Resource Management in CDMA-based Satellite Networks

Dorothy Kabagaju Okello

Department of Electrical & Computer Engineering
McGill University
Montreal, Canada

April 2004

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Abstract

There is interest, supported by successful field trials, in the use of satellite communications at the Ka band (30/20 GHz) and beyond to meet emerging demand for broadband interactive multimedia services. The key advantages of operation at Ka band are availability of bandwidth and favorable implications for terminal size, cost and mobility. We study two problems related to bandwidth management of the uplink in a multibeam, CDMA-based, GEO satellite. Our focus is on the delivery of data services with rigid constraints on bit-error rate and elastic constraints on data rate.

The first of the two problems concerns the design of the coverage areas of the satellite beams. We were interested specifically in the adaptation of beam shape to inhomogeneity in the geographic distribution of the user population, and in the impact of beam shaping on the set of transmission rates that are compatible with prescribed constraints on transmission powers and signal-to-interference ratios. Assuming that the spatial distribution of users is known, we construct an algorithm which computes beam coverage regions to equilibrate the per-beam user populations. The impact on the set of feasible bit-rate allocations is quantified through numerical experiments. Comparison with uniform beam shapes suggests that the adaptive approach is superior in terms of the number of concurrent transmissions that can be supported.

The second problem concerns the allocation of bit rates in a setting where user bit-rate requirements are assumed defined by *averages* over moving windows of constant length. We use a frame-based channel model characterized by fading coefficients which, though statistically variable, are assumed known to the controller at the start of each frame. The implied temporal elasticity in quality-of-service provides opportunity to achieve economies in transmitted power. The value of such opportunity is quantified by comparison of two extreme cases. We develop an approximate system model which allows *optimization* of the rate allocation when the number of users is small, and a heuristic which is useful when the number of users is not small. The associated performance results confirm the inverse relationship between the per-bit energy required for transmission and the length of the averaging window.

Sommaire

Les essais pratiques ont démontré que la demande grandissante pour les services multimédias intéractifs à grande largeur de bande peut être soutenue par les communications satellitaires à partir de la bande Ka (30/20GHz). L'avantage de la bande Ka est la disponibilité accrue de largeur de bande et la possibilité de construire à faibles coûts des terminaux plus petits et plus mobiles. Ce mémoire traite de deux problèmes de gestion de ressources pour le lien montant d'un réseau de satellites géo-stationnaires à multiples faisceaux électromagnétiques basé sur l'accès multiple par répartition en code (AMRC). Nous nous concentrons sur l'acheminement de services de données avec des contraintes rigides de taux d'erreur et des contraintes élastiques de débit.

Le premier des deux problèmes mentionnés ci-haut touche la conception de surface de couverture de faisceau électromagnétique satellitaire. Nous nous sommes intéressés plus spécifiquement à l'adaptation de faisceau à des distributions géographiques de population d'usagers non-homogènes et les conséquences sur l'emsemble des taux de transmission qui sont compatibles avec les contraintes de puissance et avec le rapport signal-interférence. En supposant que la distribution spatiale des usagers est connue, nous construisons un algorithme qui calcule les régions de couverture des faisceaux pour équilibrer la population d'usagers par faisceau. Les effets sur l'attribution de l'ensemble des taux binaires possibles sont étudiés par expériences numériques. La comparaison avec les faisceaux uniformes suggère que l'approche adaptative est supérieure en terme de nombre de transmissions simultanées possibles.

Le second problème touche l'attribution de taux binaires dans un scénario où les taux binaires requis par usager sont supposés être définis par une *moyenne* sur une fenêtre mobile de longueur fixe. Nous utilisons un modèle de canal basé sur des trames, qui est charactérisé par des coefficients en évanouissement qui sont supposés être connus par le controlleur au début de chaque trame par le biais d'une variable aléatoire. L'élasticité temporelle sous-entendue dans la qualité de service donne l'opportunité d'atteindre des économies de puissance de transmission. L'importance de cette opportunité est quantifiée en comparant deux cas extrèmes. Nous concevons un système approximatif qui permet *l'optimization* du taux d'attribution quand le nombre d'usagers est petit et heuristique lorsque le nombre d'usagers est plus grand. Les résultats confirment la relation inverse entre l'énergie par bit requise pour une transmission et la longueur de la fenêtre de moyenne.

Acknowledgments

I would like to express my sincerest gratitude to Professor Michael Kaplan, my thesis supervisor. His guidance and support have been invaluable for both this thesis and my personal development. I would also like to express my deep appreciation to Professor Kaplan for his patience, especially over the past few years.

I would like to thank Professor Peter Kabal and Professor Carl Tropper, members of my thesis supervisory committee, whose suggestions and input have helped to make this a better thesis.

I am very thankful to my friends and colleagues of the Telecommunications Signal Processing Laboratory, including Xumin Sun, Tamanna Islam, Hossein Najaf-Zadeh, Khaled El-Maleh, Ejaz Mahfuz and Jerome Pra from the beginning, Joachim Thiemann, Wesley Pereira, Benoît Pelletier, François Duplessis-Beaulieu, Ricky Der, Yasheng Qian and Alexander Wyglinski for programming and other technical assistance, Xiaojian Lu and Firas Jabloun my study mates, and all my labmates in McConnell 744. I am also very thankful to Benoît Pelletier for the french translation of my abstract.

I would also like to acknowledge the financial support provided by the Canadian Commonwealth Scholarship Program that enabled me to pursue my graduate studies at McGill University. I also acknowledge the support from the Department of Electrical Engineering, Makerere University, in enabling me pursue my studies.

Finally, I would like to thank my husband, Jimmy, and our son, Jordan, for their love, support and encouragement. The love and support from our parents and our families are greatly appreciated. I would also like to thank the Ugandan community in Canada for their friendship and support over the years.

To Jimmy and Jordan,
and to my mother and my father

Contents

List of Figures

List of Tables

List of Algorithms

Notation

Symbols employed only locally are defined at the point of need. The following conventions are global:

N	number of beams. In Chapter 4 and Chapter 5, N refers to frames.
M	number of users (or terminals or sources)
W	bandwidth per beam
λ	mean number of users per unit area. In Chapter 5, λ_x is the number of good channels among users with backlog x
Λ_k	mean number of users in beam k. In Chapter 5, Λ is the total number of good channels at start of a frame.
d_j	Euclidean distance between center of block j and center of the beam containing block j
A_k	set of indices of terminals allocated to beam k
M_k	cardinality of A_k
L	vector of beam center locations, $l_1, l_2, \ldots, l_N$
$F(L)$	also represented as $f(d, \Lambda)$ is a function of the beam sizes and user distribution
U_{min}	smallest (mean) number of users acceptable for any beam
μ_{min}	penalty factor associated with violation of U_{min}
U_{max}	largest (mean) number of users acceptable for any beam
μ_{max}	penalty factor associated with violation of U_{max}
φ	weighting factor
d_{th}	threshold distance away from center of beam
θ_d	penalty incurred for exceeding threshold distance
γ	signal–to–interference ratio (SIR) target or threshold
R	transmission rate
P	transmission power
p	maximum transmission power, sometimes denoted as P_{max}
$\alpha_{\ell k}$	attenuation coefficient of signal from block ℓ received in beam k
σ_k^2	satellite receiver noise at beam k

ρ_i^*	average rate threshold for user i
K_i	length of rate-averaging window for user i
$\rho_i^{(K)}(N)$	average rate at N for terminal i with averaging window K
$\mathcal{F}(\rho_i^{(K)}(N), \rho_i^*)$	function relating $\rho_i^{(K)}(N)$ to the rate threshold, ρ_i^*
X	system state space. The state at timeframe n is denoted as $x(n)$
u	a control policy
$c(x(n), u(n))$	one-stage cost function
$H(n)$	rate history at time frame n
B	buffer size
r_0	buffer input rate
θ_i	rate weighting factor for user i
a_i	relative weighting factor for user i
$\mathtt{R}$	sum of transmission rates
μ_x	number of users with backlog x at start of a frame

Chapter 1

Introduction

Due to its wide coverage area, a satellite network provides opportunity for ubiquitous service. Traditionally, satellites have been used for fixed services to provide 'bent-pipe' interconnections between terrestrial networks or end users (via the use of very small aperture terminals – VSATs) [1]. Due to the emergence of Internet and multimedia services, satellite systems have evolved to support broadband networking for high speed data transmission, as well as intelligent switching and routing via on-board processing (OBP) [1–9].

Field trials have demonstrated that the emerging demand for broadband interactive multimedia services can be addressed via satellite communications at the Ka band (30/20 GHz) and beyond [10–13]. The key advantages of operation at the Ka band are increased availability of bandwidth and the potential for smaller, less costly terminals and greater mobility [13]. Large and cumbersome mobile terminals needed to make a call were a previously citepd drawback against geostationary satellite systems ever achieving mass market penetration [14, pp. 321]. The major disadvantages of Ka band operation are increased rain and atmospheric attenuation. The objective of Ka band operations is to capitalize on the increased availability of bandwidth via appropriate resource management — without compromising the QoS requirements of the network.

1.1 Key considerations for broadband satellite networking

Rapidly growing demand for interactive multimedia services, including web browsing, bulk data transfers and video services, provides impetus to the development of versatile broadband networks capable of providing low cost direct-to-user services to fixed and mobile users irrespective of location [10, 15–19]. A network supporting such users must necessarily combine satellite and terrestrial infrastructure in order to provide access in highly diversified operating conditions, including open/shadowed, rural/suburban/urban, and indoor/low-range environments. Satellite systems offer a number of particular advantages. Satellite systems are especially efficient for multimedia broadcasting and for mobile users, particularly those on ships, planes and in remote places. Satellite systems can provide global multimedia services to end users, long before such services would be available through terrestrial networks. In many regions, terrestrial broadband networking is not commercially viable.

Increases in user population challenges both the network capacity and the server capacity. This has led to the emergence of multicast systems which are able to deliver the same information to a set of users via a single transmit operation thereby saving network capacity and reducing server requirements. In general, wireless networks are a natural medium for multicast delivery. In particular, satellite networks offer very wide coverage and, typically, one-hop infrastructure [15, 18, 20]. For service within an existing coverage area, a satellite delivery system is insensitive to the number of users; the number of users can generally be increased without any need to increase the downlink bandwidth.

By the late 1970s, it was recognized that Ku (14/12 GHz), C (6/4 GHz) and lower satellite frequencies in use at the time were becoming congested and would not provide sufficient bandwidth to meet anticipated demand for current and emerging services [10]. It was this that led to re-consideration of the Ka band, which had hitherto been ignored due to the severe rain attenuation suffered by signals at Ka frequencies. The suitability of the Ka band and of on-board processing (OBP) was conclusively shown by the ITALSAT (Italy) and ACTS (United States) satellite programs. Based on multibeam satellite systems with onboard processing, the programs proved that telecommunications requirements such as integration of digital services, assignment of resources on demand, and low-cost, handheld terminals, can be met by a single satellite equipped with OBP capabilities (demodulation, switching, remodulation) and with terminals that feature antennas small enough to be located on user premises or to be moved to remote regions at moderate cost.

The significance of on-board processing (OBP) is that the satellite uplink and downlink can be optimized separately with respect to multiplexing, modulation and error control. This enhances the signal quality as well as the bandwidth and power efficiency of the system [15, 21, 22]. Furthermore, a constellation of satellites can provide full space connectivity through the use of intersatellite link technologies at EHF (60 GHz) and/or optical frequencies. Intersatellite links also present the opportunity to bypass terrestrial networks, a possibility that is of particular interest to government and military users [15]. Challenges with OBP implementation include the lack of flexibility in adapting to various types of terminals, modulation and coding, as well as the requirement for demultiplexing at the smallest level of the individual channel in the case of functions that are provided at channel level (such as gain control and routing) [22].

In addition to processing, a key onboard technology for broadband satellites is antenna beam allocation [10, 23, 24]. Due to inhomogeneous distribution of users across the coverage area, onboard beam allocation is key to meeting the quality of service requirements for all active users. The allocation of beams independent of user distribution results in beams that require very high transmit powers relative to beams that are allocated taking the user distribution into account [25]. Onboard beam allocation allows for a beam configuration in which the power allocation per beam can be optimized with respect to the user distribution as well as the prevailing channel conditions.

The high frequencies of Ka band operation enable the generation of beams that are of high gain and small beamwidth. The high-gain spotbeams mean that user terminals can be small, low cost, handheld devices with small low–gain antennas suitable for mobile and/or remote operations [10, 23]. On the other hand, multibeam satellite antennas are required in order to provide service over a wide coverage area. Multibeam antennas are inherently more complex than single-beam antennas, and issues to be addressed range from beam architecture, beamforming complexity and reflector design complexity. A multibeam antenna can have fixed or adaptive beams.

Adaptive multibeam antennas improve coverage, capacity and switching efficiency, and are typically smaller in size. However, adaptive antennas typically require more complex system setup and beam management. Significant research has gone into the development of ‘smart antennas‘ capable of steering the antenna radiation power over the desired coverage area [23, 24, 26–32]. Examples of such antennas are active phased array antennas and digital beamforming antennas. In particular, digital beamforming is considered the future for the realization of multibeam antennas capable of generating several spotbeams with the ability to support broadband satellite applications. Beamforming via a digital signal processing unit increases the flexibility of the beam generation process compared to beamforming via conventional dynamic phased array systems which require complicated multibeam architectures [33–36]. With current technology, digital beamforming is only feasible for low bandwidth applications (< 500 MHz) [36], but we assume that high interest in broadband multimedia networks will continue to drive the development of beamforming technology.

An important design consideration for satellite antenna systems is the tradeoff between receiver antenna size (which determines the gain) and the modulation/coding scheme [16]. The objective is to achieve a cost-effective balance between three factors: the transmission bit rate, the earth terminal cost and performance, and the satellite segment (beam bandwidth and output power). Modulation schemes must be robust (to deal with rain attenuation, in particular) and of high bandwidth efficiency [15]. With appropriate channel coding, the probability of bit error can be significantly reduced such that a specified bit error rate can be attained with a smaller transmit power. Ever-more-efficient coding, modulation and detection (such as multi-user detection (MUD)) techniques mean that the cost/benefit tradeoff rules in favor of a cost-effective balance between the three factors [15, 16, 22]. Satellite systems typically use phase modulation schemes which offer high bandwidth efficiency (and in particular, QPSK which offers robust channel protection too), and use convolutional codes for channel coding.

Perhaps the biggest obstacle in deploying satellite systems is arranging the necessary finances [37]. For the first three decades of their use, satellites enjoyed a cost advantage over undersea cables – especially on long routes [17]. However, that changed with the advent of undersea fiber-optic cables. Nevertheless, given their broadcast capabilities, satellites remain the exclusive choice for television distribution both within and between countries. Reductions in the cost of terminals (typically less than $1,000) and improvements in resource management assist in bringing down the cost of operation [10, 22] and in making satellite networking an attractive research topic.

1.2 Satellite communication systems

Satellite communication systems can be classified as either geostationary or nongeostationary depending on their orbit of operation. There are three typical orbits of operation: Low earth orbit (LEO), Medium earth orbit (MEO) — sometimes referred to as the Intermediate circular orbit (ICO), and Geostationary earth orbit (GEO). Table 1.1 presents distinguishing characteristics in terms of altitude above the earth, approximate orbit period and round trip propagation delay. For the three orbits, the relation between the orbit period, T, and the satellite altitude, h, is given by [15, pp. 19]:

$$h = \sqrt[3]{\mu \left(\frac{T}{2\pi} \right)^2} - R_e,$$

where μ denotes the product of the universal gravitation constant and the mass of earth, and $R_e = 6,378$ km is the mean equatorial radius.

Table 1.1 Distinguishing characteristics of satellite orbits of operation. *Source: [7], [15, pp. 19]*

Orbit	Altitude (km)	Orbit period (hr)	Propagation delay (ms)
Low earth orbit (LEO)	500 – 1500	2	10
Medium earth orbit (MEO)	5,000 – 10,000	4 - 6	100
Geostationary orbit (GEO)	35,786	24	250

In any orbit, the nature of the propagation delay is a key issue in the design of satellite networks. GEO-based systems have long but constant propagation delays while nongeostationary LEO systems have short but highly variable delays due to their lower altitude [7, 15]. In addition, the maximum visibility of an overhead satellite is directly proportional to the orbit period. Thus, non-geostationary systems have to address issues of satellite handover, while geostationary satellites appear to be stationary to an earth terminal which requires little or no terminal antenna steering.

This thesis focuses on Ka band GEO-based satellite systems. It is estimated that there are more than 50 proposed Ka band satellite projects worldwide [17]. The overwhelming majority of these are for GEO-based systems, requiring about 270 geostationary orbit locations. Most of the proposals are targeting national or regional coverage. Table 1.2 presents system parameters of Ka band proposals from North America and Europe [5, 15, 17, 38, 39]. In the case of Astrolink and Teledesic, the notation 192 (64 HBs) and 576 (64 HBs), respectively, denotes the number of ground cells that are to be covered by the hopping beams.

All systems listed in Table 1.2 are proposed for the GEO orbit, except for Teledesic which is proposed for the LEO orbit. As noted earlier, financing is a critical issue for satellite deployment. For example, in May 1997, the US Federal Communications Commission issued twelve Ka band licenses — but by mid-1999, only five companies had announced significant funding and/or major construction for their satellite projects [37].

Table 1.2 Proposed Ka band system parameters. *Source: [5, 15, 17, 38, 39]*

System	Beams	Transponder bandwidth (MHz)	Beam Size	Capacity (Gbps)	ISL capacity
Anik F2	27 + 2 RB + 1 NB	180 36 (RB,NB)	$0.7^o \times 0.3^o$	≈ 10	None
Astrolink	192 (64 HBs) + 3 FB + 1 SB	125	1^o	9.6	1 Gbps
CyberStar	27	125	$1^o - 2^o$	4.9	1 Gbps
Echostar	24 + 1 GB	120	1^o	5.8	120 MHz
EuroSkyWay	32 (Extended Europe), 46 (North Atlantic)	170 (up), 232 (down)	$\approx 1^o$	9	56 - 64 GHz
Galaxy-Spaceway	48	125 (Ka), 46 (Ku)	$1^o - 3^o$	4.4	1 Gbps
GE*Star	44	243	$< 1^o$	4.7	None
KaStar	24 + 2 SWB + 1 GB	125	1^o	7.5	155 Mbps
Millennium	32	250	1^o	5.2	1 Gbps
Morning Star	7-10	50 (Ka), 24 (Ku)	1.6^o	0.5	None
NetSat 28	1,000	150	$\approx 0.2^o$	772.0	None
Orion	25 + 2 SBs	114	1^o	2.9	TBD
PanAmSat	12 SBs	54	$1^o, 2^o, 3^o$	1.2	None
Teledesic	576 (64 HBs)	396	"Small"	13.3	1 Gbps
VisionStar	5	40	$2^o - 5^o$	1.9	None
VoiceSpan	32 or 64	120	1^o	5.9	0.5 Gbps

FB: Fixed Beam, GB: Global Beam, HB: Hopping Beam, NB: National Beam
RB: Regional Beam, SB: Steerable Beam, SWB: Switched Beam, TBD: to be determined
ISL: Intersatellite Link

1.3 Issues for Ka band satellite communication systems

A primary concern with the high-frequency bands is the effects of rain attenuation, a particularly severe problem in tropical areas. Based on insight from two-years worth of Ku band rain fade data in three tropical countries, Cameroon, Kenya and Nigeria, it was noted that the numbers of rain fade events and total down time in Cameroon and Nigeria were significantly greater than the limits for most direct-to-home multimedia satellite services [40]. Rainfall tends to be highly localized and so the Ku band study recommended site diversity as a means of ensuring reliable communication service to all users. Where site diversity is not available or appropriate, larger receiving antennas or lower availability were proposed.

This thesis focuses on effective resource management as a means of addressing the spatial and temporal variation in both the propagation conditions and the user distribution. Specifically, we consider resource management that modulates the operation of the satellite network based on state feedback about the user distribution and the prevailing network conditions. The network control is implemented via control of network-layer parameters such as the transmission rate and power. In this case, it should be noted that the satellite network capacity is affected by physical-layer parameters such as modulation and coding.

A prerequisite for efficient network control is timely and accurate state feedback. Consequently, there are two issues that make the problem of resource management in satellite networks particularly challenging. The first challenge is the tradeoff between control and delay. The long round trip time (RTT) of about 250 ms for GEO networks can be an impediment to useful communication and result in inadequate and/or conservative control. For example, as typical for terrestrial networks, it is expected that broadband satellite networks will be supported by the Internet TCP/IP protocol suite and the ATM protocol architecture [15, 41]. The TCP flow control is a function of the window size and the round trip time. TCP transmission begins with the transmission and subsequent acknowledgment of one datagram, then two, four, etc., until the limit imposed by the maximum window size. Because of the long propagation delay in GEO satellite networks, this slow-start algorithm fails to achieve the maximum throughput achievable in networks with smaller delays. Furthermore, the satellite throughput is limited by the maximum window size permitted by the TCP protocol which is much less than the equivalent window size represented by the GEO satellite RTT.

The tradeoff between the error rate and delay is also challenging. For example, TCP ensures correct delivery of all datagrams by retransmitting datagrams for which it did not receive an acknowledgment. The long satellite RTT does not provide a sufficient window of opportunity to prevent unnecessary data retransmission in case of delayed acknowledgments. Aside for an erroneous assumption by TCP that a delayed acknowledgment is due to transmission error, even retransmission when needed only serves to exacerbate the delay for successful data transmission.

Nevertheless, significant research has gone into proposing changes to TCP/IP to overcome its limitations when operating in a satellite environment [15, 20, 41–44]. Proposals include an increase of the TCP window size and alternative methods for generating and responding to acknowledgments. In essence, the proposals seek to tailor the TCP operation taking into account the long delays characteristic of GEO satellite networks. Similarly in this thesis, we seek to develop resource management techniques that take into consideration the differing timescales of network operation and of user QoS requirements.

In this thesis we focus on the control of the uplink transmission from users to satellite, subject to the assumption that the uplink bandwidth is fixed and known. The primary issue to be addressed is multiple user access. Code Division Multiple Access (CDMA) has been proposed as the mode of access for next generation satellite networks [12], and it is what we consider for this thesis. Features driving the interest in CDMA include universal frequency reuse which eases the process of resource allocation, robust interference mitigation and flexible support for a broad range of services [45].

The satellite uplink and downlink experience different limitations. On the uplink, the signal for each of the active terminals suffers interference due to simultaneous transmission of other users. Increasing the transmission power of any of the terminals serves to increase the interference experienced by the signals of other active terminals. Hence, the uplink is interference-limited because even without constraints on the transmission power, the interference effect on other terminals places a bound on the transmission power of any given terminal. Conversely, on the downlink transmission from satellite to terminal, the interference experienced by each terminal is primarily due to the prevailing channel conditions which are generally location-dependent. To overcome poor channel conditions, the satellite could increase the power transmitted to affected terminals. However, the sum of the transmission powers to all active terminals is bounded by the satellite's maximum output power. Hence, the downlink is power-limited.

Just as the limitations differ, so too do the control strategies for the uplink and the downlink in order to achieve the respective QoS criteria. On the uplink, the control objective is to enable equitable system access among the terminals while minimizing the impact of interference due to simultaneous user transmission. On the downlink, the control objective is to improve the system throughput within the bounds of the system output power. In a GEO-based network, the system capacity is typically limited by the multiple-access uplink [45], and hence the focus in this thesis. It is however expected that multimedia applications will require an increasingly larger downlink throughput and works such as [46–50] focus on management of the downlink resources.

This thesis is also motivated by the reality that many developing countries and rural/remote areas of developed countries simply lack the terrestrial infrastructure to support communication needs. Challenges to addressing rural communications include lack of regular electrical power, difficult terrain, cost of extending terrestrial transmission systems, and the "low return on investment" [51]. Table 1.3 presents telecommunications data on three East African countries [52]: Kenya (1999), Tanzania (1999) and Uganda (1995) — where the number in brackets indicates the year in which terrestrial wireless service was introduced.

Table 1.3 Telecommunications data for East Africa (March 2003) *Source: [52]*

	Kenya	Tanzania	Uganda
Geography (sq. km)	582,650	945,087	236,040
Population	30,765,916	37,187,939	24,699,073
GDP (US$ per capita)	340	270	280
Telecoms fixed lines	354,000	230,000	55,000
Telecoms mobile	1,654,000	703,000	510,000
Telecoms operators	3	5	3

In the table we can observe that, while cellular operations are relatively recent, there are significantly more cellular subscribers compared to available fixed lines. For many areas in these countries, the terrestrial wireless network presents the only access to telecommunication services. Hence, a combination of satellite and terrestrial wireless networking should enable these countries achieve better communication access within and between themselves.

In fact, by the early 1960s, the role of satellites in enabling ready communication across Africa was noted as a number of African countries were gaining their independence [53, 54]. This role was driven by the need to foster closer political, cultural and commercial ties among regional neighbors in addition to the communication links countries had to their former colonial powers. In recent times, satellite networks remain highly valued in terms of the wide coverage and easy access to rural and to remote regions of Africa, which is where the majority of Africans reside [51, 55, 56]. For example, by year-end 2003, it is expected that there will be an internet point of presence in each of Uganda's 56 districts [57, 58]. Continent-wide, the New Partnership for Africa's Development (NEPAD) aims to utilize information and communication technologies to extend and to raise the quality of education at all levels in African education systems [59]. In both cases, internet connectivity will be achieved via satellite links — particularly for the rural areas.

A number of international satellite operators do have footprints over parts of Africa, for example, Intelsat and PanAmSat. In addition, the Regional African Satellite Communication Organization (RASCOM) has commissioned the building of the first Africa-wide satellite system that is expected to deploy in 2006. Comprised of 12 Ku and 8 C bands, the system will provide fixed voice, data communications, internet access, and satellite broadcasting services [60].

In the following sections, we describe the system model and key assumptions made in this work. We then describe the CDMA satellite resource management problem addressed in this thesis and present a summary of the key contributions. Finally, the thesis outline concludes this chapter.

1.4 System model for CDMA-based satellite network

This thesis considers resource management on the uplink of a multibeam CDMA satellite system. The key control parameters are the beam coverage geometry as well as the transmission rate and power. The beam geometry is subject to control based on the geographic user distribution. We assume that the desired beam configurations can be realized via the use of 'smart antennas'. The bandwidth per beam is fixed and so is the CDMA chip rate. Variations in transmission rate are achieved via changes in the processing gain.

