Der Schaar and S. Shankar N, "Cross-layer wireless multimedia transmission: challenges, principles, and new paradigms," *IEEE Wireless Commun. Mag.*, vol. 12, no. 4, pp. 50–58, Aug. 2005.

[19] L. Ozarow, S. Shamai (Shitz), and A. Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. Veh. Technol.*, vol. 43, pp. 359 – 378, 1994.

[20] S. V. Hanly and D. N. C. Tse, "Multiaccess fading channels - part ii: Delay limited capacity," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816 – 2831, 1998.

[21] I. Telatar and R. Gallager, "Combining queueing theory with information theory for multiaccess," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 6, pp. 963–969, Aug. 1995.

[22] R. Berry and R. Gallager, "Communication over fading channels with delay constraints," *IEEE Trans. Inf. Theory*, vol. 48, no. 5, pp. 1135–1149, May 2002.

[23] E. Yeh, "Delay-optimal rate allocation in multiaccess communications: a cross-layer view," in *IEEE Workshop on Multimedia Signal Processing*, 9-11 Dec. 2002, pp. 404–407.

[24] B. Prabhakar, E. Uysal-Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *proc. INFOCOM'01*, 2001.

[25] E. Uysal-Biyikoglu, B. Prabhakar, and A. El Gamal, "Energy-efficient packet transmission over a wireless link," *IEEE/ACM Trans. Netw.*, vol. 10, no. 4, Aug. 2002.

[26] E. Uysal-Biyikoglu, "Adaptive Transmission for Energy-Efficiency for Wireless Data Networks," Ph.D. dissertation, Stanford University, 2003.

[27] E. Uysal-Biyikoglu and A. El Gamal, "On adaptive transmission for energy-efficiency in wireless data networks," *IEEE Trans. Inf. Theory*, vol. 50, no. 12, pp. 3081–3094, Dec 2004.

[28] M. Zafer and E. Modiano, "A calculus approach to minimum energy transmission policies with quality of service guarantees," in *Proc. INFOCOM'05*, 2005.

[29] ——, "Optimal adaptive data transmission over a fading channel with deadline and power constraints," in *Conference on Information Sciences and Systems*, Princeton, New Jersey, 2006.

[30] M. Zafer, "Dynamic rate-control and scheduling algorithms for quality-of-service in wireless networks," Ph.D. dissertation, MASSACHUSETTS INSTITUTE OF TECHNOLOGY, 2007.

[31] M. Khojastepour and A. Sabharwal, "Delay-constrained scheduling: power efficiency, filter design, and bounds," in *INFOCOM 2004*, vol. 3, March 2004, pp. 1938–1949.

[32] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[33] H. A. Taha, *Operations Research: An Introduction*, 8th ed. Prentice Hall, 2007.

[34] Z.-Q. Luo and W. Yu, "An introduction to convex optimization for communications and signal processing," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 8, pp. 1426–1438, 2006.

[35] R. Knopp and P. Humblet, "Information capacity and power control in single-cell multiuser communications," in *ICC 95*, vol. 1, 18-22 June 1995, pp. 331–335.

[36] T. M. Cover, "Broadcast channels," *IEEE Trans. Inf. Theory*, vol. 18, no. 1, pp. 2–14, Jan 1972.

[37] S. Kuo and J. K. Cavers, "Optimal scheduling for wireless links under per-user maximum delay constraints," in *proc. GLOBECOM'07*, Washington DC, 2007.

[38] R. Fourer, D. M. Gay, and B. W. Kernighan, *AMPL: A Modelling Language for Mathematical Programming*. Duxbury Press, 2002.

[39] R. Andreani, E. G. Birgin, J. M. Martï£¡nez, and M. L. Schuverdt, "On augmented lagrangian methods with general lower-level constraints," *SIAM Journal on Optimisation*, vol. 18, pp. 1286–1309, 2007.

[40] ——, "Augmented lagrangian methods under the constant positive linear dependence constraint qualification," *Mathematical Programming 111*, pp. 5–32, 2008.

[41] S. Kuo and J. K. Cavers, "Performance comparison of max-delay constrained schedulers in rayleigh fading channels," in *proc. VTC'08*, Singapore, Spring 2008.

[42] R. Vaughan, P. Teal, and R. Raich, "Short-term mobile channel prediction using discrete scatterer propagation model and subspace signal processing algorithm," in *proc. VTC*, 2000.

[43] M.-S. Alouini and A. J. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *IEEE Trans. Veh. Technol.*, vol. 48, no. 4, pp. 1165–1181, Jul. 1999.

[44] W. C. Jakes, Ed., *Microwave Mobile Communications*. Wiley & Son, 1972.

[45] D. Tse and S. Hanly, "Multiaccess fading channels - part i: Polymatroid structure and optimal resource allocation and throughput capacities," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[46] A. Jalali, R. Padovani, and R. Pankaj, "Data throughput of cdma-hdr a high efficiency-high data rate personal communication wireless system," in *Proc. IEEE*

51st VTC 2000-Spring Tokyo Vehicular Technology, vol. 3, 15–18 May 2000, pp. 1854–1858.

[47] R. Cheng and S. Verdu, "Gaussian multiaccess channels with isi: capacity region and multiuser water-filling," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 773–785, May 1993.

[48] A. Goldsmith and M. Effros, "The capacity region of broadcast channels with intersymbol interference and colored Gaussian noise," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 219–240, Jan. 2001.

[49] S. Vishwanath, N. Jindal, and A. Goldsmith, "Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2658–2668, Oct. 2003.

[50] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Prentice Hall, 1996.