The satellite system provides coverage to several earth terminals that, in turn, support a variety of applications. User data is transmitted in synchronized time frames of constant period. Each frame is assumed to be larger than the user data duration by a guard-band interval. Guard bands are needed to compensate for arbitrary delays caused by signal propagation delays or clock drifts.

For each terminal, the instantaneous transmission rate and power are subject to control via an access rate schedule that is based on prevailing channel conditions and the quality of service criteria. The transmission rates are described by a traffic model that is based on three key assumptions. First, we assume that each terminal is capable of transmitting at whatever rate it is allocated; the allocated rates vary continuously. Secondly, there is no minimum rate requirement, and a terminal may have its transmission deferred at any time. Thirdly, we assume that the traffic is infinitely divisible, a characteristic of the fluid model for traffic. The fluid model is an ideal model in which it is assumed that data packets are not transmitted as entities but as infinitely divisible units of the packets [61].

Based on the traffic model assumptions, each active terminal will transmit at the exact rate allocated to it. This consideration simplifies the control algorithm in that the satellite resource management unit does not need to keep track of the *actual* rate used by a terminal, and how this differs from the *allocated* rate — since we assume them to be the same. While fluid models are ideal, they are expected to become even more useful in the evolution toward higher capacity data networks [62]. The concept of fluid in such networks is based on the assumption that the most important dynamics in high capacity networks depend on how aggregates of packets are processed and not on how individual packets are processed. In such networks, the packet size would be only a small fraction of typical buffer capacities in the network.

Network users experience variable channel conditions depending on factors such as location and number of simultaneous user transmissions. Hence the network control is constrained by a lower bound on the signal-to-interference quality of service expected by each user. While we consider the channel to be time-varying in nature, we assume that channel variations over a given time frame are small such that channel characteristics are treated as constant during each time frame. In addition, we assume that channel characteristics for each time frame are known at the start of the frame. This knowledge can be gained, for example, by measurements of a reverse channel in a duplex system, explicit feedback from the receiver, or from measurements of a pilot signal. Note that user mobility is not considered in this work. Even then, we can assume that in the case of low user mobility, channel conditions vary only slowly during each time interval and so can be considered effectively constant. Low user mobility would be exhibited by pedestrians, for example.

In a GEO satellite system, perception of change and reaction to it takes at least a round-trip propagation delay of 250 ms. This certainly poses a challenge for time-sensitive performance objectives during the operation of the satellite system. Table 1.4 presents quality of service requirements for a variety of applications in terms of the delay tolerance and the uplink and downlink data rates [5, 10]. The applications can be broadly divided into three types: (1) bi-directional messaging, (2) retrieval, and (3) conversational.

Table 1.4 Service requirements for a variety of applications. *Source:* [5, 10]

Type	Application	Max end-to-end delay	Downlink data rate	Uplink data rate
(1)	E-mail	5 min	1–5 kbps	1–5 kbps
	Paging	5 min	1–5 kbps	1–5 kbps
	PC networking	200ms	64 kbps	64 kbps
(2)	Database access	500 ms (file transfer)	2 Mbps	100 kbps
	Web browsing	500 ms	64 kbps	1–5 kbps
(3)	Telephony	250 ms	64 kbps	64 kbps
	Video Telephony	200 ms*	64 kbps – 1 Mbps	64 kbps – 1 Mbps
	Video conferencing	200 ms*	64 kbps – 2 Mbps	64 kbps – 2 Mbps
	Telemedicine	200 ms*	64 kbps – 2 Mbps	64 kbps – 2 Mbps
	Tele-education	200 ms* (1s for data)	1 Mbps	64 kbps

* According to user perception, the quality of conversational services does not degrade appreciably if transmission delay does not exceed ≈ 280 ms

This thesis focuses on delay-tolerant applications with an *average–rate* specification in which the averaging is with respect to a time horizon of prescribed length. We seek to control the network performance by exploiting timescale differences within the network. There are several time constants in play: the time constant associated with the satellite environment and related to the rate at which the channel varies; the time constant associated with the control, determined by the propagation delays; and a time constant associated with user quality of service — the rate-averaging horizon. The rate-averaging horizon is comprised of a consecutive set of the fixed length time–frames.

Remark. *The frame length is introduced for computational convenience, and perhaps also to reflect that practical systems are typically frame oriented. It is important to note the distinction between frame length and the length of the rate-averaging window, which is an integer multiple of the frame length.*

There is a tradeoff between the length of the rate-averaging horizon and packet delay. The longer the horizon, the greater the flexibility in exploiting time elasticity in improving network efficiency. However, the longer the horizon, the greater the potential for excessive delay. We do not explicitly deal with user delay.

In the following section we describe the resource management problem addressed in this thesis.

1.5 Resource management in CDMA-based satellite networks

Satellite network operations span a wide range of time scales. In general, we can consider a network to operate at three timescales: long-term, medium-term and short-term. Long-term operations, such as beam allocation across the coverage area, are based on the geographic distribution which holds steady over long time periods. Medium-term operations include fade countermeasures in response to rainy periods which can last for hours in some regions. Thirdly, short-term operations, such as call admission and scheduling, are in response to rapid fluctuations in traffic demand at call or burst level, user mobility, or channel fading. Resource management in satellite networks necessarily involves the same wide range of time scales. At network-level, the resources of interest are the bandwidth, the transmitter power and the energy. In turn, the resource management performance is measured in terms of quality of service parameters such as the data rate, call blocking and cell or frame blocking.

The choice of CDMA as the mode of access has a particular role in the formulation and solution of the resource management problem. CDMA networks are said to have ‘soft‘ capacity, that is, capacity is limited by the number of terminals permitted to transmit simultaneously rather than by the availability of CDMA codes. This limitation, as noted earlier, is due to the interference caused by the transmission power of any given terminal to other active terminals thereby impairing the signal quality received at the satellite. Consequently, the quality of service parameters of interest in the CDMA satellite network we consider are the rate and the signal-to-interference – two numbers per active terminal. Note that the one-to-one relationship between the SIR and the bit error ratio can be derived given the parameters of the transmission technology such as the channel coding and the modulation.

We address two resource management problems — at two different time scales. In particular, we consider beam management (long timescale) and cell/frame-level access control (short timescale). We assume that the effects of severe rain attenuation are mitigated via beam management, access scheduling and a lower availability specification. The overall resource management objective is to design the beamwidths and power allocation to achieve particular rate and signal-to-interference ratio vectors subject to control of the beam geometry across the coverage area and the assigned transmission powers. The challenges to be addressed include inhomogeneous user distribution, power constraints at the satellite and earth terminals, time-varying channel conditions, and long propagation delays.

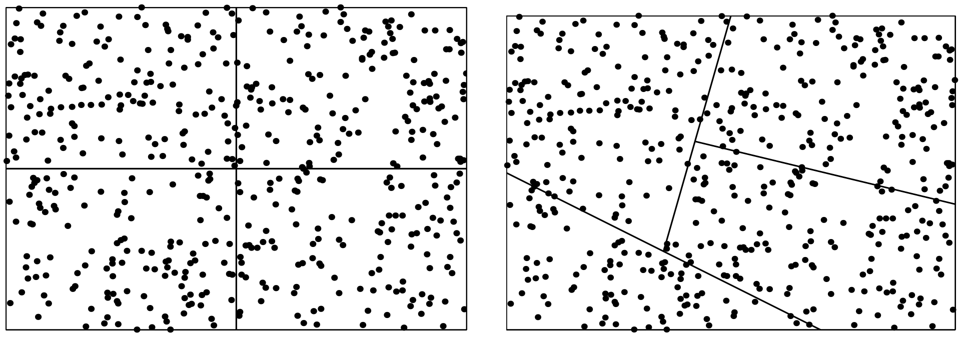
In the following subsections, we describe the beam management and access control problems addressed in this thesis. Work related to these problems is reviewed in Section 2.3 and Section 2.4 respectively.

1.5.1 Beam Management

We consider beam management that seeks to enhance network performance via an efficient distribution of users among a fixed number of beams. This distribution is achieved by tuning the shapes and sizes of the spot beams to reflect user distribution. Traditional beam configuration is of equal-sized, typically circular or elliptical, beams irrespective of user distribution. While network capacity may be dynamically allocated among the beams [4, 5, 39], the level of control and the achievable throughput are constrained by the fixed-beam geometry [25, 63].

Figure 1.1 depicts a four-beam allocation scenario for a uniform and for an adaptive beam configuration. Based on the given number of beams and the terminal distribution across the coverage area, the objective for beam configuration management is to increase the achievable rate region relative to a uniform beam allocation strategy. In either case, uniform or adaptive, the network performance is constrained by signal-to-interference targets, limits on the transmission power per terminal, and prevailing channel conditions. The achievable rate region can be defined in terms of maximum system throughput, average user throughput, and number of users permitted to transmit. A strategy based on average user throughput is generally better in terms of allowing equitable access for all system users.

In addressing the beam management problem, the first task is characterization of the user distribution. There are two approaches for this characterization: statistical and deterministic. Statistical data is easier to come by in practice, and is more stable in the sense of less susceptible to significant time variation. We assume that the user distribution across the coverage area is governed by an inhomogeneous spatial Poisson process, characterized by the feature that populations of disjoint regions are statistically independent. To each



(a) Uniform beam allocation

(b) Adaptive beam allocation

Fig. 1.1 Satellite beam configuration

unit area, the Poisson process ascribes a parameter, λ, denoting the mean number of users. The Poisson process is a standard benchmark for spatial patterns [64], and is based on three key assumptions. First, over the coverage area, the numbers of users in non-overlapping subregions are independent of each other. Secondly, the distributions of the number of users in subregions of equal area are identical. Thirdly, the probability of two or more users in a very small subregion is negligible.

Given the user distribution, the challenge is then how to generate the beam pattern over the geographic area. We address this problem by mathematical modeling and analysis of the allocation of satellite beams subject to user distribution. To ease the computation complexity, the coverage area is quantized into a grid of user locations and a beam allocation function specified to determine the configuration cost associated with a given allocation of terminals among the beams. The resulting model is a nonlinear programming problem comprising an objective function to minimize the sum of weighted configuration costs due to individual beam sizes and due to the mean number of users per beam. To solve this problem, we develop beam allocation algorithms that incorporate the Hooke & Jeeves nonlinear optimization algorithm which has been observed to provide robust solutions in a similar basestation allocation problem for a terrestrial wireless network [65]. The Hooke & Jeeves algorithm requires several iterations, however, a well-selected initial state and termination criteria can ease the computational complexity. The complexity of the problem is also eased by the grid-structure utilized, which reduces the granularity of the coverage area.

Beginning from a uniform beam allocation, which we take as the initial beam configuration, the proposed algorithms iteratively generate an improving direction and identify an optimal step length along this direction to arrive at a revised solution. This process continues until the termination criterion is satisfied. The achievable rate region of the resulting beam configuration is then compared to that of the uniform beam configuration.

1.5.2 Cell/Frame-level Access Control

Given a beam configuration, cell/frame-level access control is used to generate an access rate scheduling policy that is subject to an average-rate requirement for each terminal. We consider network operation such that the instantaneous transmission power is subject to an upper bound, and that the instantaneous transmission rates are continuously variable and subject to network control. The scheduling policy exploits time elasticity, as reflected in the length of the window that defines the averaging operation, to enhance bandwidth utilization. Such utilization enhancement is achieved by judicious allocation of the transmission rates at each time frame so as to minimize the long-run energy per bit requirements for each terminal while ensuring that the average-rate targets are achieved. Figure 1.2 presents an access rate allocation for five terminals over a window that is seven timeframes in length.

At the beginning of each time frame, the control selected is based on state feedback about previous rate allocations and prevailing channel conditions. This thesis differs from the literature [66–69] in respect of the use of such feedback, as a result of which, the user-perceived rate and the actual transmission rate are distinguished by different time constraints. While the channel capacity seen by the terminals is time-varying and stochastic, we assume that the channel conditions are known at the start of each time frame.

For a given window length, the challenge then is how to generate an optimal access rate policy that will achieve the average-rate targets with minimal energy-per-bit requirements. The setting of this access control problem is one in which dynamic programming (DP) provides a natural approach. This is because, at each time frame, the controls are selected so as to minimize the long-run system cost of operation given the uncertainty of future channel conditions — whereby the system cost is characterized by failure to achieve quality of service targets. However, the computational complexity of DP solutions rises as the state space increases. We provide exact DP solutions for a simple two-user network and use these

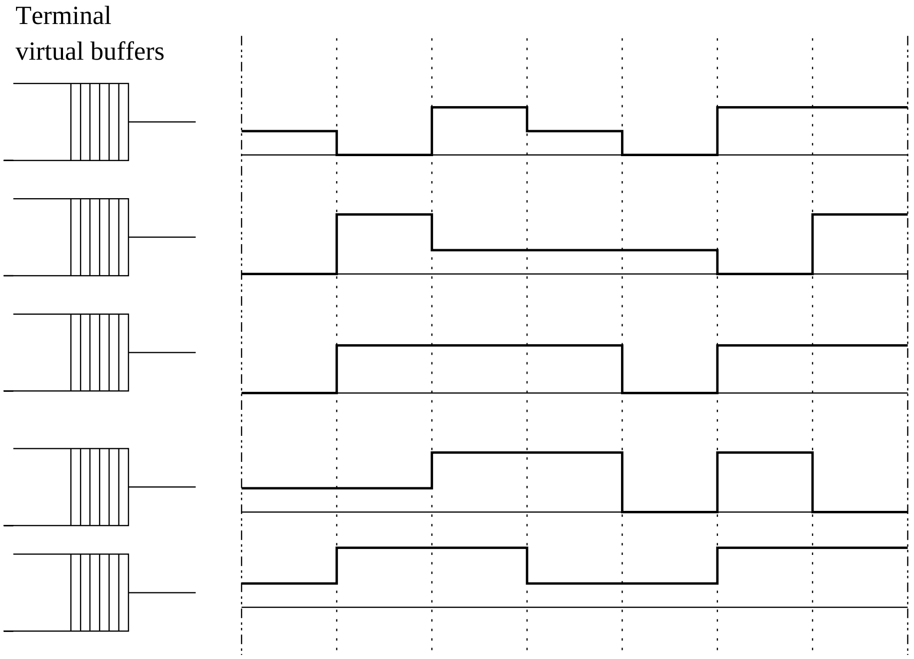


Fig. 1.2 Access rate allocation for five terminals over one window

solutions in the development of approximate solutions for a network of many users.

We are particularly interested in understanding the extent to which temporal elasticity can yield a throughput advantage. The idea that there should be an advantage is made evident by comparing the rate/power tradeoff in two extreme cases: infinite-window averaging versus single-frame averaging.

Remark. *While the problem of designing the beam shapes in a satellite network seems on the face of it similar to that of assigning users to base stations in a terrestrial cellular network, in fact there is a significant difference. In the cellular terrestrial setting, every base station hears every terminal, albeit at power levels that depend on the locations of base stations and the terminals. In principle, at least, any terminal can be assigned to any base station. On the other hand, in a satellite network in which inter-beam crosstalk can be neglected, each satellite receiver hears only those terminals found within its beam. Because beam geometries are constrained (for example, to be simply connected), the assignment of terminals to satellite receivers has geometrical constraints that are absent in the cellular case.*

The principal contributions of this thesis are reviewed in the following section.

1.6 Summary of contributions

We consider resource management for a CDMA-based GEO satellite that provides service to delay-tolerant users. We focus on two specific problems:

- Beam management - Given a nonuniform user distribution, we are interested in beam management techniques that yield roughly equal numbers (or mean numbers) of users per beam in cases where the user SIR thresholds and limits on transmission power are homogeneous. We are also interested in the impact of such techniques on the achievable rate region, compared to the achievable rate region for beam management techniques that ignore the geographical distribution of users.
- Rate allocation - We are interested in the tradeoff between transmission rate and transmission power that is achieved by exploiting temporal elasticity in user quality-of-service requirements.

The following summarizes our contributions.

Beam Management:

1. We develop two beam allocation algorithms that seek to distribute users equally among beams by beam shaping. The algorithms take as input the coverage area and state feedback on user distribution, and act to optimize an objective function that is a weighted tradeoff between the beam size and the number of users per beam. The smaller the beam size, the higher the gain, and hence the less power required per terminal.

2. The key elements of the two algorithms developed are a beam allocation function and the Hooke and Jeeves search method. In the first algorithm, the beam allocation function distributes terminals such that a terminal is assigned to the beam whose center is the shortest Euclidean distance away from the terminal. The Hooke and Jeeves method is used to search across the coverage area in search of an improving direction with respect to objective function. The second algorithm is built on the same principle. However, in addition, vector quantization techniques are applied at

each beam allocation stage to minimize the separation distance between terminals and their beam centers. The allocation performance of the algorithms developed is measured in terms of a Beam Variability Factor (BVF). The BVF is defined as the ratio of the standard deviation to the mean of the vector of mean users per beam. The more evenly balanced a beam configuration is, the smaller the BVF. The two algorithms outperform the uniform beam allocation. However, the second algorithm (incorporating beam size adjustment) provides only marginal improvement in the BVF relative to the first algorithm.

3. We compare the performance of a uniform and an adaptive allocation in terms of the *achievable rate region* and the *minimum power* required to achieve a given throughput. This thesis departs from the literature on the allocation of variable-beams in that we consider the transmission rates to be variable and controllable [10, 25, 63]. Relative to a uniform beam allocation, we show that adaptive (BVF-driven) beam allocation results in increased system throughput and achievable rate region. The adaptive scenario may require a higher power allocation because more terminals are permitted to transmit.

The allocation algorithms developed can be adapted to a variety of network performance criteria by changing the form of the objective function.

Rate allocation:

1. We formulate the problem of exploiting possible elasticity in user quality-of-service constraints for the purpose of augmenting CDMA satellite system capacity. Supposing that QoS is constrained by SIR thresholds, limits on transmission power and per-user average rates, where the averaging is over sliding windows of fixed time length, we construct algorithms that assign instantaneous rates so as to minimize the average power over users as well as over time.

2. We provide an analytic comparison of two extremal cases, corresponding, respectively, to a short averaging window and an infinite averaging window. The results quantify the potential power savings that can be had in taking advantage of user insensitivity to short-term power fluctuations.

3. In generating the rate-allocation policy, we develop a Markovian model of the access control system so as to bring the optimal rate allocation problem within the purview of Markov Decision Theory and Dynamic Programming (DP). To reduce the DP solution complexity, we propose a queue-theoretic approximation to the access control model in which (fictitious) terminal backlogs are used to represent the rate history of the system. Two cost structures are considered for the backlogs Markovian approximation: data volume lost and the probability of buffer overflow. The performance of the optimal rate-allocation policy for a simple two-user network is then compared to that of short-horizon and infinite-horizon policies. We show that temporal elasticity in the rate allocation process results in increased average throughput per user while minimizing the long-run energy per bit requirements. Specifically, it is seen that increasing the buffer size — thereby increasing the time elasticity of the access-rate control — reduces the average transmission power required to support a given rate allocation.

4. While DP techniques do yield an optimal solution, the complexity involved does not make such techniques a viable option for a general network with many users. Based on insights from the optimal rate allocation, we develop heuristic scheduling mechanisms that guide the rate allocation based on the prevailing channel conditions and the current system state. Performance evaluation reveals the same trends as obtained for the optimal two-user scenario in that time-elastic control does yield savings in the long-run average energy per bit per terminal.

5. We rework the DP formulation of the queueing model, noting that where users are homogeneous in terms of QoS requirements, there is opportunity to reduce the size of the state space. The reduction enables us to extend the optimization described in Contribution (3) to networks with more than two users. The results confirm that average power diminishes with increasing size of the rate-averaging window.

Remark. *The effect of propagation delays on the implementation and evaluation of the controls we propose requires additional research. The issue is not taken up in this thesis, not because it was deemed unimportant (on the contrary!), but because we had focused on other issues and time ran out. It is an obvious candidate as a topic for continuing work.*

1.7 Thesis outline

The remainder of this thesis is organized as follows. In Chapter 2 we describe key features of the satellite network architecture, satellite access and QoS support, and satellite channel model of a Ka band GEO network. In addition, we review beam management and cell/frame-access control techniques that have been proposed to improve the efficiency of broadband multimedia wireless networks that are capable of supporting a variety of traffic classes with different QoS requirements.

This is followed by Chapters 3– 5 in which we address the problem of resource management in a CDMA satellite network. In Chapter 3, we focus on adaptive beam configuration. We assume here that the number of beams is fixed, and that the shape of each beam is controlled by the satellite. In Chapter 4 we develop an optimal rate allocation policy that exploits the difference in timescale between the user QoS requirements and the network operation. A simplified network of two users is considered. In Chapter 5, we discuss extending the results obtained for the simple 2-user network to a general M-user network. We also consider the optimal access control problem for a homogeneous set of M users.

We conclude this thesis in Chapter 6 with a summary of the thesis and a discussion of directions in which this work could be extended.

Chapter 2

Resource Management in Satellite Networks

2.1 Introduction

Ka and higher bands enable the use of high-gain multi-spotbeams with small footprints that increase satellite power density and permit large frequency reuse — in turn enabling the support of thousands of multimedia user terminals equipped with small, inexpensive antennas [1, 10, 70, 71]. Communication at these frequencies also allows the use of smaller satellite antennas which results in a smaller satellite system and a lighter launch vehicle [72].

A principal requirement for a broadband multimedia satellite network is the capability to support a variety of traffic classes with different quality of service (QoS) requirements. While the capacity of a network is constrained by physical layer parameters (outside the scope of this thesis), in this thesis we are concerned with resource management to satisfy network layer requirements such as throughput and delay.

We begin by describing the CDMA satellite network model applied in this thesis. We then review beam management and cell/frame-level access control techniques that have been proposed to improve the efficiency of broadband satellite networks, and indicate how the thesis extends this work. In the case of beam management, we consider both fixed and variable beam allocation strategies; in the case of access control, we focus on three approaches to the rate allocation problem: class-based rate scheduling, opportunistic/greedy scheduling and fair access scheduling.

2.2 Satellite Network Model

A typical multimedia Ka band GEO satellite system is composed of a number of multibeam satellites with onboard processing (OBP) capability for switching and traffic resource management, and with intersatellite links for connections among the satellites [11, 73–76]. The satellite system supports a wide array of ground terminals including mobile and fixed satellite terminals with varying quality of service requirements and transmission rates, and gateway terminals or interworking units that act as interfaces between the satellite network and terrestrial networks, such as the PSTN and the Internet. The satellite network operations are managed via an earth-based master control station. Figure 2.1 presents a basic network architecture.

Traditionally, the ground terminals were large-antenna earth stations which served as gateways for user terminals. The first system to provide direct-to-user services was the Inmarsat system deployed in 1980 [1, 7]. This system consisted of low-power global GEO satellites and $1 - 1.5$ m mobile antennas. The trend to reduce the size of the user-antenna has continued since then; for example, the Inmarsat 3 system launched in 1996 supports mini terminals as small as laptop computers. As an example of the current variety of ground terminals, following are the components proposed for the EuroSkyWay (ESW) satellite network [76]:

- *Mobile satellite user terminals (SaT)*: of three types, SaT-A, SaT-B, and SaT-C, respectively offering uplink data rates of 160 kbps, 512 kbps, and 2.048 Mbps.

- *Service provider terminals*: fixed earth stations that connect service provider centers to the ESW satellite system. Uplink rates are multiples of 6.144 Mbps.

- *Gateway stations*: fixed earth stations that interface with the public terrestrial network through protocol adapting interfaces (for example, ISDN, ATM, PSTN) and interworking functions. Again, uplink rates are multiples of 6.144 Mbps.

- *Master Control Station*: a fixed earth station that manages the ESW satellite system by providing network management, call admission control, service management, billing, etc. Uplink data rates are multiples of 32.768 Mbps.

The downlink rate to all terminal types is 32.768 Mbps per carrier. The ESW satellite system itself is comprised of a 30/20 GHz band for the uplink/downlink connections and a

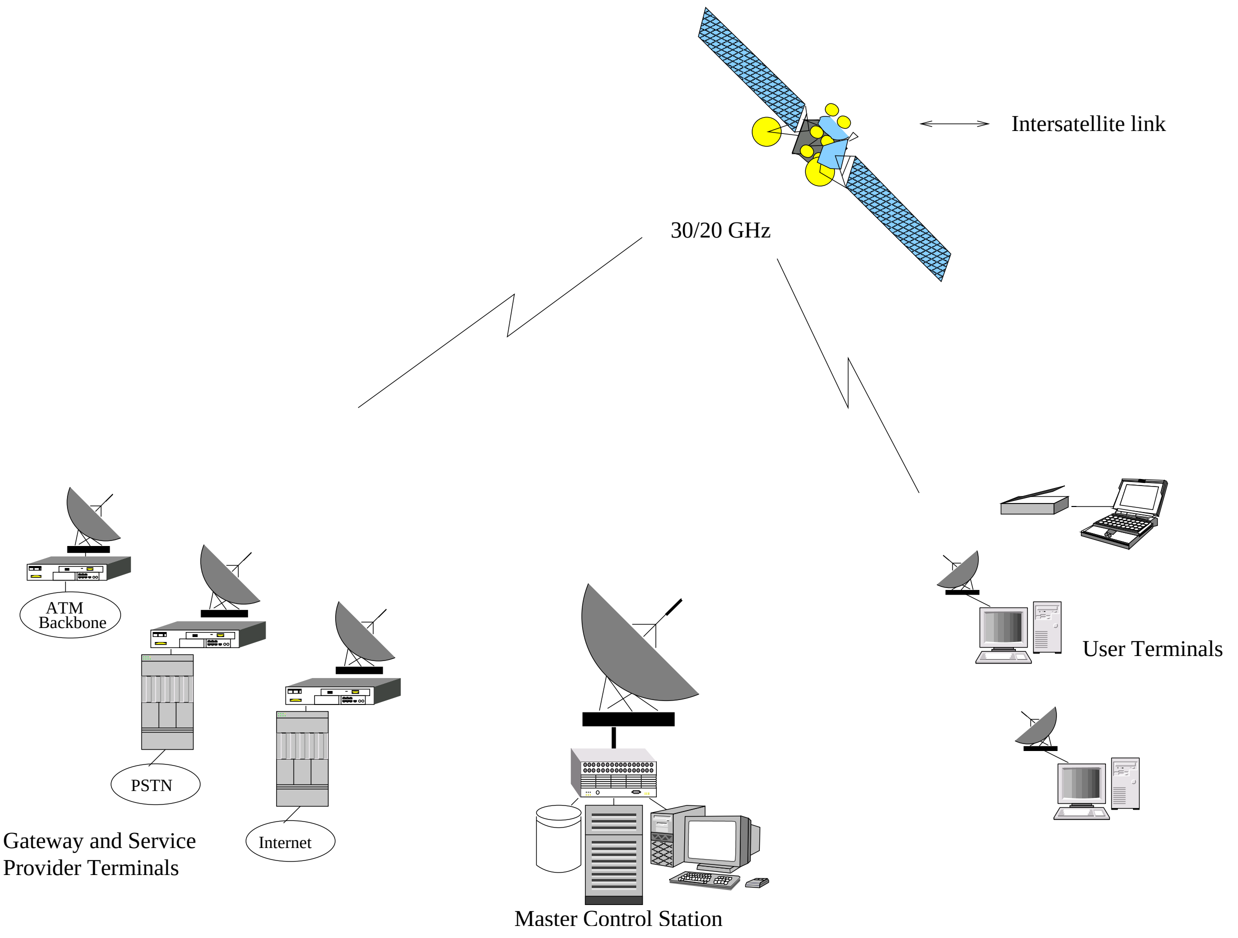


Fig. 2.1 Satellite network architecture

V band (56-64 GHz) for inter-satellite links with data rates of up to 163.840 Mbps.

As noted previously, the significance of on-board processing is that the satellite uplink and downlink can be optimized separately with respect to network performance. Our focus in this thesis is on control of the uplink performance via state feedback on the user distribution and propagation conditions, and subject to the assumption that the uplink bandwidth is fixed and known and to constraints imposed by the quality of service requirements. In the rest of this section, we provide an overview of common access schemes and of the channel model for the Ka band GEO based satellite network, and indicate how these are applied in the thesis.

2.2.1 Satellite Access and QoS Support

Connections in a multimedia GEO satellite network are generally of two types: permanent (allocated for the duration of a call) and semi-permanent (dynamically allocated on a frame-by-frame basis as needed) [74, 77]. Connection setup is handled by the earth-based master control station which is responsible for admission control as well as for enforcing routing policies and providing network management [11, 75]. Once a connection is set up, dynamic resource management is generally handled directly between the terminal (user or gateway) and the satellite resource management unit [75]. Note that signaling information suffers a delay of two satellite round trip times at the connection phase and one satellite round trip time in the resource management phase. Figure 2.2 is a schematic diagram of the functional modules for the satellite network.

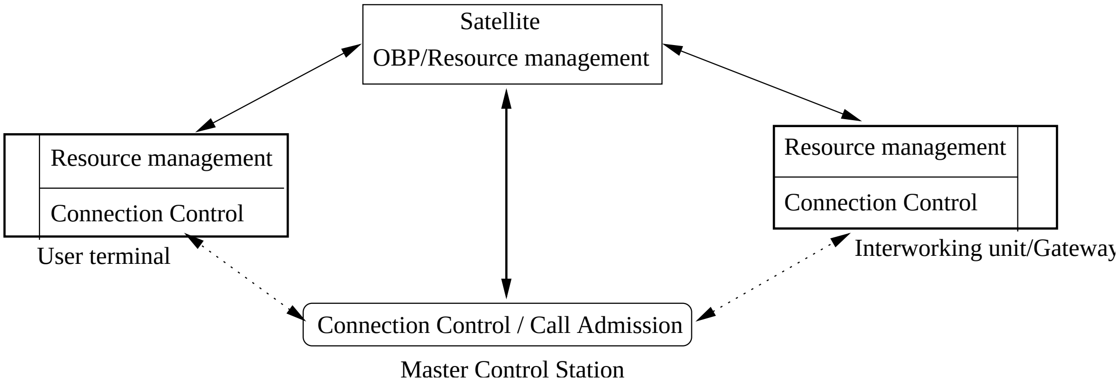


Fig. 2.2 Functional satellite network

The channel access and QoS support are key factors in the capacity of a satellite network to handle multimedia services with varying data rates. Satellite access techniques have evolved from frequency division to time division to code division [7, 78, 79]. In handling multimedia services, FDMA systems pose a greater challenge than either TDMA or CDMA because changes in channel bandwidth are required for changes in data rate [80, 81].

For the same channel bandwidth, CDMA and TDMA address changes in data rate per frame by respectively changing the processing gain or the number of assigned time slots. TDMA systems require explicit reservation for the time slots required (which introduces further delay). A connection may be severely degraded or dropped when there are no additional slots to accommodate a burst of traffic from the source. CDMA systems do not require explicit reservation at the burst level as changes in data rate are reflected in changes in the processing gain. This flexibility simplifies the implementation of access controls that seek to enhance system capacity via scheduling while providing acceptable QoS support to all terminals. However, any degradation of service due to a burst of traffic from one terminal is experienced by all active terminals. In fact, CDMA is said to have ‘soft‘ capacity because, typically, the number of active terminals is not limited by the number of CDMA codes but by the interference due to active users. This is unlike the ‘fixed‘ case of TDMA where the capacity is constrained by the number of time slots. In CDMA additional connections can be accepted until such time that the interference due to active users results in unacceptable signal-to-interference levels. Compared to a TDMA system, the CDMA approach allows for more connections to be supported and for graceful rather than abrupt degradation of service [15, pp. 163], [45].

Direct Sequence/Code Division Multiple Access (DS/CDMA) has been proposed for next generation satellite networks [12, 45, 82, 83]. In addition to flexible support of a diverse range of services, satellite CDMA systems offer universal frequency reuse (which also eases the problem of resource allocation), capability of soft satellite beam handoff, graceful degradation under loaded conditions, and a lower sensitivity to interference [45]. The International Mobile Telecommunications-2000 (IMT-2000) framework provides two modes of operation: a frequency division duplex (FDD) mode that makes use of wideband CDMA in paired frequency bands and a time division duplex (TDD) mode that makes uses of both TDMA and CDMA in unpaired frequency bands. TDD/CDMA has received considerable attention as a means of enhancing the efficiency of a wireless network supporting multimedia traffic [80, 84–87]. Via CDMA/TDD, the uplink and downlink transmissions are multiplexed on different timeslots of the same carrier whereby the slot assignment is based on traffic and QoS requirements. In satellite networks, TDD/CDMA has been considered for geostationary or elliptical regional-coverage systems that support low mobility terminals with high data rates requiring high transmission power [45] and for LEO systems [86].

We focus on an FDD wideband CDMA (WCDMA) satellite system, which is the more commonly applied model [45]. WCDMA has been proposed for broadband systems expected to support simultaneous transmissions that require varying quality of service and data rates up to 2 Mbps [88–91]. Currently, WCDMA uses a 5 MHz carrier, compared to a 1.25 MHz carrier used by narrowband CDMA. Future WCDMA systems are expected to spread over 10 – 20 MHz [90]. Based on direct-sequence CDMA, WCDMA supports variable rate allocations via variable spreading gain CDMA (VSG-CDMA) or multicode CDMA (MC-CDMA) [89]. In VSG-CDMA, the chip rate and system bandwidth are fixed, and higher transmission rates are obtained by reducing the spreading gain. The spreading gain varies with the length of CDMA code — the shorter the code, the larger the number of information bits that can be transmitted per unit time. VSG-CDMA is coupled with transmitter power control to maintain the integrity of rate allocations and SIR requirements by adjusting the energy-per-bit as the transmission rate varies. The VSG approach has increased signaling overhead because the receiver has to be informed about the changes in spreading gain. On the other hand, in MC-CDMA, each code sequence spreads a basic data rate over the entire system beamwidth. High data rates are supported by assigning multiple codes to a user. Careful selection of the spreading codes is required to limit self-interference. While MC-CDMA has a lower signaling overhead and does not require transmitter power control as does VSG-CDMA, it does require a receiver for each of the individual codes.

In this thesis, we consider the use of VSG-CDMA and assume ideal power control. Because CDMA systems are interference-limited, careful rate allocation and/or power control are necessary to make efficient use of system capacity as well as to satisfy the user QoS requirements. QoS support is a major challenge when handling multimedia services because of the diverse QoS requirements. It is widely held that satellite multimedia services will be carried over internet-based networks [11, 74, 77], where a number of models have been applied to provide QoS support; including Integrated Services (IntServ) and Differentiated Services (DiffServ). IntServ provides QoS guarantees on a per-flow basis while DiffServ enforces QoS guarantees to groups of flows. However, due to time-varying channels and long delays, ensuring QoS requirements is particularly challenging in satellite networks, and typically requires adaptation of QoS models originally designed for terrestrial application if such models are to be successfully applied within the satellite environment.

In addressing the problem of QoS support in satellite networks, the objective generally has been to enhance network utilization by enabling dynamic sharing of the network resources while giving priority to real-time traffic [73, 74, 92–95]. Data scheduling is used to exploit the temporal elasticity of delay-tolerant traffic in an endeavor to meet the QoS requirements of active users. Such scheduling is governed in a variety of ways as we shall review in Section 2.4. However, if not properly controlled, data scheduling can result in excessive delays for users with poor channels or in systems with very many users. In this thesis we assume that all users possess various degrees of temporal elasticity and the problem to be addressed is one of access control so as to meet the QoS requirements of all active users.

2.2.2 Satellite Channel Model

This thesis focuses on the Ka (30/20 GHz) and higher frequency bands, which with their high bandwidth availability provide the capability to support high data rate multimedia services. Table 2.1 presents frequency spectrum arrangements as applied to satellite and terrestrial wireless systems [96]. The frequency ranges in the table are only general in nature and, hence, do not match the frequency allocations of the International Telecommunication Union exactly.

Table 2.1 General frequency spectrum arrangement and typical applications. *Source: [96]*

Band	Freq. (GHz)	Typical applications
L	1 - 2	MSS, UHF TV, terrestrial microwave and studio TV links, cellular phone
S	2 - 4	MSS, NASA and deep space research
C	4 - 8	FSS, FSTM
X	8 - 12.5	FSS military communication, Fixed terrestrial wireless service, Earth exploration and meteorological satellites
Ku	12.5 - 18	FSS, BSS, FSTM
K	18 - 26.5	BSS, FSS, FSTM
Ka	26.5 - 40	FSS, FSTM, LMDS

MSS: Mobile satellite service
FSS: Fixed satellite service
BSS: Broadcast satellite service
UHF TV: Ultra high frequency television
FSTM: Fixed service terrestrial microwave
LMDS: Local multipoint distribution service

For bandwidths above 1 MHz, channel fading is independent of frequency and therefore may be considered flat [45, 97]. In general, the satellite propagation channel for fixed terminals can be modeled as a flat fading Ricean distributed channel [45, 97–101]. In the case of mobile terminals, the channel model also includes log-normal shadowing [100, 101]. The Ricean channel model holds particularly true at high elevation angles because signal propagation conditions improve and shadowing effects decrease with increasing elevation angle [99]. The fading on the uplink and downlink channels is usually uncorrelated since their frequency bands are separated by more than the coherence bandwidth of the channel [101]. However, it is usually assumed that the same degree of shadowing is experienced on both links.

A major problem for communication at the Ka band and beyond is attenuation due to rain, particularly in tropical and sub-tropical regions where such attenuation can easily exceed 20 dB [10, 40, 72, 102–105]. Under severe rain attenuation, the channel model is reasonably represented by a flat fading Rayleigh distributed channel [98]. On the other hand, under clear-sky conditions, the satellite channel can be approximated by an additive white Gaussian noise model, which allows one to simply focus on other users (intra-beam or inter-beam) as the source of interference [98].

2.2.3 Rain attenuation and rain fade countermeasures

As noted, rainfall is a major concern at high frequencies as it can result in severely faded channels. Raindrops absorb and scatter the signal, reducing its amplitude and distorting its phase [72]. Tropical rain regions are particularly affected due to the large amounts of rainfall received. Table 2.2 presents the rain rate distributions for the fourteen rain zones defined by the International Telecommunication Union - Radiocommunication sector (ITU-R), where regions N and P identify tropical rain zones [106].

In general, rain attenuation models are 'semi-empirical' in that available attenuation data is used in the model development [72]. Most of these models are based on the relationship $A = aR^bL(R)$, where A is the attenuation loss in dB, R is the point or surface rain rate in mm/hour, a and b are frequency- and temperature-dependent constants, and $L(R)$ is the mean radio path length through the rain in kilometers. The term aR^b (dB/km) is also referred to as the specific attenuation of rain volume or decibel loss per path length. $L(R)$ depends on the elevation angle, the rain cell size, the rain rate, and the 0^0C isotherm

Table 2.2 Rain Rate Distribution for ITU-R Rain Zones. *Source: [106]*

Percentage	*Rain Rate Distribution (mm/hour)*													
of time (%)	A	B	C	D	E	F	G	H	J	K	L	M	N	P
1.0	-	1	-	3	1	2	-	-	-	2	-	4	5	12
0.1	2	3	5	8	6	8	12	10	20	12	15	22	35	65
0.03	5	6	9	13	12	15	20	18	28	23	33	40	65	105

(Rain rate distributions are based on an average year)

height (the rain height or the height at which freezing occurs). Typically the rain rate is not constant over the duration of the storm and so $L(R)$ is modeled statistically from measured surface rain rates.

Mobile terminals will generally experience an attenuation, A_M dB, which is different from the attenuation, A_F dB, of a fixed terminal [102, 103]. A common model applied is one of equal probability of encountering rain: over the long term both fixed and mobile terminals will experience the same number of rain storms, and hence the same number of rain attenuation events such that $A_M = A_F = A$ dB. In this case, the probability distributions of the rain attenuation of fixed terminals, $P_F(A)$, and of mobile terminals, $P_M(A)$ are related by:

$$P_M(A) = \zeta P_F(A) \tag{2.1}$$

where ζ ranges in value from $0.5 - 2$ depending on the mobile terminal's experiences in the rain storm. In turn, this is dependent on whether or not the rainstorm and mobile terminal travel in the same direction and on how the mobile travels within the rain path (straight path or zig zag).

Our concern with the issue of rain attenuation is because this thesis is motivated, in part, by the issue of limited telecommunication infrastructure in Africa — much of which is covered by the tropical rain zone. As such, rain fading is a challenge that must be addressed since satellite networks can provide communication services long before they would be available via terrestrial networks. For a Ka band satellite system with a 4 dB fade margin, Table 2.3 highlights the impact of rain attenuation on four levels of satellite availability in a tropical region within the Pacific [10].

At lower frequency bands, rain fade depths are small enough to be incorporated as a fixed link margin [83, 106–108]. For example, power margins of 5–10 dB at the C band

Table 2.3 Outage time versus availability for a Ka band satellite in the tropical region of the Pacific *Source: [10]*

Satellite availability (%)	Outage per week (minutes)
99.70	30
99.30	71
99	101
98	202

(6/4 GHz) can be achieved, relatively easily, with reasonably sized antennas and within the transmitter power constraints [106, 107]. At higher frequencies, fixed margin allocations are expensive to realize and a number of rain fade countermeasures (FCM) have been proposed to mitigate rain attenuation [10, 108]. The goal is to design adaptive communication systems that can maximize the data throughput and meet bit error rate specifications.

Generally, the higher the availability requirement, the higher the required fade margin. This can result in over-utilization of the FCM resources or in poor utilization of the channel capacity. Appropriate resource management is thus required to make efficient use of the system resources. Common proposals include the use of frequency diversity — for example, switching between Ka and Ku (14/12 GHz) bands depending on the severity of the attenuation [108]; the use of adaptive resource allocation — for example, by reducing the data rate, by additional coding and by dynamic assignment of the antenna directivity pattern [10, 83, 107–110]; and the use of site diversity — applicable because rain cells tend to be highly localized in area [40, 111]. Considered especially relevant to the downlink, time diversity has been proposed for video-on-demand applications — time-delayed programme repeats are used to overcome the effects of severe deep fades typical of tropical rain regions [112].

Alternatively, rain fade countermeasures can be neglected all together. Instead, satellite networks can be designed with reasonable fade margins and for lower availability levels that are still acceptable to users [10]. For example, based on Table 2.3, a network could be designed for 99% availability with an average daily outage time of 13 minutes. This results in reasonable services charges for the users and an increase in the number of potential users for the network provider. Select customers who must have higher availability (and are willing to pay) would receive, say, larger antennas and/or power amplifiers.

In the sequel, we discuss various options for resource management in satellite networks.

Specifically, we review related work on the two problems that are the focus of this thesis: beam management and access control. Beam management, which is based on the user distribution, is assumed to hold steady for long periods of time. Access control, which is in response to traffic demand or channel fading, varies over much shorter timescales, on the order of seconds or milliseconds or less.

2.3 Satellite beam management

The traditional approach to beam allocation is to divide the satellite coverage area into regions of maximum geometric size so as to minimize the required number of beams. Typically, all the beams are of the same size. In contrast, this thesis considers adapting beam shape and size to propagation conditions and population distribution. Adaptive beam management is considered a key on-board technology for the efficient use of satellite capacity, for efforts to combat rain attenuation, and for the realization of low-cost earth terminals [10, 24, 25, 30, 33, 113]. Such efficiency, however, is dependent on the expected degree of departure from a uniform traffic distribution and the degree of penetration of terrestrial facilities [114, 115].

There are a number of parameters of interest in the beam management process: beam coverage parameters and satellite link parameters.

2.3.1 Satellite beam coverage parameters

The beam coverage of a satellite system is the geographic area over which the antenna gain exceeds some threshold value. The minimum gain required of the antennas to support a given beam allocation pattern is set by link parameters such as the satellite altitude, the data rate, losses at different segments of the link, the maximum size of the antenna reflectors, and the required signal-to-interference ratio [23].

Typically, beam coverage area is described in terms of the angular 3 dB beamwidth, the angle in degrees between the directions in which the gain falls to half its maximum value, which is given by [106, 116]:

$$\theta_{3dB} = \beta \frac{\lambda}{\phi_{ant}},$$

where β is a coefficient whose value depends on the chosen illumination law, λ is the wavelength, and ϕ_{ant} is the diameter of the satellite antenna. The maximum gain, G_{max}, is related to the 3 dB beamwidth as [116]:

$$G_{max} = \eta \left(\frac{\beta \pi}{\theta_{3dB}} \right)^2 ,$$

where η is the efficiency of the antenna. Thus, beam coverage area is inversely proportional to the required gain, and the need to provide high gain across a coverage area results in the generation of beams with small beamwidths. Furthermore, high gain satellite antenna systems enable the use of cost-effective low-gain low-power handheld terminals.

In the Ka band and beyond, where the wavelengths are small, high-gain coverage is achieved with reasonably sized antennas. Ka band satellite antennas can realize high link margins of $10 - 16$ dB with spotbeams of very narrow-diameter ($300 - 400$ km) or narrow-beamwidth ($0.25^0 - 0.75^0$) [1, 10]. For example, the EuroSkyWay Ka band GEO system for fixed and multimedia communications will provide coverage to Europe using 32 spotbeams [15, pp. 372]. On the other hand, operating at a lower frequency of 1-2 GHz, the European Space Agency L band Land Mobile (LLM) GEO system for speech and data transmission will provide coverage to Europe using 3 spot beams and 1 regional beam [15, pp. 362].

As a consequence of the narrow beam widths, multibeam antennas are required in order to provide satellite coverage across a wide geographic coverage area. Generally, a multibeam antenna can have fixed beams, adaptive beams, or a combination of these [23, 24, 28, 29, 33, 117]. Adaptive beams may reduce the size of the array feed and associated beam-forming network, but they require more complex circuitry and management relative to fixed beam antennas. The type of antenna system, fixed or variable, plays a key role in the flexibility available for adaptive beam management.

A further consequence of the narrow spotbeams is the increased likelihood that the user distribution will vary greatly from beam to beam. Thus, it becomes necessary that the resource allocation is tailored to the spatial and temporal variation of the user distribution across the satellite coverage area in order that the satellite resources are utilized efficiently. Because frequency spectrum is a limited resource, multibeam antenna systems enhance the capacity of the network by offering opportunity for frequency reuse among the various beams. In particular, CDMA systems offer the opportunity for the same frequency carrier to be used among all the beams.

2.3.2 Satellite link parameters

In addition to the beam coverage area, a number of parameters are of concern at the satellite link level. On the uplink (terminals to satellite), key parameters of concern include terminal power constraints and the effective area of the satellite receiving antenna. On the other hand, on the downlink (satellite to terminals), key concerns include satellite output power as well as the service area of individual beams — which has impact on the broadcast and multicast operations of the satellite. The parameters that affect received power in either direction are transmitter power, P_T, transmitter antenna gain, G_T, receiver antenna gain, G_R, distance between the transmitter and receiver, H, the wavelength of the carrier frequency, λ, and the receiver antenna effective size, $A_{eff} = G_R\lambda^2/4\pi$ [118]. The total signal power, C, incident on an antenna of effective area A_{eff} located a distance H from the transmitting antenna is given by:

$$C = \frac{P_T G_T}{4\pi H^2}\left(\frac{G_R\lambda^2}{4\pi}\right). \tag{2.2}$$

For reliable information transmission, the transmitter power, P_T, should satisfy a required carrier-to-noise ratio (C/N). The value of (C/N) depends on the required information rate, the required signal-to-noise ratio (for analog signals) or bit error rate (for digital signals), and the modulation system and associated bandwidth. The thermal noise present can be written as $N = kT_sW$, where T_s is the overall receive system temperature, W is the bandwidth and k is Boltzman's constant. Considering free space propagation losses, the fundamental relationship for link performance is given by:

$$\frac{C}{N} = \frac{P_T G_T}{(4\pi H/\lambda)^2}\left(\frac{G_R}{T_s}\right)\frac{1}{kW} \qquad \text{or} \qquad \frac{C}{N_0} = \frac{EIRP}{(4\pi H/\lambda)^2}\left(\frac{G_R}{T_s}\right)\frac{1}{k}, \tag{2.3}$$

where EIRP $\triangleq P_T G_T$ is the effective isotropic radiated power and $N_0 = kT_s$ is the noise power density in watts per Hz [106, 118]. In digital networks, it is typically the energy per bit to noise density ratio, E_b/N_0, that is used in specifying a required bit error rate (BER).

The relationship between E_b/N_0 and C/N is given by:

$$\frac{E_b}{N_0} = \frac{C}{N}\frac{W}{R_b}$$

where R_b is the data bit rate. For a given satellite antenna effective size, $A_{eff} = 0.5$, Table 2.4 shows the actual antenna size required to achieve an uplink C/N_0 of 49 dBHz in different satellite orbits [119]. It is assumed that the EIRP is -3 dBW and the satellite receiving noise temperature is $T_S = 600K$.

Table 2.4 Satellite antenna size for various altitudes. *Source: [119]*

Altitude (km)	System	Antenna diameter (m)
35,786	GEO	10.36
10,370	MEO	3.00
1,389	Globalstar (LEO)	0.40
780	Iridium (LEO)	0.23

Due to the different concerns on the uplink and downlink of a satellite network, the antenna and consequently the beam design differs for the two links. In general, for a satellite network in which fixed terrestrial coverage is provided, the uplink and downlink performance is constrained by the antenna size of the earth terminal [118, 120]. Table 2.5 presents uplink parameters for two antenna sizes of a GEO-based Ka band multi-beam satellite system supporting three different data rates [71]. The following are assumed: the satellite antenna aperture efficiency is 40%, the ground terminal antenna aperture efficiency is 50%, the satellite system noise temperature is 600K and the ground terminal system noise temperature is 300K.

From the table we can observe that for the same transmission power and link margin, the higher the data rate, the larger the antenna size required to provide the necessary transmission gain. At low rates, an omni-directional antenna can be used which enables communication without knowledge of the satellite's direction [119]. Such a radiation pattern requires an antenna gain of about 0 dB. At high data rates, a directive antenna is required.

In the following section we focus on beam management in terms of adapting the beam shapes and sizes to support a range of multimedia services under varying channel conditions.

Table 2.5 Uplink parameters for a Ka band satellite network ($W = 29.5$ GHz). *Source: [71]*

Satellite receive antenna	3.5m ϕ			10m ϕ		
Spotbeam coverage	120km ϕ			42km ϕ		
Transmission bit rate	8 kbps	64 kbps	1.544 Mbps	8 kbps	64 kbps	1.544 Mbps
Ground antenna	2cm ϕ	5cm ϕ	20cm ϕ	0.4cm ϕ (0 dB)	2cm ϕ	10cm ϕ
Ground transmit power	1W	1W	2W	1W	1W	1W
Uplink margin	8 dB	7 dB	8 dB	5 dB	8 dB	8 dB

2.3.3 Beam configuration management

The goal of future wireless networks is to bring to the untethered and/or remote user a broad range of multimedia services — each with different processing gains and SIR requirements. Hence, dimensioning and planning for network coverage must address the user distribution, the propagation conditions and the interference control mechanisms. Fig 2.3 presents a beam management model adapted from general planning models for multimedia wireless networks [121, 122].

A number of techniques have been proposed for beam management. These techniques can be broadly classified as based on a uniform-beam or variable-beam configuration. In uniform-beam configuration, the beam shape and size are fixed but the bandwidth allocated to each beam can vary with the active user profile. In variable-beam configuration, the beam shape, size and bandwidth are adapted in accordance with the active user profile.

Uniform-beam configuration

While the beam size and shape is the same for all beams, the beam bandwidth is dynamically allocated in order to match the spatial and temporal variation of traffic over the coverage area. Ka band proposals with uniform-beam configurations include the North American Multimedia Satcom, Anik F2 (Canada), and the EuroSkyWay satellite system [4, 5, 12, 39, 114].

The North American satellite systems propose a combination of beam hopping and dynamic channel allocation to adapt bandwidth allocation to the active user profile. For example, across Canada, the Anik satellite system will have a beam configuration that consists of twenty-seven identical small high-gain elliptical ($0.7^o \times 0.3^o$) beams over southern

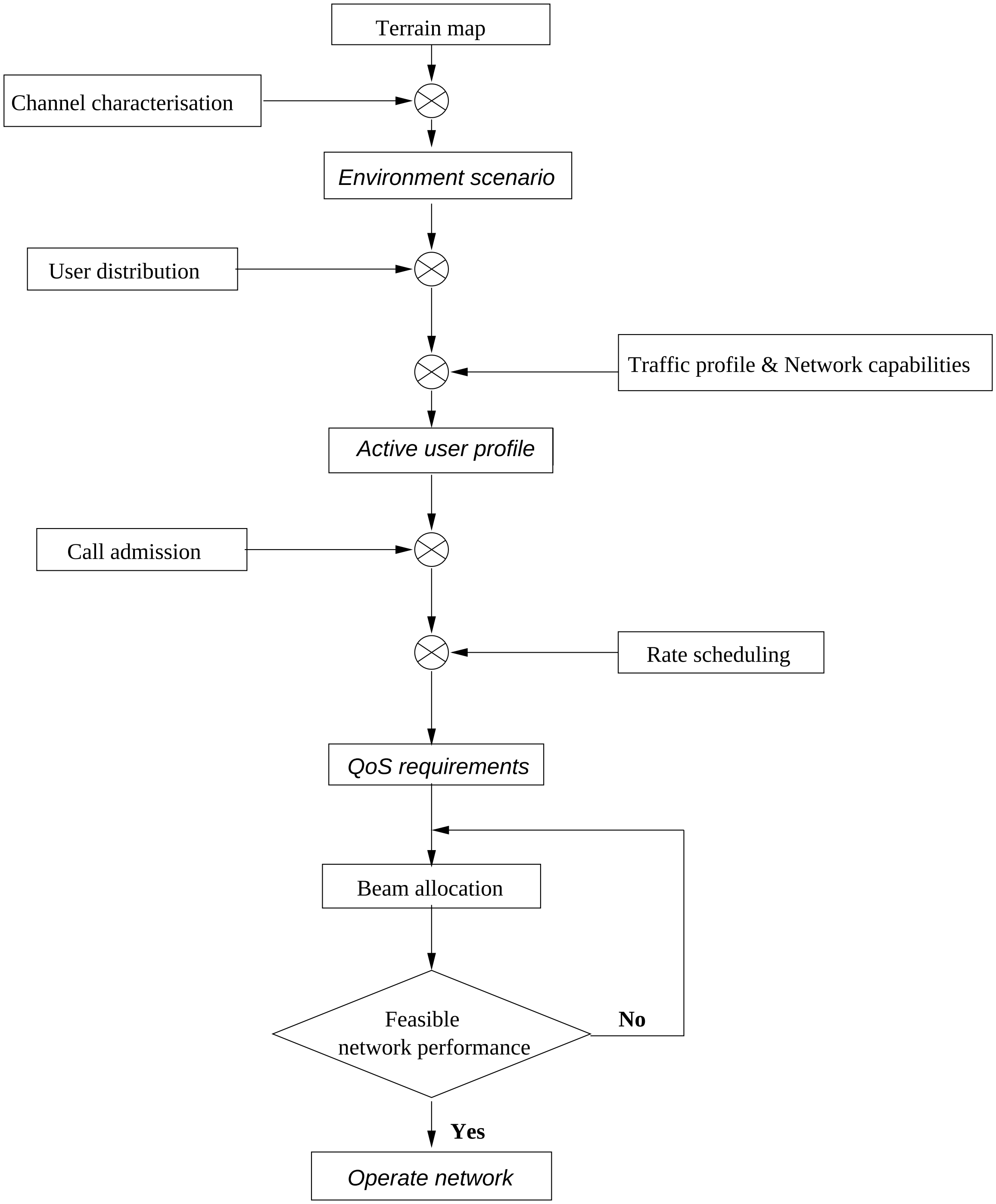


Fig. 2.3 Beam allocation model for future generation wireless networks

Canada and two larger northern beams [39]. Adjacent spotbeams are combined into beam groups consisting of two, three and four spot beams [4, 39]. The beam groups operate as independent sub-networks and the satellite bandwidth is shared equally among the beam groups.

Within each beam group, capacity is allocated according to the forecasted traffic distribution. On the uplink, multi-frequency (MF) TDMA is used to share the available capacity among active users within the group. For example, for a group of four spotbeams that has been assigned seven frequency channels, four channels would be allocated to the spotbeam with the most traffic at any given time and the other spotbeams would each be allocated a single frequency channel [4]. On the downlink, the allocated capacity will be shared via time-division multiplexing (TDM) whereby the downlink carrier for the group is connected to a destination spotbeam for the duration of the traffic burst to that beam.

The principal difference between the North American satellite systems and the EuroSkyWay system is that, in the latter, the satellite capacity is distributed among the beams based on the traffic distribution requirements in each individual beam [5, 12]. Given that the initial phase of the EuroSkyWay system will comprise of thirty-two spotbeams, the level of complexity for the resource allocation process is significantly increased compared to the North American systems that allocate resources on a sub-network basis.

The resource allocation objective in the EuroSkyWay system is to locate and dimension the satellite system based on a market-driven approach for millions of low cost user satellite terminals [123]. For each beam, the traffic distribution is estimated on the basis of the Gross Domestic Product (a measure of the overall size of a country's economy) and the covered area per spot beam (for example, ignoring uninhabited areas).

The dimensioning of the EuroSkyWay satellite system is based on an approach proposed by the Advanced Communications Technologies and Services - Satellite EHF Communications for Multimedia Mobile Services (ACTS-SECOMS) project [12, 123, 124]. In SECOMS, the spotbeams are uniformly assigned across the coverage area and allocated bandwidth based on the traffic distribution in each beam. Two access schemes have been considered for the uplink within each beam: MF-TDMA and slotted CDMA whereby multirate services will be multiplexed in time. The downlink capacity will be shared in TDM fashion.

For the TDMA uplink system, the required power and bandwidth per spotbeam is calculated beam-by-beam sequentially, based on the traffic density, the modulation and channel coding technique, and the frequency reuse factor. For the CDMA uplink system, because of the impact of multiple access interference, calculation of the power and bandwidth per beam must be conducted simultaneously across all beams. A resource allocation scheme based on genetic algorithms is used to determine the number of carriers to be allocated to each spotbeam as well as the frequency carrier and polarization distribution over the coverage area [12]. The objective is to optimize the power distribution over the coverage area in order to minimize the maximum bandwidth and the power requirements.

The performance of the optimization algorithm for a synchronous CDMA system was compared to the bandwidth requirements for a TDMA system. In both cases, the same traffic distribution and combined modulation and coding scheme was assumed. In general, it was observed that the total bandwidth required was lower for the CDMA system compared to that required by the TDMA system [12]. However, the CDMA system required higher average power per terminal.

Nevertheless, in either mode of multiple access, the proposed dimensioning of a fixed-beam system yields an optimal resource allocation in terms of a lower cost system for both service providers and users relative to a uniform-beam system with uniform capacity allocation [123]. Such cost savings represent a major issue for commercialization and viability of satellite services.

Uniform-beam configurations place a constraint on capacity allocation in that beam allocation (and hence satellite gain) is fixed independent of user distribution. The proposals described here address the challenge by allowing for dynamic power and bandwidth allocation in order to achieve the QoS targets for all users irrespective of location. The alternative provided by variable-beam configuration is to use variable satellite gain allocation as an additional degree of freedom in the quest for optimal resource allocation.

Variable beam configuration

Variable beam configurations seek to adapt the beam shape and size to the user distribution and propagation conditions. Typically, the beam coverage area is defined by the location and number of terminals to which the beam antenna must offer a minimum gain value [10, 25, 63]. We consider two cases, one for the uplink and the other for the downlink.

For the uplink scenario, Cances *et al.* consider two Ku band multibeam systems that support 64 kbps telephony-type traffic via FDMA [63]. One system applies frequency reuse and is comprised of thirty-two contoured beams. The second system does not apply frequency reuse and is comprised of five contoured beams. In both systems, the shape of each individual beam is changed according to variations in the traffic so as to adapt the capacity of the beam allocation to the traffic demand at call-level. Beam allocation is reconfigured when either the frequency channels available within a particular beam can no longer support all the traffic within its coverage area or when a terminal within the satellite coverage area does not receive sufficient gain by any beam. Beam reconfiguration is conducted by shifting the contours of the beams in the neighborhood of the *new* terminal and/or the *overloaded* beam until a configuration that supports all terminals is obtained.

The performance of the variable-beam systems was compared to that for uniform beam allocations with dynamic capacity allocation. The number of beams was the same in both cases and the performance was assessed in terms of call blocking probability and mean waiting time. It was shown that:

- when a system with no frequency reuse is considered, the uniform coverage system performs better in that it has both a lower mean waiting time and a lower call blocking probability. However, the uniform coverage system cannot provide service to terminals outside of its beam coverage area.

- when a system with frequency reuse is considered, at high demand, the variable coverage system performs better than the uniform coverage system. It is observed that the variable system also enables better control over inter-beam interference among beams with the same frequency bands. This is because it is able to balance the demand from beam to beam.

On the downlink, broadcasting/multicasting is a key application for satellite networks because a single beam can provide coverage to a very wide area. Hence, a typical beam allocation objective is to divide the satellite service area into beam coverage areas of maximum feasible size (satisfying a minimum gain requirement) so as to minimize the number of beams required. The problem with uniform beam allocation is that the resulting power requirements in some of the beams may exceed the output power levels of the satellite antennas [25]. Variable beam allocation based on user distribution and propagation conditions can be used to minimize the number of beams required for a given constraint on the satellite output power.

Shimada *et al.* propose variable beam allocation based on the geographic population, rain attenuation and maximum beam power constraints [10, 25]. Let the coverage area require M channels. If the power per channel is p, assume that $P_{max} < pM \leq 2P_{max}$. Then, the service area can be subdivided into two beams such that *beam* #1 supports M_1 channels and *beam* #2 supports M_2 channels where $pM_1 \leq P_{max}$, $pM_2 \leq P_{max}$ and $M_1 + M_2 = M$. The number of beams required is minimized by selecting the largest service area per beam for a given P_{max}. The satellite power requirement per beam coverage area, P_{req}, is defined by:

$$P_{req} \propto (\text{maximum rain margin}) \cdot (\text{population density}) \cdot (\text{beam coverage area}),$$

where the population density is used as an indicator of the number of channels required and the beam coverage area is a measure of the satellite antenna gain provided to the service area.

Given the population distribution and rain margin data (assuming 99% link availability) and beginning from one end of the satellite coverage area, a search program is used to map out the beam allocation. Any beam allocation yielded by the search depends on the direction of the search (north to south, south to north, or from center). The performance of the proposed scheme was compared to that of a uniform allocation scheme in providing service to Japan. For the same antenna size, five beams were required for the uniform allocation, compared to seven beams for the variable beam allocation. However, the uniform beam system required additional power equipment in order to satisfy the higher power requirements.

In summary, the uplink and downlink cases described highlight the gains in resource utilization accruing from variable beam allocation. In either case, the allocation was based on fixed power and/or rate considerations. This thesis focuses on the case where both the rate and power allocated are subject to change. We are also interested in a flexible search algorithm in which allocation optimization parameters (such as coverage area, beam size, data rates) can be varied in order to generate particular beam allocation patterns.

While we have focused on beam management in satellite networks, the optimal allocation problem has also been considered in terrestrial wireless networks for assigning users to base stations or determining base station locations [65, 125, 126]. A key difference, as earlier noted (Section 1.5.2), is that the allocation of terminals to beams in satellite networks is primarily a function of their location whereas in terrestrial wireless networks the allocation is primarily a function of received power or received signal-to-interference ratio. The timescale of the terrestrial allocation process is typically much shorter, in part, because of terminal mobility.

In the following section, we review work related to the second problem addressed in this thesis: cell/frame-access control. The essence of solutions proposed for this problem can usually be applied to either satellite or terrestrial wireless networks.

2.4 Cell/frame-level access control

As is the case for terrestrial wireless networks, the objective of access control in satellite systems is typically to improve the network capacity or throughput with transmission priority extended to real-time traffic [13, 73, 76, 92, 93, 127, 128]. Achieving this objective requires knowledge or estimation of the channel conditions. Timely feedback can be useful in anticipating channel degradation or congestion and in countering the effects on quality of service.

The issue of response and feedback times poses a particularly significant problem for satellite networks. This is due to the long propagation delays of GEO-based networks and the large delay variations of the LEO-based networks [74]. Delay is variable in LEO networks because the distance from user to satellite varies with the motion of the satellites and with changes in the dynamic assignment of users to satellites [7]. Control in the LEO case is further complicated by the handover process by which responsibility for a particular connection is transferred from one satellite to another. Our work focuses on the GEO setting.

In wireless CDMA networks, satellite or terrestrial, access control techniques typically involve the control of any (or a combination) of the transmission power, the target signal-to-interference ratio or transmission rate (or codes) so as to enhance network capacity or throughput and to support user QoS. First-generation commercial CDMA systems provided voice service, or voice with data having homogeneous QoS targets [129]. For such systems, the admission control and power control problems were dealt with as independent problems. Admission policies for CDMA were based on measurements of the system load as represented by the number of active users or on the multiuser interference. Due to the delay-intolerant nature of voice, the goal of system design was to provide constant rates, using power control to compensate for channel fades and propagation path loss [130]. With the growing necessity to support multimedia services, next-generation systems have supported joint rate and power allocation schemes in order to enhance system capacity while ensuring support for user quality of service requirements. CDMA schemes lend themselves to adaptive rate allocation in a simple manner by using multiple codes, multiple processing gains, or multirate modulations [130]. Our work focuses on delay-tolerant services in GEO-based CDMA multimedia networks in which changes in data rate are effected via changes in the processing gain.

In this section, we review related work on the access control problem and highlight how this thesis extends the body of work. Access control literature can be divided into two broad categories. In the first category, we shall consider solutions in which power control is used to maintain SIR targets for a given resource allocation, for example, to mitigate the near/far effect or rain fades. In the second category, we shall consider solutions in which dynamic power and/or rate control is used to enhance network throughput in a system constrained by the various user QoS requirements. Of particular interest is understanding the extent to which temporal elasticity can yield a throughput advantage in systems with varying QoS requirements and time-varying propagation conditions. In this thesis, we show that such advantage can be gained via access control that relies upon state feedback of the active user profile and the propagation conditions.

2.4.1 Access control to mitigate the effects of fading and interference

We distinguish between fast and slow power control. Fast power control is conducted at the physical layer to counter short-term fading effects and near/far problems; in a terrestrial CDMA voice network, for example, the result of such control is that all transmissions are received with equal powers. Slow power control, on the other hand, exploiting the essentially one-to-one correspondence between rate and power, is for purposes of rate allocation. It is conducted at the medium access control layer.

Algorithms implementing power control can be categorized as centralized or distributed. In a centralized scheme, the power levels for all users are administered by a single controller (for example, a base station or satellite) [131–133]. For a distributed scheme, each user controls its own power levels based on information local to the user [134–139]. While centralized schemes require significant global information management and may induce system vulnerability, distributed schemes take longer to converge to optimal power levels and may cause system instability.

Network access control can be done as a two-step process whereby the first step is to obtain an efficient distribution of users to beams followed by the dynamic power allocation step. Section 2.3.3 discusses approaches for conducting the first step in satellite networks. An example for terrestrial wireless networks is the ‘cell breathing‘ approach whereby the set of users assigned to a base station expands to include new users, if the cell is lightly loaded, or contracts to hand-off some users, if the cell is heavily loaded [140]. The changes in number of cell users occurs as the users are switched among the beams in search of an allocation that satisfies the SIR requirements of all active users with minimal power. This allocation is attained when no user can independently decrease their allocated power by changing from one cell to another. The ‘cell breathing‘ approach is analogous to the beam management problem with respect to the shaping of cells based on user distribution and channel conditions.

In CDMA networks the objective of dynamic power allocation is to ensure respect for SIR constraints and counteract the near/far problem in which a strong transmitter close to the receiver can swamp out the desired signal from a distant transmitter. The near/far problem is of greater concern in terrestrial wireless networks than it is in satellite networks since satellite propagation paths are approximately the same due to the great distance between satellite and terminals [100]. The challenge in satellite networks is effective tracking of channel conditions due to the long propagation delays. In general, one would need to account for issues such as variable satellite antenna gain due to different pointing angles among the terminals, log-normal shadowing for mobile users and Rayleigh fading variation [45, 100]. The short-term effect of Rayleigh fading variation is usually too fast to be controlled, and hence, closed-loop (fast) power control is considered impractical in satellite CDMA [100, 101, 141, 142]. However, the Rayleigh fading degrades a CDMA system by almost halving the number of users that can be supported even after the large-scale variations have been controlled [100]. It is thus necessary to deal with this effect head-on in order to enhance network efficiency and throughput.

Solutions to address the channel tracking problem for the Ka (30/20 GHz) band include frequency scaling of the attenuation on the 20 GHz downlink to estimate the uplink attenuation or adding a 30 GHz beacon at the satellite [107, 142]. The latter scheme offers significantly better estimates of the uplink attenuation — at the expense of acquiring the beacon. Two proposals to measure the degree of attenuation are:

- Use of a channel state predictor [100]. The satellite measures the mean of the received power over a time period — this mean is the result of shadowing and path loss. It then measures the instantaneous power variation due to Rayleigh fading. Assuming these measurements are very accurate, the receiver uses this collected data to predict the uplink Rayleigh fading amplitude after a given time.

- Use of a BER-driven, dual loop power control system comprising an inner and an outer closed loop [45, 127]. The inner loop which is based on the signal-to-interference ratio is used to compensate for large scale signal variations due to path loss, interference and shadowing on a user-by-user basis. To counteract the fast fading components, an outer loop is then used to maintain the target signal-to-interference so that all users obtain the desired performance in terms of average bit error rate.

In this thesis, we assume that every terminal transmits at the power assigned to it. We focus on rate allocations that exploit temporal elasticity to enhance network throughput. In the following section, we review different approaches for access control to enhance throughput. Here, the overall objective is for efficient network operation via control of who gets access to the network and for how long such access is permitted.

2.4.2 Access control for enhanced network throughput

In this section we consider solutions proposed for efficient transmission by delay-tolerant users in wireless networks. Of particular interest is the manner in which (if at all) state feedback in terms of the prevailing channel conditions and the terminal backlogs are applied in the control process. The literature can be broadly categorized as class-based access control, opportunistic (‘greedy‘) access control, and ‘fair‘ access control.

Class-based access control

Typically, users are classified according to the nature of their traffic, for example, voice/data or realtime/delay-tolerant, and the transmission rate per class is fixed. The quality of service requirement is in form of an signal-to-interference (SIR) target for all active users. The control variable is the power which is allocated so as to maximize network throughput while minimizing the interference due to simultaneous user transmission.

A common form of access control is data scheduling which is used to improve throughput and to ease traffic congestion by delaying the transmission of non-realtime users. Typically, a stationary channel and unlimited terminal buffers are assumed. Then, given the data rates and SIR targets, the number of non-realtime users allowed to transmit is constrained by the need to maintain the total network interference below a given threshold, while giving priority to real-time services [82, 143–145].

For example, consider a system with bandwidth W and with k_V voice users and k_D data users. Let γ_V and γ_D denote the SIR targets for the voice users and data users respectively, ν_i and ζ_i respectively denote the activity indicators for the i-th voice and data users, and $G_V = W/R_V$ and $G_D = W/R_D$ denote the processing gains for voice and data respectively. Assuming ideal power control, the number of data users permitted to transmit is controlled by the specified probability of outage, P_{out}, given by [143]:

$$P_{out} = Pr\left[\left\{\frac{\gamma_V}{G_V}\sum_{i=1}^{k_V}\nu_i + \frac{\gamma_D}{G_D}\sum_{i=1}^{k_D}\zeta_i\right\} > (1-\eta)\right]$$

where η is a function of the interference and background noise.

To maximize the throughput of delay-tolerant users, a round-robin style access schedule can be applied whereby only a fraction of eligible data users are permitted to transmit at each time [144]. In this way, users can be assigned higher rates but require no more average power than the case where all eligible data users are permitted to transmit. For example, let M_1 and M_2 respectively denote the number of users who are are delay intolerant and delay tolerant. It is shown that the minimum power allocation such that all users meet their target SIRs, assuming unconstrained peak transmit powers, is obtained when:

$$\frac{M_1}{\left(\dfrac{W}{R_1\cdot\gamma_1}+1\right)} + \frac{M_2^{max}}{\left(\dfrac{W}{R_{min}\cdot\gamma_2}+1\right)} < 1$$

where W is the system bandwidth, γ_1, γ_2 are the respective SIR targets, M_2^{max} is the maximum number of delay tolerant users that can be supported if each uses a minimum rate of R_{min}. It is then shown that the throughput can be enhanced (doubled or even higher) by permitting only a fraction of the delay-tolerant users to transmit, but at rates $R_2 > R_{min}$, such that:

$$\frac{M_1}{\left(\dfrac{W}{R_1 \cdot \gamma_1} + 1\right)} + \frac{k_2}{\left(\dfrac{W}{R_2 \cdot \gamma_2} + 1\right)} + \frac{M_2 - k_2}{\left(\dfrac{W}{R_0 \cdot \gamma_2} + 1\right)} < 1$$

where k_2 are the delay tolerant users transmitting at any time and with rate R_2, and the remaining $(M_2 - k_2)$ users are simply maintaining synchronization with the base using an idle rate $R_0 < R_{min}$.

Given the emergence of multimedia traffic with varying QoS requirements, such approaches are not readily adaptable to meet this demand. Secondly, because channel variation is not accounted for, non-realtime users may experience excessive delays.

Opportunistic (‘greedy‘) access control

Allowing for variable rates is more representative of the demands by multimedia traffic. It also affords the access control more flexibility in exploiting the delay-tolerance of active users so as to enhance the ‘soft capacity‘ of a CDMA network while ensuring that target SIR are met.

With the flexibility of variable rate allocation, the objective is to maximize network throughput while minimizing the energy per bit requirements. This is generally achieved by allocating transmission rates in proportion to the channel quality and results in an opportunistic or ‘greedy‘ system whereby, at each time frame, users with good channels get high data rates while users with poor channels do not transmit [47, 146–151]. Because previous rate allocations or terminal backlogs are not considered, a few users may get high rates in consecutive frames while others suffer from poor channels. Opportunistic access solutions typically assume that the network has infinite buffers or that the length of the scheduling window is such that it is considered likely that a time-varying channel may improve conditions for users experiencing poor channels.

In some access control schemes, it is the target SIR rather than the data rate that is variable. Typically, the focus is on delay-tolerant sources which are expected to handle assigned target SIR values via coding or data retransmission, if necessary [152–154]. For example, consider an M-user system with bandwidth W and data rate R_i for each terminal i. Subject to constraints on the transmission power, the objective is then to select the target SIR, γ_i, for each terminal so as to maximize the total information rate, R_T, given by:

$$R_T = \sum_{i=1}^{M} R_i f\left(\frac{\gamma_i W}{R_i}\right)$$

where $f(x)$ is an increasing function of x, and is convex for small x and concave for large x [152]. If the number of users, M, is large or the bandwidth is large ($W >> \sum_j R_j$), then the optimal solution is shown to be close to a 'proportional' solution:

$$\frac{\gamma_i^*}{R_i} \approx c, \ \forall i, \quad c \text{ is a constant,}$$

whereby, the optimal SIR is proportional to the user data rate. In cases where $W/\sum_j R_j$ is not large enough, the proportional SIR allocation results in a violation of the transmission power constraints, and hence, some users will be assigned their minimum SIR target values or blocked from transmission during the current time frame. It is important to note that access control schemes that advocate for retransmissions or extra coding as an integral part of the scheme would generally not be recommended for GEO satellite networks due to the additional delays they introduce into the network.

Assuming stationary channels and with no constraints on the delay or power, the optimal policy for maximizing system throughput is, in fact, by scheduling one user at time [47, 151]. One-by-one scheduling also minimizes the time taken for all users to complete transmission of their data. Still the question remains as to how users should be selected to transmit. Similar in essence to the round-robin scheme, and still assuming stationary channels, a ‘relatively-best‘ scheme has been proposed in which the transmission schedule is ordered according to channel quality [155]. The time fraction to transmit may be the same for all users, as is the case with round-robin, or it may vary for different users depending on the throughput requirements.

In summary, while it is usually assumed that the scheduling window is large enough to permit all active users to complete transmission of the data, opportunistic scheduling can result in excessive backlogs or long delays for users with poor channels and/or in cases where there are several users wishing to transmit. To achieve a more sociable response among users, techniques including utility based pricing and game theory have been proposed whereby each user seeks to maximize its utility or level of QoS satisfaction within the network [136, 138, 156]. These concerns are a core issue of the QoS specification when ‘fair‘

access control is considered, as we do in this thesis. Generally, the objective is to ensure that all active users achieve QoS requirements in form of an average performance metric over a finite scheduling period or window.

“Fair” access control

In multimedia wireless networks, cell/frame-level access control has to contend with a multitude of services with varying QoS requirements as well as with time-varying channel conditions. This raises the question of how to conduct *fair* resource management such that user QoS requirements are satisfied in spite of the propagation conditions. *Fairness* is variably defined within the literature, but we note that fairness *per se* is not an operating objective. This is because, as also pointed out by Hayes *et al.* [157], fairness cannot be a direct measure of QoS as perceived by the user since a user is not directly aware of such fairness. In this section, we consider ‘fair‘ access control as that in which users are denied service during bad channel conditions but are compensated when their channels improve — such that in the long-run, each user is served according to their rate and delay requirements.

One approach to the ‘fair‘ wireless networking problem is to formulate it as an adaptation of wireline fluid fair queueing (FFQ) whereby the objective is to provide long-term fairness to terminals experiencing poor channel conditions, while maintaining short-term fairness for channels experiencing ’clean’ channels [158–165]. In FFQ, for an arbitrary time window $[t_1, t_2]$ during which any two terminals are backlogged with $X_i(t_1, t_2)$ bits for each flow i, the service received, R_i, in terms of bits drained is such that [61, 164, 165]:

$$\frac{X_i(t_1, t_2)}{R_i} = \frac{X_j(t_1, t_2)}{R_j}, \ \forall i, j.$$

However, such FFQ instantaneous fairness cannot be achieved in wireless networks because it may not always be feasible to transmit as scheduled. Furthermore, in FFQ, flows that do not transmit during a given interval are not compensated later on.

In adaptations of FFQ for wireless networks, the system distinguishes between service missed due to no backlog and service missed due to channel error. Typically, the service status is managed via a log of *leads* (service received above that of a virtual fair-queueing model) and *lags* (service lost relative to the virtual fair-queueing model). Compensation is then provided to address service missed due to channel error. In essence, the goal is to weight the users based on tolerable delay and transmission rate instead of simply the rate as is the case with FFQ [161, 164]. How the compensation is applied marks a primary difference among the solutions proposed, and in general the problem is to provide compensation without undue penalty to terminals that have previously received good channels.

A natural approach for the formulation of 'fair' access control problems is provided by dynamic programming (DP). DP solutions satisfy the Bellman's principle of optimality such that, independent of allocations made in previous timeframes, subsequent allocations constitute an optimal policy over the rest of the window [166]. At each time interval, the applied control is selected based on state feedback and in consideration of the impact on the overall system performance given the uncertainty of future propagation conditions.

For finite-sized windows, the optimal access control is generally shown to be a threshold-type policy that seeks to balance the gains of opportunistic one-by-one scheduling with the transmission demands due to the delay bound requirements. The threshold can be characterized by factors such as the backlog, number of active users, and/or level of interference [50, 66, 67, 69, 154, 167, 168]. For example, for single-user operation, a critical backlog state, x_s^*, can be associated with each channel state, s, where x denotes the backlog [154]. At each interval, the amount of data transferred, $u \leq x$, is given by:

$$u^* = u^*(x,s) = \begin{cases} 0, & \text{if } x \leq x_s^* \\ x - x_s^*, & \text{if } x_s^* \leq x \leq x_s^* + P_{max}/w^s \\ P_{max}/w_s, & \text{if } x > x_s^* + P_{max}/w^s \end{cases}$$

where P_{max} is the maximum transmission power, and w^s is a function of the power and rate required in a given channel state and decreases with improving channel quality. In between the extreme cases of $x^* = 0$ and $x^* = x_{max}$, the critical backlog value increases monotonically with increasing tolerable average delay.

Alternatively, the threshold can be characterized by the channel conditions at a given backlog state [69]. For a given backlog, b, the optimal power allocation at each interval, n, is then obtained as:

$$p_n^*(i,b) = \begin{cases} \frac{1}{\alpha}\left(\sqrt{\beta X_n(b)i} - \beta i\right), & i < \dfrac{X_n(b)}{\beta} \\ 0, & i \geq \dfrac{X_n(b)}{\beta} \end{cases}$$

where $X_n(b)$ is a function of the minimal expected cost (averaged over the interference) if the system were to evolve from time $n+1$ with backlog b to time N, and $s(p,i)$ is the probability of successful transmission given power, p, and interference, i such that:

$$s(p,i) = \frac{p}{\alpha p + \beta i}, \ \alpha \geq 1, \ \beta > 0.$$

The challenge with fair access control is the complexity required to obtain optimal solutions, particularly in real time. To address this problem, one approach is to obtain the optimal control policy considering various combinations of user QoS requirements and channel conditions. This policy is then maintained as a look-up table that can be consulted as the system operates in real time [154]. Alternatively, the optimal policy of a simplified user network can be used to provide insight into an efficient operating policy for the regular network [67].

2.5 Concluding Remarks

Ka band frequencies offer the opportunity for broadband satellite networking. The high frequencies enable the generation of high-gain spot beams that allow for the use of cost-effective earth terminals. As has been described, beam management and access control enhance system efficiency in terms of improved system throughput or average throughput per user.

In summary, the two problems that will be addressed in this thesis are to develop a flexible beam configuration algorithm to shape the beams based on user distribution and propagation conditions (Chapter 3) and to develop ‘fair‘ access control schemes that balance the throughput gains of opportunistic scheduling with user performance metrics averaged over a finite-sized scheduling window (Chapter 4– 5). The thesis extends existing work by considering variable power and rate allocations for the beam management problem and by considering a general multiuser network for the ‘fair‘ access control problem in which channel conditions are time-varying and there are constraints on the maximum transmission power. We consider ‘fairness‘ in the sense that all admitted users will get the long-run average rate they expect. Responsibility for ensuring that the average-rate allocation is feasible lies with the admission control, which is outside the scope of this thesis.

Chapter 3

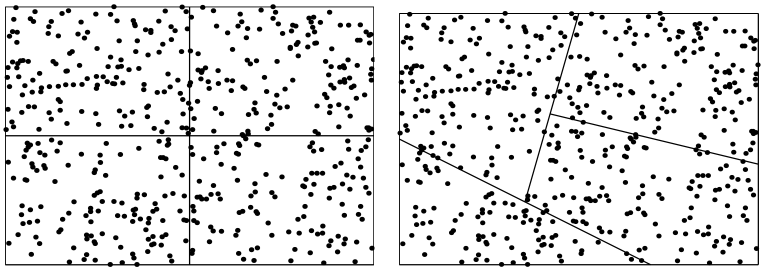
Beam Configuration Management

3.1 Introduction

With the emergence of Internet and multimedia services, satellite systems have evolved to support broadband networking incorporating high speed data transmission as well as intelligent switching and routing. Field trials have demonstrated that the demand for multimedia services can be addressed by multibeam satellite communications at the Ka band (30/20 GHz) and beyond [10]. The key advantages of operation at such high frequencies are the increased availability of bandwidth as well as the potential for smaller low cost terminals for fixed and untethered satellite access.

Two key features of a broadband satellite system are on-board processing and antenna beam allocation. With on-board processing, the satellite uplink and downlink can be optimized separately with respect to multiplexing, modulation and error control. On-board beam allocation can help to meet the link availability requirement at all locations within the coverage area, particularly given inhomogeneous spatial user distribution. Furthermore, adaptive beam allocation can be used to mitigate the effects of rain attenuation, a particularly severe problem in tropical regions.

The design of beam coverage areas for broadband satellite systems must address the need to support an inhomogeneous spatial and temporal user distribution as well as a wide range of quality of service requirements. This calls for a beam configuration system that is responsive to the inhomogeneous user distribution and traffic demand while offering a minimum gain value to all users within the satellite service area. Two approaches have been considered for this problem. Both approaches consider dynamic allocation of satellite resources. However, in one case the beam allocation is fixed and in the other, the beam allocation is adaptive. While dynamic allocation of satellite capacity in a fixed beam system enhances the network efficiency compared to a traditional "fixed beam – fixed resources" system, the achievable throughput and the flexibility of resource management are constrained by the fixed beam geometry. In addition to dynamic resource allocation, broadband multimedia satellite systems require reconfigurable beam coverage areas in order to support inhomogeneous spatial and temporal demands for the satellite services [10, 169]. Figure 3.1 depicts a four-beam allocation scenario for a uniform (fixed) and for an adaptive beam configuration.



(a) Uniform beam allocation (b) Adaptive beam allocation

Fig. 3.1 Satellite beam configuration

We focus on adaptive beam management as a means of effective utilization of satellite resources, given the user distribution and quality of service requirements. In particular, we consider beam management that seeks an efficient distribution of users among the beams by tuning the shape of the spot beams to reflect user distribution. The challenge with adaptive beam allocation schemes is the computational complexity of the allocation algorithms to ensure that all terminals within the satellite service area are provided with sufficient gain. The beam allocation algorithms should define beam sizes that are large enough to minimize the number of beams (and hence the on-board antenna payload) required, yet small enough to meet minimum satellite receiver gain requirements so that the multimedia mobile and fixed terminals can transmit with reasonable power allocations.

The adaptive beam allocation problem has been considered in [10, 25, 63]. The solutions proposed include one in which beam contours are continuously shifted in the neighborhood of a new terminal and/or an overloaded beam until a beam configuration that supports all terminals is obtained [63]. A homogeneous set of voice traffic terminals is considered. A second proposal allocates the downlink beam coverage based on the user distribution and required rain margin, given the constraints on satellite output power and the need to minimize the number of beams [10, 25]. Beginning from one end of the satellite service area, beams are allocated in keeping with the design constraints and considering that the transmission to all terminals is at maximum power.

The contribution of this chapter is the development of a flexible beam allocation algorithm in which allocation optimization parameters (such as beam size, data rates, user distribution, satellite service area) can be varied in order to generate efficient beam allocation patterns. We assume that the number of beams is fixed and the shape of each beam is controlled by the satellite. Our focus is on the relationship among rate, power and beam configuration. Satellite delay constraints are ignored to the extent that the beam management is viewed as a slow process. The goal is to design a beam-shaping algorithm — an algorithm which, on the basis of the demand specification and the prevailing constraints QoS on power and signal-to-interference ratio (SIR), computes a set of beam contours. The effectiveness of any such algorithm is measured by the associated tradeoff between rate and power; for example, by the average power required to achieve a given average rate.

There are two approaches to characterization of the user distribution — *deterministic*, based on the actual geographical location of active terminals, and *statistical*, based on a probability density function defined over the coverage area. Statistical data is easier to come by in practice, and is more stable in the sense of being less susceptible to significant time variation. We assume here a statistical user distribution that is governed by an inhomogeneous spatial Poisson process.

Assuming that users are homogeneous with respect to quality of service requirements (SIR and power constraints), we are particularly interested in *equalizer* algorithms. For convenience, we assume that beams are disjoint.

Definition. *An equalizer algorithm in the deterministic case ensures that the numbers of active terminals in the various beams are approximately the same; in the statistical case, the action is to equalize the corresponding means.*

We assume that the equalizer beam configuration results in realizable antenna systems for the satellite and the terminals, but do not consider the design of such systems. We suggest that the resulting beam configurations can be realized by beam shaping techniques such as those proposed or under development in [24, 28, 29, 33, 34, 170]. In any case, our primary interest is in the throughput advantage to be had from adaptive beam allocation relative to uniform allocation — in which the beams have equal area independent of the distribution of traffic demand.

Note that beam shaping in the satellite context is analogous to base station assignment in cellular wireless, a beam in the one case corresponding to a cell in the other. The two problems are nonetheless different: beam shapes are constrained to be simple and the assignment of terminals to beams is determined by geometry while in a cellular network, at least in principle, the assignment of terminals to base stations is unconstrained – for example, see [140].

In the sequel, we outline the beam configuration problem and present the equalizer algorithm. A comparative performance evaluation of the uniform and equalizer beam allocations is also presented.

3.2 Beam configuration problem

As noted in Section 2.3.3, in order to conduct beam allocation in a WCDMA network, the traffic profile, the channel characteristics and the interference control mechanisms should be considered. In our work, we assume that the traffic and the channel characteristics are of known statistical distribution. The interference control is realized by ensuring that the target SIR is met.

We consider the following beam configuration model. Assume that the geographic coverage area can be quantized to a grid $\{(x\Delta, y\Delta), x, y \text{ integers}\}$ resulting in equal-sized blocks across the coverage area. Discretization of the problem eases the computational complexity associated with mapping the users (or their means, in the statistical case we consider here) to the satellite service area. The coverage area of a spotbeam is defined by the number of blocks where the antenna gain exceeds some threshold value. We shall refer to the point of maximum gain within the beam coverage area as its beam center. Figure 3.2 depicts a beam configuration of four beams where d_j marks the distance between the center of the jth block and the center of the beam in which the block is located. Each block within the quantized service area is ascribed a parameter λ denoting the expected number of users, which we assume is governed by an inhomogeneous spatial Poisson process.

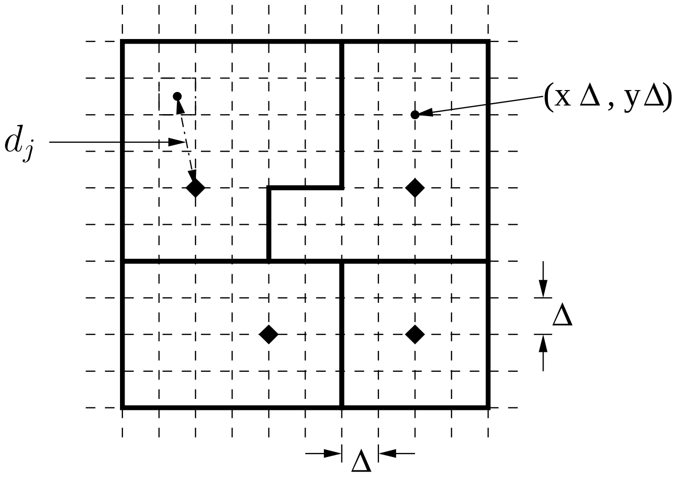


Fig. 3.2 Four-beam allocation over quantized coverage area

The problem is then to allocate beams across the grid so that the various beams have about the same statistical mean number of users and the beams are as small as feasible. As shown in Section 2.3.1, the smaller the beam size, the higher the beam gain and hence the smaller the power required per terminal. However, very narrow beams might mean that more beams are needed if all regions of the satellite service area are to receive the minimum required gain.

In the following section we present the design of an equalizer algorithm and a comparison of its performance relative to uniform beam allocation. The metrics for the performance evaluation are the realized SIR and the rates assigned to active terminals. The system performance is controlled by the assigned power and the beam geometry.

3.3 Adaptive beam allocation algorithm — *Equalizer algorithm*

The principal elements of the equalizer algorithm are a beam allocation function and a search method to scan the grid for potential locations at which to center the beams. Beams are immobile relative to short-term fluctuations such as the traffic demand at call or burst level, user mobility, or channel fading.

3.3.1 Equalizer optimization criterion

In the proposed algorithm, we search for a set of beam shapes that will minimize a weighted sum of two beam allocation functions. The first function is a measure of the beam sizes while the second function is a measure of the statistical distribution of users among the beams. The number of beams is fixed at N.

As a measure of beam size, we use a function of the distance between a beam center and each of the terminals assigned to the beam. Let r_i denote the position of terminal i within the grid, where r_i is a random variable. For $x, \beta(x)$, where $\beta(x)$ is the center of the beam containing x, we define a measure of the beam sizes as

$$\mathbf{E}\left[\sum_i \|r_i - \beta(r_i)\|\right], \tag{3.1}$$

where the expectation is with respect to the inhomogeneous spatial Poisson distribution.

Remark. *Equation (3.1) does not represent a beam size. Rather, it is more a function of the population size and spread within a beam, averaged over the beams.*

Supposing that there are M blocks across the grid, we can rewrite Equation (3.1) as

$$\mathbf{E}\left[\sum_{j=1}^{M} \sum_{i \in \text{block} j} \|r_i - \beta(r_i)\|\right].$$

Let d_j denote the Euclidean distance between the center of block j and the center of the beam containing block j. We assume that each block, receiving satellite coverage, is wholly contained within a single beam. If Δ is so small that $\|r_i - \beta(r_i)\| \approx d_j$ for all terminal i in block j, then Equation (3.1) reduces to

$$\sum_{j=1}^{M} d_j \mathbf{E}[\text{terminals in block } j] = \sum_{j=1}^{M} d_j \lambda_j.$$

Beam contours are determined from beam centers by a nearest-neighbor rule: each block is associated with the center closest to it in the sense of Euclidean distance. If the shortest distance is the same for several such centers, the block is assigned randomly to one of them. The size of each beam is constrained by the minimum gain to be provided, where the minimum gain required is set by link parameters such as the satellite altitude, the transmission rate, the propagation losses, and the required signal-to-interference ratio. Let d_{th} (threshold distance) be defined by the property that all users within d_{th} of the beam center will receive at least the minimum gain value. Let vector $d = (d_1, \dots, d_M)$. We define the beam size measure, $f_1(d)$, by

$$f_1(d) \triangleq \sum_{j=1}^{M} d_j \lambda_j + \theta_d \sum_{j=1}^{M} (d_j - d_{th})^+, \tag{3.2}$$

where $(x)^+ \triangleq \max(o, x)$ and θ_d is the penalty incurred for exceeding the threshold distance.

The function $f_1(\cdot)$ forms one element of our optimization criterion. The second element of the optimization criterion brings in the equalizer property. Given an allocation of blocks to N beams, let the statistical mean number of users per beam be denoted by vector $\Lambda = (\Lambda_1, \dots, \Lambda_N)$. The more balanced the beam allocation, the smaller the sample variance of Λ, which we denote as $v(\Lambda)$. Based on channel conditions and power constraints, as we shall discuss in Section 3.4.1, let U_{max} denote the largest (statistical mean) number of users acceptable for any beam. Let U_{min} denote the smallest mean number of users acceptable for any beam. Write μ_{min}, μ_{max} for penalty factors associated with violation of the U_{min} and U_{max} thresholds. We then define a second function as

$$f_2(\Lambda) \triangleq v(\Lambda) + \mu_{min}(U_{min} - \min(\Lambda))^+ + \mu_{max}(\max(\Lambda) - U_{max})^+. \tag{3.3}$$

Definition. *The beam configuration problem is the search for a set of N beam shapes which together minimize the weighted sum*

$$\min_{d,\Lambda} f(d, \Lambda) = \varphi f_1(d) + (1 - \varphi) f_2(\Lambda), \tag{3.4}$$

where $\varphi \in (0, 1)$ is a design parameter which weights the impact of $f_1(\cdot)$ and $f_2(\cdot)$ on $f(\cdot)$.

3.3.2 Equalizer beam contour selection

In selecting a beam configuration, an exhaustive search for a solution to Equation (3.4) is not computationally feasible. In the solution we propose, the equalizer algorithm employs a beam allocation function and a Hooke and Jeeves search function [65, 171] to explore the coverage area for N beam center positions that optimize Equation (3.4).

We consider two alternatives of the beam allocation function for the equalizer algorithm presented in Algorithm 2. Algorithm 1 allocates terminals to beams based on a shortest-distance rule, where the distance is given by the Euclidean distance between the center of a block and a beam center. Algorithm 3 takes the results of Algorithm 1 as its starting point and then adjusts the beam center position so as to minimize the average distance between the beam center and the terminals assigned to the beam.

The equalizer search begins from the set of N beam center positions for a uniform beam allocation. If an exploratory direction is an improving direction with respect to Equation (3.4), an iterative search in the exploratory region is conducted to obtain a minimal solution to the optimization criterion. This iterative search is repeated as further improving directions are identified. The algorithm terminates when no further search directions improve the current solution. Note that with the Hooke and Jeeves method, the optimization criterion is not required to be continuous or differentiable.

3.3.3 Equalizer algorithm

The objective of the equalizer algorithm is for an evenly distributed beam configuration with respect to Equation (3.4). Let $L = [l_1, \ldots, l_N]$ denote the locations of N beam centers, $l_1, \ldots, l_N$. The initial L is given by the beam centers obtained from a uniform allocation. For each block j of the grid structure, we assume that a single terminal i located at the center of the block is responsible for handling all the block traffic. A_k is the set of indices of those terminals assigned to beam k, $k = 1, \ldots, N$, and is of length M_k.

Given L, Algorithm 1 presents the non-iterative *beam_allocate* function used to distribute terminals among the beams. Each terminal is assigned to a beam with the shortest distance between the terminal and the beam's center. In Algorithm 1, r_j represents the position of the center of block j, $j = 1, \ldots, M$ and $D_j = (d_{j,1}, \ldots, d_{j,N})$ represents the separation (Euclidean) distance of block j from the centers of beams $1, \ldots, N$. Given the shortest-distance vector $d = (d_1, \ldots, d_j, \ldots d_M), j \in A_k$, obtained as a result of the beam allocation operation, we define $F(L)$ as the weighted sum $f(d, \Lambda)$ given beam centers L — where $f(d, \Lambda)$ is given by Equation (3.4).

The equalizer algorithm is an iterative algorithm that utilizes the *beam_allocate* function

Algorithm 1 *beam_allocate* function

```
Objective: Allocate terminals to beams based on nearest distance to beam center.
  for $k = 1$ to $N$ do
    for $j = 1$ to $M$ do
      $d_{j,k} \leftarrow \|r_j - l_k\|$ {ordinary Euclidean norm}
    end for
  end for
  $j \in A_k \iff \min D_j \leftarrow d_{j,k}$
  $F(L) \triangleq f(d, \Lambda),$ given L
```

and a search function to approach the objective of an evenly distributed beam allocation. Beginning with the beam center locations, L_u, for a uniform beam allocation, the equalizer algorithm proceeds in search of a set of beam center locations, L, such that $F(L) \leq F(L_u)$. Algorithm 2 presents the equalizer algorithm, in which we apply the following definitions:

- accuracy bound, $\varepsilon > 0$
- step size vector, $h_z, \ z \in \{1, 2\}$
- +/- direction vector, $s_q, \ q \in \{1, 2\}$
- search direction vector at extrapolation t, $v(t)$
- progress marker, $p \in 0, 1$

Note that in order to ensure that the search is restricted to the coverage area, whenever a point outside the region is identified as a possible search point, it is replaced by its nearest valid point (on a coordinate-by-coordinate basis).

As given in Algorithm 1, the *beam_allocate* function obtains the shortest distance between the blocks and a given set of beam centers, L. An option is to consider further reduction of the average separation distance between blocks and their beam centers by obtaining an optimal location for each of the beam centers after the *beam_allocate* operation. For the same allocation $A_k, k = 1, \ldots, N$, such an option would yield a lower value of the beam size measure $f_1(d)$ given by Equation (3.2).

Algorithm 3 presents the *beam_center* function that can be used in place of *beam_allocate* in the equalizer algorithm. Given an allocation $A_k, k = 1, \ldots, N$, the *beam_center* function obtains the beam center coordinates that minimize the average weighted separation distance between the terminals and their respective beam centers.

Algorithm 2 Equalizer algorithm

```
Define: $\hat{l} \triangleq$ test point, $\tilde{l} \triangleq$ new point, $l(t) \triangleq$ best point at iteration $t$
Initialize: $z = 1,\ t = 1,\ f_t(L) = F(L_u),\ f_{new} = F(L_u),\ k = 1,\ \hat{l} = l_1,\ l(t) = l_1$. {Repeat
  for $k = 2$ to $N$}
  for q = 1:2 do {SEARCH along the direction of the axes for beam center $\hat{l}$}
    $\tilde{l} \leftarrow \hat{l} + s_q \cdot h_z$
    $L : l_k \leftarrow \tilde{l}$
    beam_allocate
    if $F(L) < f_{new}$ then
      $\hat{l} \leftarrow \tilde{l}$
      $f_{new} \leftarrow F(L, l_k = \tilde{l})$
    else
      $\tilde{l} \leftarrow \hat{l} - s_q \cdot h_z$
      $L : l_k \leftarrow \tilde{l}$
      beam_allocate
      if $F(L) < f_{new}$ then
        $\hat{l} \leftarrow \tilde{l}$
        $f_{new} \leftarrow F(L, l_k = \tilde{l})$
      end if
    end if
  end for
  if $f_{new} < f_t(L)$ then {Move forward. Adjust parameters and repeat
  SEARCH.}
    $t \leftarrow t + 1$
    $l(t) \leftarrow \hat{l}$
    $v(t) \leftarrow l(t) - l(t-1)$
    $[s_1\ s_2] \leftarrow \mathbf{sign}(v(t))$
    $f_t(L) \leftarrow f_{new}$
    $\hat{l} \leftarrow l(t) + v(t)$ {Coordinates $(x\Delta, y\Delta)$ in coverage area}
    $p \leftarrow 0$
  else {Move back}
    if p = 0 then {Adjust parameters and repeat SEARCH}
      $p \leftarrow p + 1$
      $\hat{l} \leftarrow l(t)$
    else {Check step sizes}
      if $h_z \leq \varepsilon, z \in \{1, 2\}$ then {Terminate condition}
        $l_k \leftarrow l(t)$
        STOP
      else {Adjust step sizes and repeat SEARCH}
        $h_z \leftarrow h_z/2,\ z \in \{1, 2\}$
      end if
    end if
  end if
```

Algorithm 3 *beam_center* function

```
Objective: Minimize average separation distance per beam.
  beam_allocate {Allocate terminals to nearest beam center}
  for $k = 1$ to $N$ do {Obtain new $L$}
    $S_k \leftarrow \sum_{m \in A_k} \lambda_m$
    $l_k \leftarrow \frac{1}{S_k} \sum_{m \in A_k} \lambda_m r_m$
  end for
  $F(L) \overset{\Delta}{=} f(d, \Lambda),$ given L
```

3.3.4 Evaluation of the equalizer algorithm

The point of the equalizer algorithm, as noted, is to equalize the user distribution among beams. Also as noted, our ultimate concern is with the rate/power tradeoff. It seems reasonable, as an auxiliary step, to evaluate to what extent equalization is achieved. To that end, we introduce the Beam Variability Factor (BVF), defined as the ratio[1] of the sample standard deviation to the sample mean

$$BVF = \frac{std(\Lambda)}{mean(\Lambda)}, \tag{3.5}$$

where $\Lambda = [\Lambda_1, \ldots, \Lambda_N]$ is the vector of mean beam populations. In this section we compare the uniform and two equalizer allocations on the basis of BVF, and then in Section 3.4 we evaluate the performance in terms of rate and power.

The more evenly balanced the beam configuration, the smaller the BVF. We compare the allocation performance of the equalizer algorithm to a uniform allocation algorithm in terms of how evenly users are distributed among the beams. We consider three algorithms applied to an N-beam satellite network:

(1) Uniform allocation where the coverage area is divided into N equal regions and a beam assigned to each region.

(2) Equalizer allocation, where N beams are dynamically assigned using the equalizer algorithm with the *beam_allocate* function (Algorithm 2).

[1] Amounts to the square root of the sample-path version of what in the statistics literature is called the “coefficient of variation”.

(3) Equalizer algorithm with the *beam_center* function (Algorithm 3), where N beams are dynamically assigned and within each beam, the beam center location is adjusted so as to minimize the average separation distance between the terminals and the beam center.

The three beam allocation strategies are compared based on a 4-beam satellite network with a coverage area comprising a grid structure of 24 blocks. The comparative performance results are presented in terms of the beam variability factor (BVF). The smaller the BVF, the better balanced the beam allocation, and scheme (1) provides the benchmark allocation performance. It is expected that the equalizer-based schemes should perform better than the uniform scheme.

For the beam configuration function described by Equation (3.4), the parameters used for the experiments are presented in Table 3.1. In the table, Δ_x and Δ_y refer to the dimensions of each block — we consider a general case where the x and y dimensions are not the same. The smaller the beam size, the higher the gain provided and hence the smaller the power required per terminal. Hence, the function $f_1(\cdot)$ was given a slightly higher weight, φ, in Equation (3.4).

Three experiments were conducted with the maximum λ parameter for any block, $\Lambda_{max} \in \{10, 50, 100\}$. For a given experiment, the performance was averaged over 1,000 sets of 24-block λ's generated within the interval (0, Λ_{max}).

Table 3.1 Parameters for allocation experiments

Parameter	Value(s)
M	24
N	4
λ	$0 < \lambda < \Lambda_{max}$
Δ_x	25
Δ_y	16.7
d_{th}	20
θ_d	5
μ_{min}	50
μ_{max}	50
U_{min}	$\sum \lambda_i/N,\ i = 1, \cdots, M$
U_{max}	$\sum \lambda_i/N,\ i = 1, \cdots, M$
φ	0.6

Table 3.2 presents the BVF obtained for the three allocation schemes examined. The number in brackets after BVF indicates Λ_{max}. From the table, we can make the following observations:

1. Both versions of the equalizer algorithm result in better performance relative to the uniform allocation, as shown in Table 3.2.
2. Between the two equalizer versions, the one with the *beam_center* function provides only marginal performance improvement.

Table 3.2 BVF results for different allocation schemes

Scheme	BVF(10)	BVF(50)	BVF(100)
Uniform allocation	0.2774	0.2776	0.2736
Equalizer allocation (with *beam_allocate*)	0.1096	0.1089	0.1099
Equalizer allocation (with *beam_center*)	0.0930	0.0895	0.0911

The results obtained show that the adaptive allocation yields a relatively even users per beam distribution compared to the uniform allocation. In the sequel, we define the adaptive algorithm to be the Equalizer algorithm with *beam_allocate* function. This is because the *beam_center*-based algorithm takes a longer time to process but yields only marginal performance improvement.

In the following section, we characterize the relationship between the control (beam shape, power constraints) on the one hand and the performance (SIR, rate) on the other for a given beam configuration. Again, our primary interest is of a comparative performance evaluation of the adaptive and uniform allocation schemes with respect to rate and power.

3.4 Performance evaluation of beam configuration management

The objective of beam configuration management is to enable the network match the geographic and/or time distribution of the traffic demand over the coverage area to the available satellite resources. Variable-beam configurations provide two degrees of flexibility for beam management, since both the beam shape and the user rates can be adapted in support of efficiency. The ability to match the beam sizes to the demand is additionally significant in CDMA-based networks because it allows the network to influence the interference experienced by any one user.

The network performance measures we consider are signal-to-interference ratio (SIR) and rate. For each beam configuration, performance is obtained in terms of the *achievable rate region*, which is comprised of those rate allocations that meet the target SIR and satisfy the power constraints. Given a set of beam configurations, how can one compare the achievable rate region of the different configurations? To evaluate the performance of a beam configuration, we consider the following criteria:

- average throughput per user and number of active users
- average transmission power per user for a feasible rate allocation, given constraints on the maximum transmission power per user

In general, network performance depends on the form of the SIR model which, in turn, depends on the receiver architecture. In this thesis we consider the single-user receiver architecture. A single-user receiver estimates the signal of the desired user by treating the other users as noise. In the following sections, we describe the resulting SIR model and discuss the criteria by which the achievable rate regions of the uniform and the equalizer configurations are compared. This is followed by a discussion of the comparative performance results obtained.

3.4.1 SIR model

The satellite network has N beams serving a total of M terminals. The results of the beam allocation process earlier described are summarized by an N- vector $\mathbf{A} = (A_1, \ldots, A_N)$, A_k denoting the set of indices of those terminals assigned to beam k. We write M_k for the cardinality of A_k.

We use a simple multiplicative model for the effects of propagation and antenna losses on the signals received at the satellite. Specifically, we assume that a signal transmitted at power P from block ℓ generates power $\alpha_{\ell k} P$ in beam (k) at the satellite. As suggested by the notation, the *attenuation coefficient* α, while depending on originating block and receiving beam, does not (by assumption) depend on the position of the originating terminal within its block. In the ideal case, which is the one we use in our calculations and simulations, $\alpha_{\ell k} = 0$ unless block ℓ lies within beam k.

In line with at least a subset of published literature on related problems (including, for example, [138, 146, 152, 153, 172]), we use a much simplified model for quality-of-reception. Write SIR_i for the *signal-to-interference ratio* associated with satellite reception of transmissions from terminal i. Recall that the assumed signaling format is CDMA. We assume

$$SIR_i = \frac{P_i \alpha'_{ik} W / R_i}{\sum_{j \neq i}^{M_k} P_j \alpha'_{jk} + \sum_{\ell \neq k}^{N} \sum_{j}^{M_\ell} P_j \alpha'_{\ell k} + 1} \qquad (i \in A_k), \tag{3.6}$$

where W is the signaling bandwidth, R_i is the signaling bit-rate, P_i is the transmitted power and

$$\alpha'_{jk} \triangleq \frac{\alpha_{jk}}{\sigma_k^2 W}$$

is the previously defined attenuation coefficient, normalized now by the total power $\sigma_k^2 W$ attributable to noise (as distinct from multi-access interference) in the receiver at beam k.

Remark. SIR$_i$, *as defined by Equation (3.6), is obtained from the error-rate analysis of the optimal single-user detector for extraction of a signal from additive white Gaussian noise. Its application in the present setting is motivated by arguing (1) that the multi-access interference is at least approximately Gaussian if the number of interferers is large and the power attributable to any particular interferer is small relative to the aggregate; and (2) that the effect of the pseudo-random spreading sequences in the CDMA signal is (roughly) to whiten the signal spectrum. There are far more sensitive analyses available (for example, [100, 143, 173–175]), and more sophisticated receiver architectures (including the variety of multi-user detectors that are the subject of much recent attention in the literature (for example, [176–178]). In focusing on network management issues, as opposed to physical layer, we have elected for simplicity in the physical layer models.*

The QoS requirement in terms of SIR is $SIR_i \geq \gamma_i$ for each i, where γ_i is some positive constant. From the analysis of the single-user receiver in white Gaussian noise one can, as noted, relate SIR_i to the probability of bit error for user i, and thus obtain γ_i from information on the character of user i's signal and its tolerance of bit errors. We assume that all the γ_i's are fixed and known to the network management system. Typically we expect that the constraints on SIR will be satisfied with *equality* in any particular realization of the system — simply because inequality would mean that the power levels at the various transmitters in the system are in excess of what is demanded for user satisfaction.

In the case, alluded to above, that there is no leakage of power from one beam to another ($\alpha_{ik} = 0$ when terminal i is outside beam k), the formulas simplify further:

$$SIR_i = \frac{P_i \alpha'_{ik} W / R_i}{1 + \sum\limits_{j \neq i}^{M_k} P_j \alpha'_{jk}} = SIR_i = \gamma_i. \tag{3.7}$$

In the following subsection we develop the relationships among rate, power, attenuation coefficients, and number of users that we use to quantify the comparative performance between the adaptive and uniform beam allocations presented in Section 3.4.2.

Achievable rate region and required transmission power

The problem here is to assess the effect of beam configuration on system capacity, where our notion of system capacity is based on the M-dimensional achievable rate region – the set of user rates that are compatible with the power constraints. We wish to get a sense of the extent to which beam configuration impacts a set of achievable rates.

For an achievable rate region $\mathcal{R}$, let R_i denote the rate of terminal i in beam k. Within an N-beam set, each beam k has a rate and power allocation described by M_k-vectors $R = (R_1, \ldots, R_{M_k})$, $P = (P_1, \ldots, P_{M_k})$. For powers constrained within $\mathcal{P}$, a given rate vector R is admissible ($R \in \mathcal{R}$) if and only if

$$\frac{R_i \gamma_i}{W} = g_i(P, \alpha) \quad (i = 1, \ldots, M_k), \tag{3.8}$$

where α denotes the normalized attenuation vector, $\alpha = (\alpha'_{1k}, \alpha'_{2k}, \ldots)$,

$$g_i(P,\alpha) \triangleq \frac{P_i\alpha'_{ik}}{1 + \sum_{j\neq i}^{M_k} P_j\alpha'_{jk}} \qquad \text{and} \qquad G_i(P,\alpha) \triangleq \frac{P_i\alpha'_{ik}}{1 + \sum_{j=1}^{M_k} P_j\alpha'_{jk}}. \tag{3.9}$$

The admissible rate region $\mathcal{R}$ is given by the range of the function

$$g_k(P,\alpha) \triangleq (g_1(P,\alpha), \ldots, g_{M_k}(P,\alpha)) \quad (P \in \mathcal{P},\ \alpha \text{ fixed},\ k = 1, \ldots, N).$$

after the scaling of each $g_i(\cdot)$ by W/γ_i for all k.

Let us define $\eta_i = G_i(P,\alpha)$, and

$$\frac{1}{\eta_i} = 1 + \frac{W}{R_i SIR_i}. \tag{3.10}$$

We see that η_i increases monotonically with $R_i SIR_i$, and thus $\mathcal{R}$ can be obtained via the set of feasible Γ's; where $\Gamma = (\eta_1, \ldots, \eta_{M_k}),\ k = 1, \ldots, N$. Denote this set of feasible Γ's as $\mathcal{G}$. Since $G(P,\alpha)$ is continuous, $\mathcal{G}$ is connected, closed and bounded. If we continue with the assumption that there is no inter-beam interference, then the achievable rate region is a convex polyhedron [179]. It follows that any linear function of the rate vectors, such as the total or mean rate, assumes its maximum on a vertex of the polygon defined by the appropriate scaling of $\mathcal{G}$, which in turn corresponds to a vertex of the power region. The conclusion is that any throughput-maximizing control is a *bang-bang* control: each terminal is either turned off or signaling at maximum power. Any such control is completely characterized by the list of active terminals (see for example, [148]).

If the user population is homogeneous in terms of the distribution of attenuation coefficients, α, maximum powers and SIR constraints, as we consider here, then the optimal policies (for a given beam configuration) are characterized simply by the *number* of active transmitters, the number of (equivalent) optimal policies being identical to the number of ways in which the active transmitters can be chosen from the total user population. Our interest then is to compare the achievable rate regions of the adaptive and uniform allocations based on the number of active users for given rates.

Within a beam k, the maximum number of users that can be supported, M_k^{max}, corresponds to the situation when each active terminal gets a minimum rate, R_{min}. If we consider that $\mathcal{G}$ is contained within a unit multidimensional cube, from Equation (3.9) we have that $\mathcal{G}$ is bounded by the hyperplanes

$$\sum_{j=1}^{M_k} \eta_j \leq 1, \quad k = 1, \ldots, N.$$

From Equation (3.10), for a homogeneous set of users within a beam k with $SIR_j = \gamma$, for all $j \in A_k$, we then have that

$$M_k^{max} \leq \left\lfloor \left(1 + \frac{W}{\gamma R_{min}}\right) \right\rfloor.$$

For a beam k with M_k terminals, the number of active terminals, M_k^*, is

$$M_k^* \triangleq \min\{M_k, M_k^{max}\}. \tag{3.11}$$

The required transmission power to support a feasible rate allocation is determined as follows. We begin by making the following definitions:

$$\begin{aligned} \overline{\eta}_k &\triangleq \sum_{j}^{M_k} \eta_j \\ \overline{\mu}_k &\triangleq \frac{\overline{\eta}_k}{1 - \overline{\eta}_k} \end{aligned}$$

From Equation (3.9), we have that

$$\eta_i = \frac{P_i \alpha'_{ik}}{1 + \sum_{j=1}^{M_k} P_j \alpha'_{jk}} \qquad \text{or} \qquad P_i = \frac{\eta_i}{\alpha_{ik}} \left(1 + \sum_{j=1}^{M_k} P_j \alpha'_{jk}\right).$$

By solving for P_i in terms of $\overline{\eta}_k$, when $\overline{\eta}_k < 1$, we obtain

$$P_i = \frac{\eta_i}{\alpha_{ik}} \frac{\overline{\mu}_k}{\overline{\eta}_k} \quad \equiv \quad \frac{\eta_i/\alpha_{ik}}{1-\overline{\eta}_k}, \quad i \in A_k. \tag{3.12}$$

In the following section, we present a comparative performance evaluation of the adaptive and the uniform beam allocations based on the relationships developed here.

3.4.2 Comparative performance evaluation: adaptive vs. uniform

Ideally, the shape of the antenna beams should take into account the character of the propagation environment as seen by each user, in addition to the geographical distribution of users over the satellite coverage area. The beam shaping algorithm previously described, based exclusively on the latter, can probably be improved. We are nonetheless interested, its limitations notwithstanding, in determining the extent of its impact on system capacity. System capacity, as we define it for our purposes here, is identified with the achievable rate region in M space, where M (as before) is the number of users. A rate vector is achievable provided that it is compatible with the SIR and power constraints. The smaller the number of users within a beam, all other things being equal, the larger the rates that can be assigned to individual users. So one would expect that a shaping algorithm that tries to balance the number of users among the various beams might have some beneficial effect on rate. Here we try to quantify that effect, using the relationships among rate, power, number of users and fading coefficients summarized in the previous sub-section.

We assume that the user population is homogeneous in respect of SIR requirements and power constraints. The spatial distribution of users is modeled by a two-dimensional inhomogeneous Poisson process. Our numerical experiments compare two beam shaping algorithms — the uniform one, and the one calculated by the algorithms previously described — on the basis of

- Beam Variability Factor (Equation (3.5))
- Achievable rate region (Equation (3.11))
- Required transmitter power (Equation (3.12))

The model selected for the numerical experiments was inspired by the European multimedia Ka band satellite network [12], [15, pp. 324]. We used the parameter values in Table 3.3.

Table 3.3 Parameters for performance evaluation experiments

Parameter	Value(s)
Uplink carrier frequency	30 GHz
Number of beams, N	4
Number of blocks in coverage area, M	24
Individual terminal rates	160 kbps, or 512 kbps, or in the range 1 to 6 Mbps

The Poisson means for each of the M blocks were drawn randomly (and independently from block to block) from the set $\{10, 20, 50, 200, 500\}$. The corresponding Poisson variables were used to generate the user population in each block. We ran the experiment 10,000 times and averaged the results to produce the tables and graphs included below.

The first parameter of interest is the beam variability factor (BVF). Table 3.4 presents the average BVF values obtained for the adaptive and the uniform beam configurations. As expected, the adaptive allocation has the lower BVF and we observe that the equalizer algorithm results in an improvement of about 70% in terms of balancing the users among the beams.

Table 3.4 BVF results for an adaptive and a uniform beam configuration

Scheme	BVF
Uniform allocation	0.6527
Adaptive allocation	0.207

In the following sub-sections we present the comparative performance in terms of the achievable rate region and the minimum power required to transmit a given rate allocation.

(1) Achievable rate region

Recall that the users are assumed homogeneous with respect to SIR and power constraints and with respect to the distribution of attenuation, α. We distinguish two cases. In the first case, all users transmit at a fixed rate R_{min}, and performance is evaluated in terms of the mean number of users permitted to transmit. In the second case, where permitted by the SIR and power constraints, some or all active users will have transmission rates greater than R_{min}. In that case we also measure the achieved throughput. In both cases, we assume a single-user receiver architecture at the satellite and ignore interbeam interference. Transmission is scheduled on a block-by-block basis: all terminals within a block are either transmitting or turned off.

In the first case, we are interested in the *active user ratio* which relates the mean number of users that are permitted to transmit to the mean user population. Figure 3.3 reveals that as R_{min} increases, the active user ratio drops for both uniform and adaptive algorithms. However, the adaptive (equalizer) algorithm results in a larger achievable region since it enables a greater ratio of mean users to transmit.

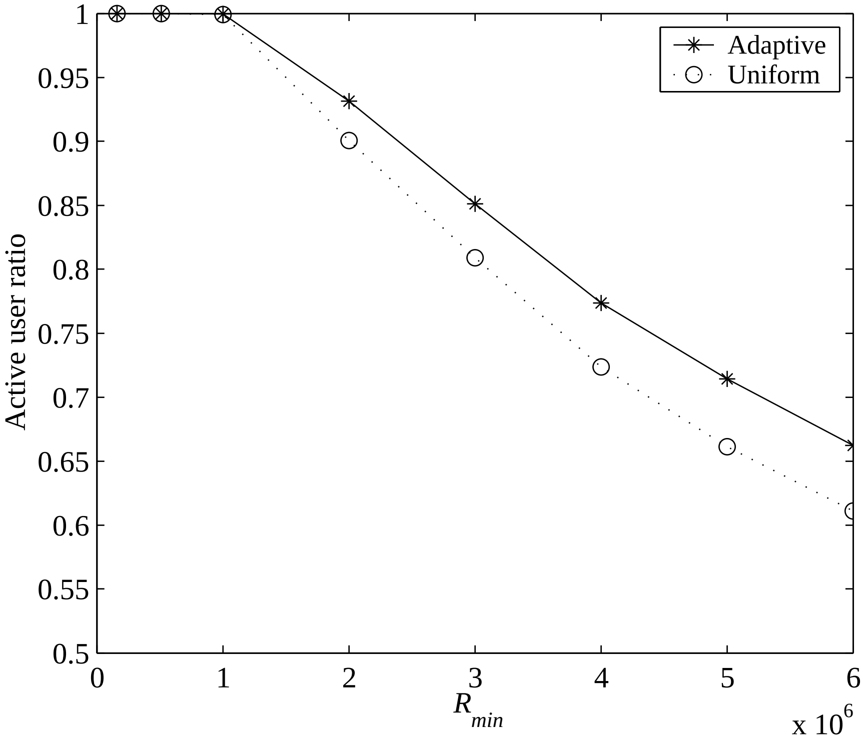


Fig. 3.3 Average user ratio versus R_{min}

In the second case, in general, the fewer the mean number of users in a beam, the more likely that all will get to transmit and possibly at higher rates too. Since the active user ratio versus R_{min} remains as shown in Figure 3.3, one may expect that the adaptive allocation performs poorly in terms of average throughput per active user because it balances the user distribution among the beams and has a higher active user ratio. In fact, from Table 3.5, we observe that the average rate per active user is basically the same, even while the adaptive algorithm has a higher mean active user ratio. It seems that adaptive beam allocation — and the associated beam geometry — can be useful in expanding the achievable rate region for a satellite network.

Table 3.5 Average throughput per active user for $R_k \geq R_{min}$ case

R_{min} (Mbps)	0.16	0.512	1	2	3	4	5	6
R_k (*Uniform*)	0.16	0.5121	1.0002	2.0006	3.0009	4.0014	5.002	6.0025
R_k (*Adaptive*)	0.16	0.5121	1.0002	2.0006	3.001	4.0015	5.0019	6.0023

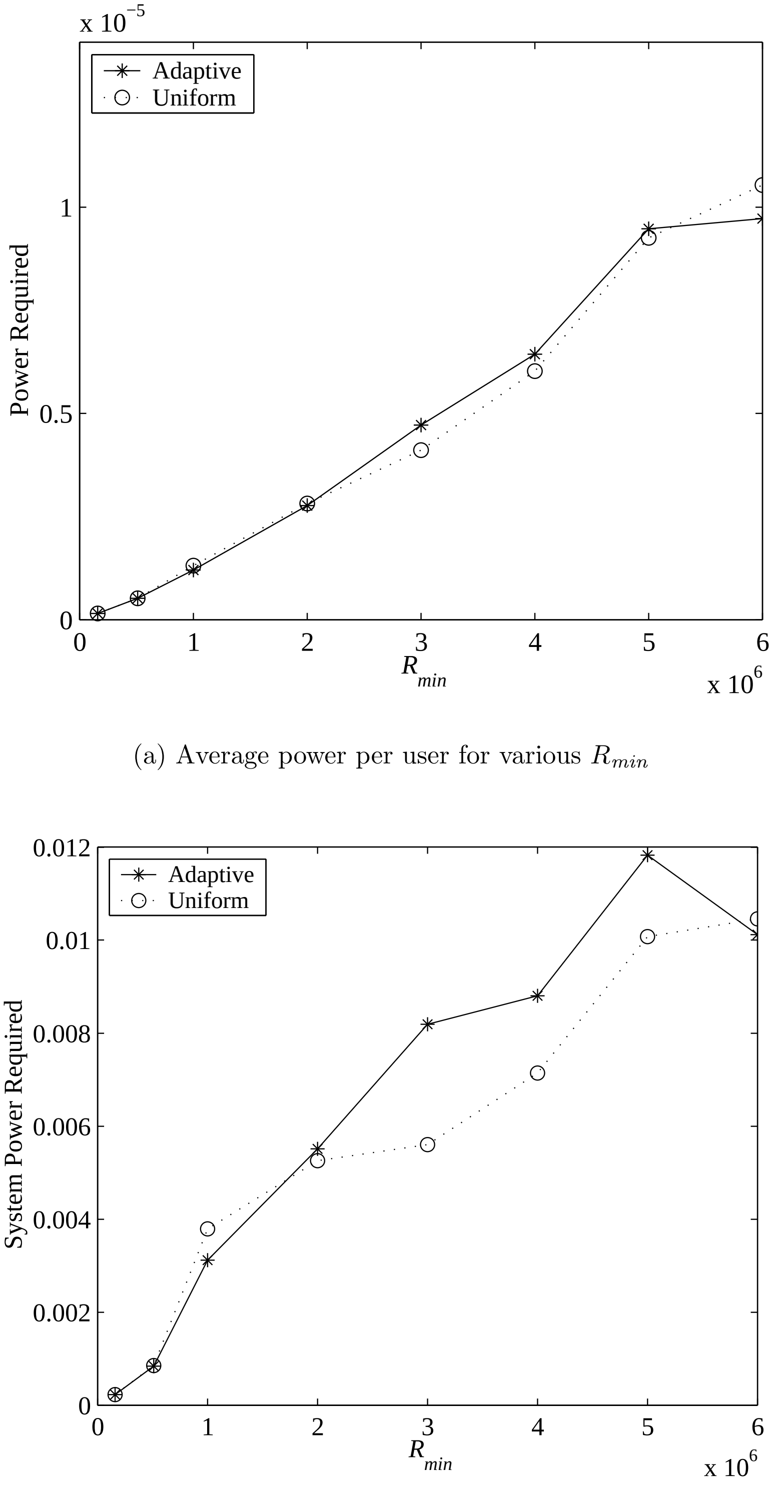
(2) Transmission power required for given rate allocation

In this case we consider the minimum power required when the transmission rate is fixed at R_{min} Mbps. In Figure 3.4, we observe that the same or slightly higher transmit power per user is required for the *adaptive allocation* - a notable exception is at $R_{min} = 6$ Mbps.

Since more users are carried in the *adaptive allocation*, it can be expected that the total system power will be higher in this case. Indeed we notice that in the case of $R_{min} \leq 1$ Mbps, where all users are active for both allocations, the *uniform allocation* actually requires more or the same power than the *adaptive allocation* case. We observe then that the gains achieved by the equalizer algorithm in terms of expanding the achievable rate region (Figure 3.3) are not accompanied by a significant increase in the power requirements (Figure 3.4(a)).

3.5 Concluding Remarks

We have presented a flexible beam allocation algorithm that combines the requirement to provide a minimum satellite gain value with the need to provide a statistically balanced beam configuration for a homogeneous set of users. By replacing the beam function $f_2(\cdot)$



(a) Average power per user for various R_{min}

(b) System power for various R_{min}

Fig. 3.4 User and system power requirements versus R_{min}

in Equation (3.3), one can generate alternative selection criteria, say in the case of inhomogeneous users.

In the case considered here for homogeneous users, use of the *adaptive allocation* results in a higher active user ratio and hence larger achievable rate region. While the system power requirements are also higher, we observe that the extra transmit power per user required in the *adaptive allocation* case is not significant.

In conclusion, adaptive beam allocation has been shown to enhance the efficiency of satellite resource utilization by enabling a higher active user ratio without degrading the average user throughput. The evaluation and results presented clearly illustrate that beam geometry and transmitter power are important controls in expanding the achievable rate region for a satellite network.

Chapter 4

Access Rate Scheduling

4.1 Introduction

Rapidly growing demand for interactive multimedia services, including web browsing, bulk data transfers and video services, provides impetus to the development of versatile broadband networks capable of providing cost-effective direct-to-user services to fixed and mobile users irrespective of location [10, 15–19]. Satellite systems are especially efficient for multimedia broadcasting and for mobile users, particularly those on ships, planes and in remote places. Satellite systems can provide global multimedia services to end users, long before such services would be available through terrestrial networks. In many regions, terrestrial broadband networking is not commercially viable.

This thesis considers resource management on the *uplink* of a multibeam CDMA satellite system, that is, from terminals to satellite. The key control parameters are the beam coverage geometry as well as the transmission rate and power. The beam geometry is subject to control based on the geographic user distribution as discussed in Chapter 3. We assume that the desired beam configurations can be realized via the use of adaptive "smart antennas" [24, 28, 29]. In this chapter we focus on the cell/frame-level access control problem for which the overall objective is to assign transmission powers so as to achieve particular rate and signal-to-interference ratio vectors. We consider that the bandwidth per beam is fixed and so is the CDMA chip rate. Variations in transmission rate are achieved via changes in the processing gain.

As noted previously, CDMA is the access technology of choice for future generation networks. However, CDMA is capacity-limited by interference due to simultaneous transmissions within the same bandwidth. To minimize such interference and enhance system capacity, a variety of techniques are proposed in the literature. Such techniques can be broadly classified as follows:

- Physical layer techniques: These involve code and receiver design so as to facilitate the separation of one signal from another [180–182].
- Network layer techniques: These involve the control of the number and/or rate of simultaneously active terminals or the control of the transmitter power for each terminal or the joint control of the power and rate allocated to active terminals. Section 2.4 presents a review of such techniques.

Our interest in this work is for network layer techniques. Consider a system whereby the various terminals in the system see channels of different and randomly varying capacities. One obvious way in which to counter the effect of deep fading on any particular channel is to increase power at the corresponding transmitter – but in so doing to increase interference on all other channels, prompting all other transmitters to increase power as well. An alternative strategy is in fact to *diminish* rate when the channel weakens, and to compensate with augmented rate when the channel is strong. Such an alternative, exploiting temporal elasticity in the definition of user QoS, may be useful in settings where the time constants associated with user-QoS are long relative to fade durations.

In this chapter, we formulate a rate-allocation strategy along the lines suggested above and quantify its performance as a function of the quality of service time constraints. Our work departs from the literature (for example, [66–69]) in that in our particular instance of the rate management problem, in which user-perceived rate and actual transmission rate are distinguished by different time constants, the long-term requirements are specified in terms of *average-rate*. The scheduling policy exploits time elasticity, as reflected in the length of the window that defines the rate averaging operation, to enhance bandwidth utilization. Such utilization enhancement is achieved by judicious allocation of the transmission rates at each time frame so as to minimize the long-run energy per bit requirements for each terminal while ensuring that the average-rate targets are achieved. Figure 4.1 depicts the access rate allocation for five terminals over an averaging window that is seven timeslots in length. In this case, the rate allocation is restricted to three rate levels.

In evaluating the performance of the access rate scheduling policy we consider the extent to which temporal elasticity yields a throughput advantage. The idea that there should

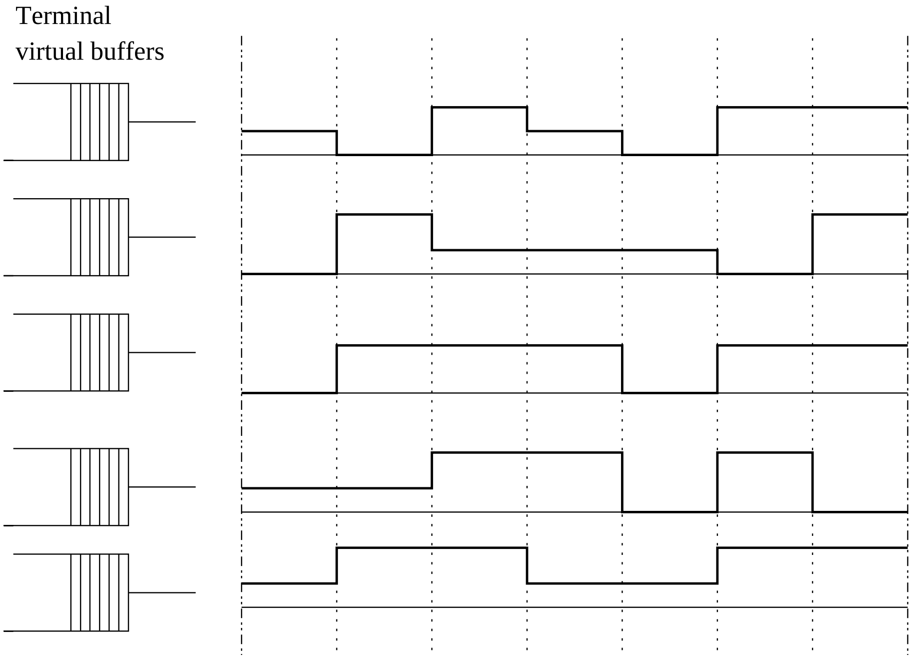


Fig. 4.1 Access rate allocation for five terminals over averaging window

be an advantage is made evident when we compare the throughput due to long-horizon scheduling policies to that due to short-horizon policies, as we shall show in Section 4.3.

Armed with this motivation, we proceed to develop a Markovian model of the access control system so as to bring the optimal rate allocation problem within the purview of Markov Decision Theory and Dynamic Programming (DP). Without loss of generality, we consider the system state at each interval to be the channel conditions and rate allocations up to that interval within the window. At each interval, the control (in form of a rate allocation) selected by the scheduling policy is governed by the current system state and state transition probabilities so as to minimize the long-run network operation costs in terms of failure to achieve the average-rate requirements. To reduce the DP solution complexity, we propose a Markovian approximation to the access control model in which we utilize terminal backlogs to embody the rate history of the system. We conclude the chapter with a performance evaluation of the rate scheduling policy in order to highlight the value of exploiting temporal elasticity to enhance network performance.

In the following section, we outline the access rate scheduling problem.

4.2 Access rate scheduling problem

We consider network operation such that the instantaneous transmission power is subject to an upper bound, and assume that the instantaneous transmission rates are continuously variable and subject to network control. We focus on *uplink* transmissions, from terminals to satellite receiver. At the beginning of each time frame, the access control selected is based on state feedback about previous rate allocations and prevailing channel conditions. We consider a reduced version of the problem, in which user-QoS is specified in terms of average rate alone.

Consider a single satellite beam serving M terminals and with uplink signaling bandwidth, W. Terminal i transmits at instantaneous rate R_i bps and at power P_i limited by the power constraint, $P_i \leq p_i$. The signal from terminal i is received at the satellite receiver with power $\alpha_i P_i$. The attenuation coefficients, α_i, depending on the prevailing channel conditions are assumed known and *iid* from frame to frame. While this is a significant assumption, we consider that the channel conditions can be estimated on a frame-by-frame basis by satellite measurement mechanisms such as described in Section 2.4.1.

Remark. *We acknowledge that the assumption that attenuation coefficient, α, is known is a strong assumption made to simplify the computational model. The assumption is made for simplicity, in the expectation that such simplicity will facilitate exploration of the issues of primary interest.*

Within the satellite beam, channel errors are due to satellite receiver noise and user interference. The former is assumed white and Gaussian, with intensity σ^2 watts/Hz. The latter is typically more complex, but in this thesis — as noted in Section 3.4.1 — we assume that the interference is white and Gaussian. These assumptions allow for use of the SIR formula given in Equation (3.7)

$$\frac{P_i \alpha_i W / R_i}{W\sigma^2 + \sum_{j \neq i}^{M} P_j \alpha_j} = \gamma_i \quad i = 1, \cdots, M.$$

The rate allocation policy decides who will transmit and at what rate, given the delay tolerance, SIR requirements, power constraints and time-varying channel conditions.

We assume that time evolves in consecutive, non-overlapping intervals of constant length. Time interval n is $[T_n, T_{n+1})$. For each interval n, the control parameters are the rate allocation vector $R(n) = (R_1(n), \ldots, R_M(n))$ and the power allocation vector $P(n) = (P_1(n), \ldots, P_M(n))$. The foregoing relationship between powers, rates and SIR thresholds can be expressed in the form

$$\frac{R_i \gamma_i}{W} = g_i(P, \alpha) \quad i = 1, \cdots, M, \tag{4.1}$$

where α denotes the attenuation coefficient vector, W is the system bandwidth, and γ_i is the minimum SIR requirement for terminal i. Where g is invertible, if a given rate allocation is feasible we can calculate $P = h(R, \alpha),\ h(\cdot) \stackrel{\Delta}{=} g^{-1}(\cdot)$. At each interval n, we assume that the attenuation coefficient α_n is independent of α_m, for all $m < n$, and is also independent of all previous rate and power allocations. In addition, we assume that α_n holds constant over the time interval n. We further assume that the probability distribution of α is known.

The access rate scheduling policy has two objectives. The first objective is to achieve prescribed *average-rate* constraints. The constraints in question are formulated in terms of parameters K_i (the length of the rate-averaging window) and ρ_i^* (the rate threshold). For each terminal i, we require

$$\frac{1}{K_i} \sum_{n=N-K_i+1}^{N} R_i(n) \geq \rho_i^* \quad \text{for all } N \tag{4.2}$$

Secondly, we need to achieve the *average-rate* constraints with minimum transmission powers. The objective is to minimize the long-run average energy per bit required per terminal, based on the rate and power allocations made at each interval n, and given by

$$\limsup_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \frac{P_i(n)}{R_i(n)} \tag{4.3}$$

Problem. *Let $\rho_i^{(K)}(N)$ denote the left-hand side of Equation (4.2). For each terminal i, we seek to optimize a combination of the average energy per bit $E_b(i)$ and the function $\mathcal{F}(\rho_i^{(K)}(N), \rho_i^*)$, where $\mathcal{F}(\cdot)$ indicates how well (or not) the rate allocation strategy satisfies* *the average-rate constraints.*

In Section 4.4, we develop a Markovian model of the access control system so as to bring the rate allocation problem within the purview of Markov Decision Theory and Dynamic Programming. Before that, we show the case for time-elastic control by comparing the performance of an infinite-horizon policy to that of a short-horizon policy.

4.3 The case for time-elastic access control

We look at a simple model by which to assess analytically the potential benefits of time-elastic access control. We compare the short-horizon policy, which confers no elasticity, to an infinite horizon one, in which the elasticity is unconstrained. In working out the details, we assume that there are M sources, that the average rate requirement, r, is the same for all sources, and that the α's (as before) are *iid*. In effect, we consider

- Short horizon — $K = 1 : R(n) = r$ (all n)
- Infinite horizon — $K = \infty : \bar{R} = \lim_{K\to\infty} \frac{1}{K} \sum_{i=0}^{K-1} R(n-i) = r$

The rate and power requirements are obtained as follows. We denote the required transmission power as $P(\alpha)$ and the allocated transmission rate as $R(P(\alpha), \alpha)$, and assume a homogeneous SIR requirement, γ.

$K = 1$: For each interval n and each terminal i, $R_i(P(\alpha), \alpha) = r$. That is, the $K = 1$ case represents constant rate transmission. The transmission power can be determined from Equation (3.7)

$$\frac{r\gamma}{W} = \frac{P_i\alpha'_i}{1 + \sum_{j\neq i} P_j\alpha'_j},$$

where, as before, the α' is normalized with respect to the satellite receiver noise, σ^2.

$K = \infty$: In this case,

$$\bar{R} = \int R(P(\alpha), \alpha) dF(\alpha) = r, \qquad \bar{P} = \int P(\alpha) dF(\alpha).$$

Assuming stationary channels and with no constraints on the delay or power, the optimal policy for maximizing system throughput has been shown to be that of scheduling one user at time with priority to users with good channels [47, 151]. In our particular case, where the channel is time-varying and there are constraints on the transmission power, we can assume that due to the infinite length of the averaging window, each terminal will eventually experience good channel conditions. Furthermore, our interest is to achieve average-rate constraints. Consider a rate allocation policy such that, at each interval n, the terminal with the largest α gets to transmit. For a system with M terminals, the probability a terminal will transmit is given by $1/M$. Let α_{max} denote the maximum α at a given interval. If permitted to transmit ($w.p.\ 1/M$), the transmission rate for terminal i is $R_i(P(\alpha), \alpha) = Mr$ and, from Equation (3.7), the transmission power is

$$P_i(\alpha) = \frac{Mr\gamma}{W} \frac{1}{\alpha_i}. \tag{4.4}$$

Otherwise, $P_i(\alpha) = 0$.

In the following subsections, we show the long-run average power requirements for $K = 1$ and for two cases of $K = \infty$ – without enforced idleness (example considered above) and with enforced idleness (transmission permitted only if $\alpha_{max} \geq a$, a given threshold). These average power requirements are given in Equations (4.5), (4.6), and (4.7), respectively. A comparison of the power requirements reveals that the $K = \infty$ policies require less power than the $K = 1$ policy. However, the transmission delay, transmission rate and transmission power for each terminal grow linearly with M in the case of $K = \infty$.

While we do not prove the optimality of the infinite horizon policies, our primary interest here being to highlight the advantages of time-elastic control, we shall use the $K = \infty$ power requirements as benchmarks for the time-elastic control policies we develop later in this thesis.

4.3.1 Long-run average power requirements for $K = 1$

In the case of $K = 1$, all M terminals are required to transmit at their prescribed average rate at each time interval. Assume that all terminals have the same target average rate, r, and SIR constraint, γ. From the SIR formula (Equation (3.7)), we have

$$\frac{1}{\eta_i} = 1 + \frac{W}{r_i \gamma_i}, \quad i = 1, \ldots, M.$$

where

$$\eta_i \triangleq \frac{P_i \alpha_i'}{1 + \sum_{j}^{M} P_j \alpha_j'}.$$

It can be shown that a feasible power allocation should satisfy $\sum_{j=1}^{M} \eta_j \leq 1$.

Suppose terminal $i = 1$ is taken as the reference terminal, and that it has a normalized attenuation coefficient, α_1'. If $P_1 = p$, then a feasible power allocation can be obtained from

$$\left(\frac{1 + p\alpha_1'}{p\alpha_1'}\right) \eta_1 + \eta_2 + \ldots + \eta_M = 1.$$

Given that the terminals have the same rate and SIR constraints, we can rewrite the power feasibility condition as

$$\left[\left(\frac{1 + p\alpha_1'}{p\alpha_1'}\right) + M - 1\right] \eta_1 = 1,$$

from which we obtain

$$p = \frac{1}{\alpha_1}\left(\frac{\delta}{1 - M\delta}\right),$$

where $\delta = \left(1 + \frac{W}{\gamma r}\right)^{-1}$.

At each interval n, the total transmission power is $P(\alpha) = Mp$ because all terminals are transmitting. Let F_1 denote the marginal distribution of each α — given that the α

are *iid*. The long-run average power is given by

$$\begin{aligned}\bar{P} &= \frac{M\delta}{1-M\delta}\int \frac{1}{\alpha_1} dF(\alpha) \\ &= \frac{M\delta}{1-M\delta}\int \frac{1}{\alpha_1} F_1^{M-1}(\alpha) dF_1(\alpha) \qquad (4.5) \\ &= \frac{M\delta}{1-M\delta}\mathbb{E}\left[\frac{1}{\alpha_1}\right]\end{aligned}$$

In the sequel we compute the average power requirements for $K = \infty$ for comparison with Equation (4.5).

4.3.2 Long-run average power requirements for $K = \infty$ (without enforced idleness)

For $K = \infty$ without enforced idleness, at least one user gets to transmit at each time interval. In the case we consider here, we assume that at each interval, the user i with the largest α gets to transmit at power $P_i = Mr\gamma / W\alpha'_i$ (Equation (4.4)). Let the largest α' at each interval be denoted by α_{max}. Then the long-run average power is given by

$$\begin{aligned}\bar{P}_i &= \int P_i(\alpha) dF(\alpha) \\ &= \int I_i \cdot \frac{Mr\gamma}{W}\frac{1}{\alpha'_i} dF(\alpha) \\ &= \frac{Mr\gamma}{W}\int \frac{I_i}{\alpha'_i} dF(\alpha)\end{aligned}$$

where I_i is an indicator function such that $I_i = 1$ whenever $\alpha'_i(n) = \alpha_{max}$.

Focusing on the integral term, we have that

$$\begin{aligned}
\int \frac{I_i}{\alpha_i'} dF(\alpha) &\equiv \mathbb{E}\left[\frac{1}{\alpha_i'}|I_i = 1\right] Pr[I_i = 1] \\
&= \mathbb{E}\left[\frac{1}{\alpha_i'}|\alpha_i = \alpha_{max}\right] \\
&= \int \frac{1}{\alpha_i'} dF_{\alpha|\alpha_i'=\alpha_{max}}(\alpha|\alpha_i' = \alpha_{max}) \\
&= \int \frac{1}{\alpha_i'} \frac{dF_{\alpha,\alpha_i'=\alpha_{max}}(\alpha, \alpha_i' = \alpha_{max})}{dF_{\alpha_i'=\alpha_{max}}(\alpha_i' = \alpha_{max})}.
\end{aligned}$$

Given that α is *iid*, let F_1 denote the marginal distribution for each α. We obtain the long-run average power for $K = \infty$ as

$$\begin{aligned}
\bar{P}_i &= \frac{Mr\gamma}{W} \cdot M \int \frac{1}{\alpha_i'} F_1^{M-1}(\alpha) dF_1(\alpha) \\
&= \frac{Mr\gamma}{W} \mathbb{E}\left[\frac{1}{\alpha_{max}}\right] \qquad (4.6) \\
&\equiv \frac{M\delta}{1-\delta} \mathbb{E}\left[\frac{1}{\alpha_{max}}\right],
\end{aligned}$$

where, as before, $\delta = \left(1 + \frac{W}{\gamma r}\right)^{-1}$.

A comparison of long-run average power requirements as given in Equation (4.5) and Equation (4.6) reveals the following:

1. For all M and δ, the case of $K = 1$ has the larger coefficient for the expectation operator.

2. For any arbitrary distribution, $\mathbb{E}\left[\frac{1}{\alpha_1}\right] \geq \mathbb{E}\left[\frac{1}{\alpha_{max}}\right]$.

Hence, a larger average power is required for the case of $K = 1$. However, while this comparison highlights the case for *infinite-horizon* considerations in terms of power required to meet given rate objectives, we should also note the following in the case of $K = \infty$. The transmission delay, burst rate and transmission power for each terminal grow linearly with M. With increasing M, such a system would no longer be sustainable in a case where, say, constraints are placed on the maximum transmission power per terminal.

4.3.3 Long-run average power requirements for $K = \infty$ (with enforced idleness)

In the strategy for $K = \infty$ above, we have considered a work-conserving policy in which there is a transmission scheduled in each and every frame. Call this the MAX-α policy. In this section, we show that the optimal policy when $K = \infty$ is in fact achieved when non work-conserving policies that involve forced-idleness are considered.

The basic model for the optimal policy is the same as for the MAX-α policy: time evolves in frames; α's are constant over individual frames and *iid* from frame to frame; sources are homogeneous in the sense of sharing a common distribution for α and common objectives for SIR and long-run average rate. No doubt the model could be generalized without sinking the whole enterprise, but then the conclusion, no longer valid as stated, would have to be reformulated. In any case, it is important to move the frame structure outside the scope of design and optimization — to foreclose the otherwise attractive possibility of making the frames arbitrarily short. So we make the assumption that the frame structure is fixed from the outside: α's are constant within frames and *iid* between frames.

Consider the following non work-conserving policy. Let a denote the largest possible value of α, and let transmission for user i be permitted only when $\alpha_i = \alpha_{max} = a$. We can show that enforcing idleness in the access control strategy has the effect of replacing the factor $\mathbf{E}[1/\alpha_{max}]$, in Equation 4.6, by $1/a$. In this case the long-run average power with enforced idleness is given by

$$\bar{P}_i = \frac{\gamma\sigma^2}{a} r. \tag{4.7}$$

Hence, long-run average transmission power is minimized by the access control strategy of $K = \infty$ with enforced idleness.

In the rest of this sub-section, we show the validity of Equation (4.7). We begin by showing that simultaneous transmission is sub-optimal. Then, we proceed to show the effect of enforced idleness and provide the proof of Equation (4.7).

A. Concurrency is sub-optimal

Without constraints on time and/or power, and subject to the assumption that interference is appropriately modeled by the SIR model (Equation (3.7)), the simultaneous transmission of users is sub-optimal — note that the assumption made here excludes receiver techniques such as multiuser detection (MUD) [176, 183]. Indeed, in addition to the preceding section on the MAX-α policy, a variety of works show that single-user transmission per timeslot will maximize system capacity, see [47] and [151] (and references therein).

A.1 Fact. *Any strategy in which two or more terminals transmit simultaneously can be improved by separating those transmissions in time. The improvement is strong: each terminal sees unchanged average rate at reduced average power. The cost of such improvement is in terms of peak rate and peak power, both of which are substantially increased.*

Proof. Consider two systems, A and B, operating over a time frame of length T seconds. Each system supports M terminals. In System (A) all M terminals transmit together, at rates $R_1^A, \ldots, R_M^A$. System (B) separates the transmissions in TDM fashion: the T-second frame is partitioned into M (disjoint) slots of lengths $T_1, \ldots, T_M$, where $T_1 + \cdots + T_M = T$; terminal (i) transmits at rate R_i^B in slot (i) and otherwise is silent. In both systems the transmit powers are selected so that the M SIR constraints are satisfied with equality. Assuming that the average rate per terminal per frame is the same in both systems, we compare the corresponding average powers.

We begin by making the following definitions:

$$\epsilon_i \triangleq \frac{T_i}{T}, \qquad \delta_i \triangleq \left(1 + \frac{W}{\gamma_i R_i^A}\right)^{-1}, \qquad \Delta \triangleq \sum_i \delta_i,$$

where γ_i stands for the SIR target associated with terminal (i). The two systems are related by

$$\epsilon_i R_i^B = R_i^A, \; i = 1, \ldots, M.$$

The average and instantaneous powers in System (A) are the same: $\bar{P}_i^A = P_i^A$. In System (B) they are related by $\bar{P}_i^B = \epsilon_i P_i^B$, where P_i^B is the power expended during actual transmission. Power and rate are in turn related in the usual way, through the SIR constraints

$$\frac{\gamma_i R_i^A}{W} = \frac{\alpha_i P_i^A}{W\sigma^2 + \sum_{j \neq i} \alpha_j P_j^A}, \qquad \frac{\gamma_i R_i^B}{W} = \frac{\alpha_i P_i^B}{W\sigma^2}.$$

From these you get that

$$\bar{P}_i^A = \frac{W\sigma^2}{\alpha_i} \frac{\delta_i}{1 - \Delta}, \qquad \bar{P}_i^B = \frac{W\sigma^2}{\alpha_i} \frac{\delta_i}{1 - \delta_i}.$$

In particular,

$$\frac{\bar{P}_i^A}{\bar{P}_i^B} = \frac{1 - \delta_i}{1 - \Delta} > 1. \tag{4.8}$$

□

B. The effect of enforced idleness:

Assume here that all sources have the same SIR threshold γ. The tradeoff under strategy MAX-α between long-run average power and long-run average rate is given by

$$\bar{P} = \gamma\sigma^2 \mathbf{E}\left[\frac{1}{\alpha^*}\right] \bar{R},$$

where α^* (depending on M) denotes the random variable $\max_{i \leq M} \alpha_i$.

Consider a to be the largest possible value of α. We can define

$$a \stackrel{\Delta}{=} \min\{u : F_\alpha(U) = 1\}.$$

Inasmuch as F_α is a cumulative distribution function (*cdf*) for α, and therefore right-continuous by construction, the use of min, as opposed to inf, is well-founded. In particular, because the quantity a is the largest possible value of α, we have that $F_\alpha(u) < 1$ for $u < a$.

B.1 Fact. *The very best power/rate tradeoff of which MAX-α is capable, and which indeed is achieved when M is large, is simply*

$$\bar{P} = \frac{\gamma\sigma^2}{a}\bar{R}. \tag{4.9}$$

Proof. From Equation (4.6), for reference terminal i, we have that

$$\mathbf{E}\left[\frac{1}{\alpha^*}\right] = M\int \frac{1}{\alpha_i}F^{M-1}(\alpha)dF(\alpha)$$

As $M \to \infty$, we have that

$$F_\alpha^{M-1}(u) \to \begin{cases} 1, & u = a \\ 0, & u < a \end{cases}$$

Thus, when M is large,

$$\mathbf{E}\left[\frac{1}{\alpha^*}\right] \longrightarrow \frac{1}{a}$$

□

Claim. *A tradeoff identical to Equation (4.9) is (almost) achieved for arbitrary M by modifying MAX-α so as to allow for forced idleness.*

Observe, in light of (A.1), that the inquiry can be conducted ignoring all sources save one. Focus on Source (1). Write $I(n) = 1$ if Source (1) is permitted to transmit in frame (n), and $I(n) = 0$ otherwise. Let $R(n)$, $P(n)$ respectively denote the Source (1) rate and power in frame (n). Then

$$\frac{1}{N}\sum_1^N R(n) = \frac{\sum_1^N I(n)}{N}\cdot\frac{\sum_1^N I(n)R(n)}{\sum_1^N I(n)}, \qquad \frac{1}{N}\sum_1^N P(n) = \frac{\sum_1^N I(n)}{N}\cdot\frac{\sum_1^N I(n)P(n)}{\sum_1^N I(n)}.$$

Assuming that the limits exist in the limit of large N, we have that

$$\bar{R} = \nu\bar{R}', \quad \bar{P} = \nu\bar{P}',$$

where ν is the long-run proportion of frames that are allocated to Source (1), *unprimed* overbar denotes long-run average over *all* frames, and *primed* overbar denotes long-run average over Source (1) frames.

B.2 Punch line. *If $\bar{P}'$, $\bar{R}'$ are linearly related, then so are $\bar{P}$, $\bar{R}$, and the relationships in both cases are identical.*

In other words, the optimal location of forced idle periods can be derived from a *single-source* system model. The resulting $\bar{P}'$, $\bar{R}'$ tradeoff can be transferred without change to a multi-source environment subject to a, the maximum value of α.

B.3 Fact. *For every M, and for every $\epsilon > 0$, there is a MAX-α policy with forced idleness for which*

$$\bar{P} \leq (1+\epsilon)\frac{\gamma\sigma^2}{a}\bar{R}.$$

Motivated by (B.2), here the underlying model has just one source which owns all frames. This being so, for convenience we shall drop the primes. Note that frame indices are not conserved in moving from the multi-source to single-source model. The goal is to minimize $\bar{P}$ for given $\bar{R}$. Moreover, there being only one source, we have that

$$\gamma\sigma^2 R(n) = \alpha(n)P(n),$$

the argument n indicating that the associated variables refer to frame (n). Hence, the problem can be reduced as follows.

Problem. *Choose $P(\cdot)$ so as to minimize $\bar{P}$ subject to $\overline{\alpha P} = c$, where $c \triangleq \gamma\sigma^2\bar{R}$ is a given constant — $P(n)$ being a function of $\alpha(n)$.* **Equivalently:** *Given a random variable α, find a function $P = g(\alpha)$ which minimizes $\mathbf{E}[P]$ while satisfying $\mathbf{E}[\alpha P] = c$, where $\mathbf{E}[\cdot]$ denotes expectation relative to the distribution of α.*

Typically there is no such $g(\cdot)$, as will be clear shortly. Consequently, the set of possible controls is not closed. Nevertheless, $\mathbf{E}[P]$ has a well-defined infimum which is arbitrarily closely approachable.

B.3.1 Fact. $\mathbf{E}[P] \geq a^{-1}\mathbf{E}[\alpha P] = c/a$. *The inequality is strict unless $(a-\alpha)P = 0$ with probability 1 (w.p. 1) — the probability distribution being that of α. In other words, the* *inequality is strict unless transmission is restricted essentially to those frames in which $\alpha = a$.*

B.3.2 Fact. $\inf_g \mathbf{E}[P] = \dfrac{c}{a}$.

Proof. Choose $\epsilon > 0$. Define

$$q \triangleq 1 - F_\alpha(a - \epsilon), \qquad Q \triangleq \mathbf{E}[\alpha \,|\, \alpha > a - \epsilon]$$

and

$$g(u) \triangleq \begin{cases} \dfrac{c}{qQ} & \text{if } u > a - \epsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that $q > 0$, by definition of a. Set $P \triangleq g(\alpha)$. Then $\mathbf{E}[\alpha P] = c$, as required, and

$$\mathbf{E}[P] = \frac{c}{Q} \le \frac{c}{a - \epsilon} = \frac{\gamma\sigma^2 \bar{R}}{a - \epsilon}.$$

Since ϵ can be arbitrarily small, it follows that

$$\inf_g \mathbf{E}[P] \le \frac{c}{a} = \frac{\gamma\sigma^2}{a}\bar{R}. \tag{4.10}$$

Combining Equation (4.10) with (B.3.1) concludes the proof of both (B.3.2) and (B.3). □

C.1 Conclusion. *Allowing forced idleness in the access control strategy has the effect of replacing the factor* $\mathbf{E}[1/\alpha^*]$*, where it appears in the slope of the* $\bar{P}$ – $\bar{R}$ *graph, by* $1/a$*.*

In summary, analysis of the power requirements for the cases of $K = 1$ and $K = \infty$ supports the case for time-elastic access rate scheduling. As previously noted, we shall use the $K = \infty$ power requirements as benchmarks for the access control policies we develop in this thesis. In particular, we shall consider the case of enforced idleness and make use of Equation (4.10). In the following section, we develop a Markovian model of the access control system.

4.4 Markovian model of time-elastic access control

The access control we seek decides who will transmit and at what rate, given the delay tolerance, SIR requirements, power constraints and time-varying channel conditions. As discussed in Section 4.2, the rate allocation objectives are to

(a) achieve prescribed *average-rate* constraints such that for each terminal i we have

$$\frac{1}{K_i} \sum_{n=N-K_i+1}^{N} R_i(n) \geq \rho_i^* \quad \text{for all } N \tag{4.11}$$

(b) minimize the long-run average energy per bit required per terminal, based on the rate and power allocations made at each interval n, and given by

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \frac{P_i(n)}{R_i(n)} \tag{4.12}$$

As before, let $\rho_i^{(K)}(N)$ denote the left-hand side of Equation (4.11). As noted previously, for each terminal i, we seek to optimize a combination of the average energy per bit $E_b(i)$ and the function $\mathcal{F}(\rho_i^{(K)}(N), \rho_i^*)$, where $\mathcal{F}(\cdot)$ indicates how well (or not) the rate allocation strategy satisfies the average-rate constraints. We develop a Markovian model of the access control in order to apply techniques from Markov Decision Theory and Dynamic Programming.

Definition. *A Markov Decision Process (MDP) is a controllable Markov process with associated costs or rewards. We are interested in the discrete-time variant. At each stage (time instant), a control action is selected, a next state is selected according to transition probabilities that depend on the control, and a reward is issued (or a cost incurred) again depending on the control [184, 185].*

From the rate allocation objectives in Equation (4.11) and Equation (4.12), and the SIR model in Equation (3.7), at each interval n, the rate allocation $R(n)$ depends on the prevailing channel conditions (as described by the attenuation factor) and past rate allocations; that is, $R(n) \triangleq f(R(n-1), \ldots, R(n-K+1), \alpha(n))$. We assume that the rate allocation process is stationary and non-anticipatory (that is, independent of future channel conditions). Without loss of generality, we can consider the system state, $x(n)$, at interval n to be the channel conditions and rate allocations up to that interval. Assuming the channel attenuation coefficients to be *iid* random variables of known distribution, we shall assert the state variable to be $x(n) = [R(n), \ldots, R(n-K+1)]$ — only a function of the rate allocation. For an M-terminal system, this results in an MK dimensional state space, X.

In the following subsections we describe the system control, and discuss the dynamic programming formulation of the access control problem (Section 4.4.2). Based on this formulation, we discuss the computation of an optimal access control policy (Section 4.4.3). Due to the complexity of this computation (as we shall show), in subsequent sections, we proceed to define Markovian approximations to the access control model which we shall use in this thesis to obtain an optimal access rate scheduling policy.

4.4.1 Describing the access system control

At interval n, the system control $u(n)$ determines a rate allocation $R(n)$ from a set of possible rate allocations, $A(n)$. The traffic model used in this work assumes that each terminal is capable of transmitting at whatever rate it is allocated and that the rates vary continuously.

The system control is governed by two conditions:

- The prevailing channel conditions. At each interval n, the conditions are described by the attenuation coefficient, $\alpha(n)$, and are independent of the state of the system. Again, we assume the governing probability density function for the attenuation coefficient is known. Further, we assume the system is non-anticipative in that future values of the attenuation coefficient, $\alpha(m)$, $m > n$, remain unknown and statistically independent of the past and the present.

- The nature of the past rate allocations. This can be determined, for example, by the value of an indicator function or a rate history function (RHF) that relates the left-hand side of Equation (4.11), $\rho_i^{(K)}(N)$, to the prescribed average-rate, ρ_i^*. For instance, RHF could indicate whether the average of the rates in a given state is greater or less than the prescribed average-rate constraint, ρ_i^*, for each terminal i.

Under these conditions note that, at each interval, the randomness exhibited by the control is due to the randomness of the channel conditions, α. The evolution of the state process can be described by a function ϕ of the current state, the prevailing channel conditions and the applied control:

$$x(n+1) = \phi(x(n), u(n)),\ \ n = 1, 2, \ldots . \tag{4.13}$$

The control $u(n)$ in interval n is specified by a policy, $\pi : X \times \alpha \mapsto A$, where A is the set of possible rate allocations such that $R_i(n) \triangleq u_i(x(n), \alpha(n))$ for each terminal i. We assume that the policy π is independent of time: that is, π is a stationary policy.

4.4.2 Dynamic programming formulation of access control problem

At interval n, let the control described above be denoted $u(n) \triangleq \pi(x(n), \alpha(n))$. The rate allocation problem is then to optimize the long-run average of a function $\psi(x(n), u(n))$ that represents a combination of $E_b(i)$ and $\mathcal{F}(\rho_i^{(K)}(N), \rho_i^*)$ for all i. This optimization is subject to the constraint $0 \le P_i(T_n) \le p_i$, $i = 1 \ldots M$, where p_i is the maximum transmission power for user i. The average is taken over the distribution of the attenuation coefficient, α. We can write this optimization problem as

$$\begin{aligned} \text{optimize} \qquad & \limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{E}_\alpha[\psi(x(n), u(n))] \\ \text{subject to} \qquad & 0 \le P_i(T_n) \le p_i, \quad i = 1 \cdots M. \end{aligned} \tag{4.14}$$

Let us consider the problem above as follows. As a result of the chosen rate action at each interval n, let the system incur a one-stage cost of $c(x(n), u(n))$ equivalent to $\psi(x(n), u(n))$. The goal is then to optimize the *conditional average cost* incurred by the system which is defined by $\bar{c}(x(n), u(n)) \triangleq \mathbf{E}_\alpha[\psi(x(n), u(n))|x(n) = x]$, for all possible values $x \in X$. Equation (4.14) then defines an average cost Markov Decision Problem (MDP), and can be solved using dynamic programming techniques [166, 186]. For the

MDP, we seek an admissible control policy π that will minimize the long-run average of function $\psi(\cdot)$ — via minimization of the long-run average cost for the system, $\bar{c}(x(n), u(n))$. A fundamental theorem of dynamic programming is as follows.

Theorem. If *the following conditions are true:*

- *the one-stage cost is a function of* x
- *the objective is to minimize the long-run average cost per stage*
- x *is Markov when control* u *is a function of* x

then u *can be chosen to be a function of only* x *without loss of optimality.*

In addition to the cost structure $\bar{c}(x(n), u(n))$ defined above, we need the transition probabilities from one state to the next. Let $J_\pi(i),\ i \in X$, the state space, denote the value function that estimates the long-run average cost — given policy π and an initial state $x(1) = i$. Then

$$J_\pi(i) = \lim_{N\to\infty} \frac{1}{N} \mathbf{E}_\pi \left\{ \sum_{n=1}^{N} \bar{c}(x(n), u(n)) | x(1) = i \right\} \ i \in X. \tag{4.15}$$

Given the set of all admissible policies, Π, the optimal control policy π^* is that which minimizes the average operating cost and is obtained as

$$J^*_\pi(x) = \inf_{\pi\in\Pi} J_\pi(x). \tag{4.16}$$

Let V denote the optimal value function, J^*_π . V satisfies the *Optimality Equation* [166, pp. 31], [185]

$$\begin{aligned} V(i) &= \min_u \left[\bar{c}(x(n), u(n)) + \sum_j p_{ij}(u) V(j) \right], \quad \forall\ i \in X \\ &\equiv \min_u \left[\bar{c}(x(n), u(n)) + \mathbf{E}\left[V(j) | x = i \right] \right], \quad \forall\ i \in X, \end{aligned}$$

where $p_{ij}(u)$ is the transition probability from state i to state j under control u:

$$p_{ij}(u) = P[x(n+1) = j | x(n) = i,\ u(n) = u,\ \alpha(n) = \alpha),\ \forall i, j \in X. \tag{4.17}$$

For a controlled Markov process, the transition probabilities are determined by the actions due to the rate-dependent control function u. Let the rate history at state n be denoted by

$$H(n) \triangleq [R(n-K+1), R(n-K+2), \ldots, R(n-1)].$$

Then, dependent on the control function u, the transition probability from state $x(n) = (R(n), H(n))$ to $x(n+1)$ is given by a function ϑ of the control applied, the rate history and the prevailing channel conditions

$$p_{ij}(u) \triangleq P[(R(n+1) = r', H(n+1) = h') | (R(n) = r, H(n) = h)] = \vartheta(r, h, \alpha). \tag{4.18}$$

Observe that the randomness in $x(n+1)$ is due to r', which in turn is random as a result of the randomness of the channel conditions. In conclusion, the transition probabilities are governed by the probability distribution of the attenuation coefficient, α.

With the cost structure and the transition probabilities defined, we are now in position to solve for the optimal policy that will minimize the long-run average cost of the system. As we anticipated, the computational complexity involved is significant.

4.4.3 Computation of time-elastic access control policy

In the case of finite or sufficiently quantized control u, there are two conditions that result in the existence of a stationary optimal policy. Let i, j denote two states. The conditions are that there exist a bounded function $h(i), i \geq 0$ such that $|h(i)| \leq Z,\ Z > 0$, and a constant λ such that

$$\lambda + h(i) = \min_{u} \left[\bar{c}(i, u) + \sum_{j=0}^{\infty} p_{ij}(u) h(j) \right],\ i \geq 0 \tag{4.19}$$

where $\bar{c}(i, u)$ is the one stage average cost for state i and control u [166, pp. 93], [186].

Then the optimal stationary policy, π^*, is obtained as

$$J_\pi^*(i) = \lambda, \quad \forall i \geq 0 \tag{4.20}$$

and for each i, π^* is any policy with controls that minimize the right-hand side of Equation (4.19).

An optimal solution to the average cost problem of Equation (4.19) can also be obtained as a special case of the solution to the ν-discounted cost problem, where $\nu \in (0, 1)$ [166]. This is done as follows. Assume a focal state, say state 1. Let J_ν denote the optimal expected ν-discounted cost function

$$J_\nu(i) = \inf_u \left[\bar{c}(i, u) + \nu \sum_j p_{ij}(u) J_\nu(j) \right], \ \forall\ i \in X. \tag{4.21}$$

Condition. *If there exists $Z < \infty$ such that $|J_\nu(i) - J_\nu(1)| < Z$ for all $\nu > 0$ and i, then*

1. *there exists a bounded function $h(i)$ and a constant λ that satisfy Equation (4.19);*
2. *for some sequence $\nu_n \to 1,\ h(i) = \lim_{n\to\infty} [J_{\nu_n}(i) - J_{\nu_n}(1)]$;*
3. $\lim_{\nu \to 1} (1 - \nu) J_\nu(1) = \lambda$.

Proof on [166, pp. 93].

The condition above satisfies the existence of a stationary optimal policy whereby the optimal cost per stage is independent of the initial stage. In this case, λ is the minimal average cost. Let u^* be any policy with controls that minimize the right-hand side of Equation (4.19) for all $i \in X$. Applying $J_u^*(i) = \lambda$ for all $i \in X$ (as in Equation (4.20)), we have the following approximation:

$$J_u^*(i) \approx (1 - \nu)\ \inf_u \left[\sum_{n=0}^{\infty} \nu^n \mathbf{E}_\alpha \left[\bar{c}(i, u) | x_0 = i \right] \right], \ \forall i \in X. \tag{4.22}$$

Hence, one computational approach to solving the optimal control problem is *value iteration* (also known as Successive Approximation) which is in essence the dynamic programming algorithm. Let $J_0(i)$ be any arbitrary bounded function and define $J_1(i)$ by:

$$J_1(i) = \inf_u \left[\bar{c}(i,u) + \nu \sum_j p_{ij} J_0(j) \right], \ \forall \ i \in X.$$

Then for iteration index, $n > 1$,

$$J_n(i) = \inf_u \left[\bar{c}(i,u) + \nu \sum_j p_{ij} J_{n-1}(j) \right], \ \forall \ i \in X. \tag{4.23}$$

J_n is the minimum expected discounted cost of an n-stage problem that confers a terminal cost $J_0(j)$ if the process ends in state j. J_n converges uniformly to J as $n \to \infty$, that is, $\inf_i \| J_n(i) - J(i) \| \to 0$. Finally, as $n \to \infty$, the optimal access control policy is obtained as that which approaches the long-run average cost J.

As can be observed from Equation (4.23), J_n is optimized over all possible control u for each of the MK states and for each value of α. Furthermore, the value function for each state and control is obtained by solving a system of simultaneous linear equations. Solution of this problem is computationally intensive with increasing dimensionality of the MK, control and/or α spaces. It is such complexity that leads us to consider Markovian approximations to the time-elastic access control model.

4.5 Markovian approximations to the time-elastic access control model

With solution complexity as indicated above, direct calculation for an optimal rate allocation policy satisfying Equation (4.19) is computationally intensive. Instead we consider approximations to the time-elastic access control problem.

We begin by reducing the dimensionality of the state space, MK. We suppose, temporarily, that there are only $M = 2$ users, so that the state space is small and the set of rate and power vectors consistent with the given SIR constraints has an analytical description. In such a case, it is possible to construct an optimal strategy via dynamic programming as we shall show. System performance will be measured by the (two-dimensional) set of feasible *average-rate* vectors. Figure 4.2 depicts an access rate allocation for two terminals with rate averaging window of size K.

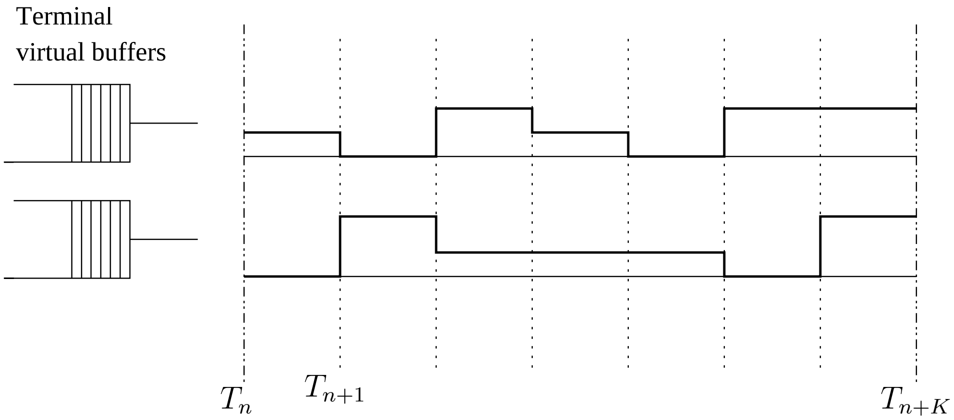


Fig. 4.2 Access rate allocation for two terminals over a window of size K

We further consider two additional options that will simplify the control policy computation. First, we suppose that instead of being allowed to vary continuously, the allocated rate can be quantized to a specific number of values. In this case, for a simple 2-user network, the computational complexity remains a function of only the size of the rate averaging window which represents the state space per terminal. We discuss this option in Section 4.5.1, where we also present the cost structure and transition probabilities associated with the model.

The second option, which is our primary focus for the Markovian approximation to the access control model, is to use terminal backlogs as the state variable instead of the rate history. In so doing, we reduce the state space per terminal from K (the window size) to a single state variable (the terminal backlog). We discuss the backlogs approximate model in Section 4.6.

4.5.1 Quantized-rate approximate model

Assume that the rate allocations are quantized to some finite number of values, instead of allowed to vary continuously. This reduces the dimensionality of the control space and, in turn, the computational complexity. In the general M case, the quantized-rate model does not alter the state space of the original Markov model, it only alters the control space. As before, at each interval n, the state variable for each terminal, $m = 1, \ldots, M$, is denoted by $x(n) = [R(n), \ldots, R(n-K+1)]$.

Consider a system that can support only three rates (in decreasing order of magnitude): $R_{max}, R_{min}, R_{sync}$. At each interval n, the control $u(n)$ determines the rate allocation $R(n)$ in a manner that is dependent on the past K allocations as well as the prevailing channel conditions. However, in this approximate model, $u(n) \stackrel{\Delta}{=} \mathcal{R}(x, \alpha) \in \{R_{max}, R_{min}, R_{sync}\}$. As before, we assume that the attenuation coefficients, α, are *iid*.

The evolution of $x(\cdot)$ is described as follows. Let $x(n+1) = \mathbf{y}$ and $x(n) = \mathbf{x}$, where $\mathbf{y} = y_0, y_1, \ldots, y_{K-1}$ and $\mathbf{x} = x_0, x_1, \ldots, x_{K-1}$, then

$$y_j = \begin{cases} x_{i-1}, & i = 1, \ldots, K-1, \quad j = i \\ u(\mathbf{x}, \alpha), & j = 0. \end{cases}$$

Note that, again, the randomness in the evolution of $x(\cdot)$ is due exclusively to the randomness of $R(n)$ — itself a random variable as a result of the randomness in the α-process.

Thus, given $x(n) = \mathbf{x}$ and rate allocation policy u, $x(n+1) = \mathbf{y}$ if and only if for each m,

$$u_m(\mathbf{x}, \alpha) = y_0, \quad y_0 \in \{R_{max}, R_{min}, R_{sync}\}. \tag{4.24}$$

Equation (4.24) provides the basis for computing the transition probabilities in $x(\cdot)$ associated with a particular policy $u(\cdot)$. To complete the problem formulation for the quantized-rate approximate model, we proceed to specify the cost structure and the transition probabilities.

Cost structure for the quantized-rate approximate model

At each interval n, we require that the cost $c(x(n), u(n))$ for any strategy u be a *scalar* function of the rate allocation, $x(n)$, the channel parameters, $\alpha(n)$, and the control u. The system operating cost is the sum of the costs of achieving the prescribed average rate and of the energy-related costs. We assume that the energy-related costs are a non-decreasing function of $P_m(n)/R_m(n)$, for each terminal m, where $P_m(n)$ is the transmission power. The costs of achieving the prescribed average rate are obtained as follows. As in the original problem, we require that for each terminal m:

$$\frac{1}{K}\sum_{n=N-K+1}^{N} R_m(n) \geq \rho_m^*$$

Again, let $\rho_m^{(K)}(N)$ denote the left-hand side, and let $\mathcal{F}(\rho_m^{(K)}(N), \rho_m^*)$ denote a function that relates the two sides of the equation. For example, we could have $\mathcal{F}(\rho_m^{(K)}(N), \rho_m^*)$ denotes an indicator function that tracks if the prescribed average-rate is not attained, such that

$$\mathcal{F}(\rho_m^{(K)}(N), \rho_m^*) = \begin{cases} 1, & \rho_m^{(K)}(N) < \rho_m^* \\ 0, & o.w. \end{cases}$$

Alternatively, we could have that $\mathcal{F}(\rho_m^{(K)}(N), \rho_m^*)$ is a non-decreasing function of the difference between the prescribed average-rate and the achieved average-rate, such that

$$\mathcal{F}(\rho_m^{(K)}(N), \rho_m^*) = \begin{cases} f(\rho_m^* - \rho_m^{(K)}(N)), & \rho_m^{(K)}(N) < \rho_m^* \\ 0, & o.w. \end{cases}$$

Then we can consider a cost structure as follows:

$$c_u(n) = \sum_m c_{u,m}(n) = \sum_m \mathcal{F}_n(\rho_m^{(K)}(N), \rho_m^*).$$

In the context of the optimization of the rate allocation policy u, the significant attribute of $c_u(\cdot)$ is the *conditional average* for all possible values of $\mathbf{x}$,

$$\begin{aligned} \bar{c}_u(x) &\triangleq \mathbb{E}_\alpha[c_u(n)|x(n) = \mathbf{x}] \\ &= \int c_{u,\mathbf{x}}(n) f_\alpha(\alpha) d\alpha \end{aligned} \tag{4.25}$$

Next, we turn our attention to the transition probabilities.

Transition probabilities for the quantized-rate approximate model

The transition probabilities, $p_{\mathbf{xy}}(u)$, from state $\mathbf{x}$ to state $\mathbf{y}$ under control u, are developed from the evolution of $x(\cdot)$ as given in Equation (4.24). The randomness in the transition from $\mathbf{x}$ to $\mathbf{y}$ is a consequence of the randomness of the attenuation coefficient, α, and so we have that:

$$\begin{aligned} p_{\mathbf{xy}}(u) &\triangleq \int Pr\{\mathbf{y} = y_0, y_1, \dots |\mathbf{x} = x_0, x_1, \dots, u, \alpha\} f_\alpha(\alpha) d\alpha \\ &= \int p_u(\mathbf{x}, \mathbf{y}|\alpha) dF(\alpha) \qquad (4.26) \\ &\equiv Pr\{\alpha : u(\mathbf{x}, \alpha) = y_0\}. \end{aligned}$$

We can now proceed to solve the optimal rate allocation for this approximate model using DP techniques such as value iteration. However, we focus on the backlogs approximate model discussed in the following section.

4.6 Backlogs approximate model

The system state as defined above is a segment of the rate history and described by MK-dimensional vectors. The approximation to be developed here leads to an alternative state definition requiring only M scalar integers. The idea is to attach a fictitious buffer to each terminal and to use the backlog in that buffer as an approximation to the rate history of that terminal.

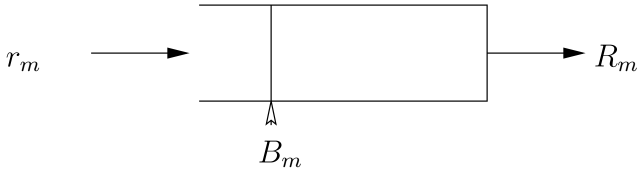


Fig. 4.3 Schematic diagram of a buffer

Important. *The backlog model is proposed here as a* computational device *by which reasonable controls may be derived for the original rate allocation problem. We will show* *among other things (Section 4.6.1) that the appropriate input rate to the fictitious buffers is exactly the target average rate that appears in the quality of service specification. The input rate to the buffers is thus* constant *even though the actual source transmission rates in the original model are not.*

As depicted in Figure 4.3, the m-th buffer is characterized by its length B_m and its fixed, constant input rate, r_m, that is equivalent to the target average-rate requirement. How these values are selected we take up later by referring to the original problem. The associated depletion rate R_m is time varying and state dependent. The vector $R \triangleq (R_1, \cdots, R_M)$ is the control we wish to optimize. In systems where the vectors R and P are related by the SIR constraint in closed form fashion, such as the SIR model (Equation (3.7)), P is an equivalent control variable.

As before, time is assumed to evolve in consecutive, non-overlapping frames of constant length. The time scale is chosen so that the frame length is unity; the integer-valued time instants accordingly coincide with frame boundaries. The depletion rates R_m, and the parameters of the M channels between terminals and satellite, are assumed constant over the course of a single frame.

The dynamics of the system are described by a number of discrete-time random processes:

1. *State* processes $\{x_m(n)\}_n$, $m = 1, \ldots, M$, where $x_m(n)$ stands for the backlog in buffer m at time n (the start of frame n).

2. *Channel* processes $\{\alpha_m(n)\}_n$, $m = 1, \ldots, M$, where $\alpha_m(n)$ is the (assumed constant) attenuation coefficient in channel m during frame n.

3. *Control* variables $R_m(n)$, $m = 1, \ldots, M$, for all n, where $R_m(n)$ specifies the rate (in bits per frame) at which terminal m transmits in frame n.

Note that the α_m and R_m are assumed constant in each frame. On the other hand, the variables $x_m(n)$ are the samples at frame boundaries of backlog processes that do vary over the course of a frame.

In Section 4.6.1, we show how terminal backlogs are representative of the rate history of the terminal. Following this, we discuss the formulation of the access control problem for the backlogs approximate model, including definition of the cost structure and transition probabilities. Results of the performance evaluation of the resulting access control are presented.

4.6.1 Equivalence between original model and backlogs approximate model

We show how the original model and the backlogs approximate model are related.

The terminal m backlog at the end of frame n is given by

$$x_m(n) = \max\left\{A_m(s,n) - D_m(s,n)\right\}, \tag{4.27}$$

where $A_m(s,n)$ is the traffic applied in intervals $[s,n]$ — or time $[T_s, T_{n+1})$ — and $D_m(s,n)$ is the traffic drained in $[s,n]$:

$$\begin{aligned} D_m(s,n) &\triangleq \sum_{u=s}^{n} R_m(u)\Delta T \\ &\equiv \sum_{u=s}^{n} b_m(u), \quad b_m(u) \triangleq R_m(u)\Delta T, \end{aligned}$$

where ΔT is the length of a unit time interval. Overflow is avoided by maintaining $x_m(n) \leq B_m$; that is by ensuring that

$$\begin{aligned} A_m(s,n) - D_m(s,n) &\leq B_m, \quad \text{for all } m, n, s \leq n \\ &\mathit{or} \\ \sum_{u=s}^{n} b_m(u) &\geq (n-s+1)B_{m_0} - B_m, \quad \text{for all } m, n, s \leq n, \end{aligned}$$

where $B_{m_0} \triangleq r_m \Delta T$.

This condition can be rewritten in the form

$$\frac{1}{k}\sum_{n=N-k+1}^{N} b_m(n) \geq B_{m_0} - \frac{1}{k}B_m, \quad \text{for all } n, k, m. \tag{4.28}$$

This is to be compared with Equation (4.11), which for convenience is reproduced here in the form

$$\frac{1}{K}\sum_{n=N-K+1}^{N} b_m(n) \geq \rho_m^*, \quad \text{for all } n, m. \tag{4.29}$$

Note that Equation (4.29) holds for a particular K, unlike Equation (4.28) that holds for all k. B_m is K frames of data (applied at rate r_m) scaled by a factor β to be chosen. We want to choose β so that Equation (4.29) approximates Equation (4.28). In terms of β, Equation (4.28) becomes

$$\frac{1}{k}\sum_{n=N-k+1}^{N} b_m(n) \geq B_{m_0}\left(1 - \frac{\beta K}{k}\right). \tag{4.30}$$

The inequality has teeth only when $k > \beta K$. If $\beta \geq 1$, then Equation (4.30) is certainly satisfied for $k = K$. The two parameters r_m and β can be varied in attempting to bridge the gap between Equation (4.28) and Equation (4.29). In our numerical studies the approach was to take $r_m = \rho_m^*$ and $\beta = 1$, selecting a rate allocation policy in the approximate model so as to control the probability of buffer overflow.

In the sequel we describe the access control problem for the backlogs approximate model.

4.6.2 Rate allocation problem formulation

We detail the rate allocation problem formulation for the backlogs approximate model. We have asserted that the vector process $x \triangleq (x_1, \ldots, x_M)$ forms a state process, and thus, in particular, that $x(\cdot)$ has Markovian statistics. This assertion will be validated by the definition of the model's cost structure c and by the following restrictions on the channel parameters α and control variables R

- For each m, the variables $\alpha_m(n)$, $n = 0, 1, 2, \ldots$, are *iid*.
- For each m and n, $R_m(n)$ is a function of the M-vectors $\alpha(n)$, $x(n)$, both of which are assumed known to the controller at the start of the frame.
- The input rates r_m (bits per frame), the output rates R_m (bits per frame) and the

buffer capacities B_m (bits) are all integer-valued. This, for computational convenience.

The randomness in the evolution of $x(\cdot)$ is due exclusively to the randomness in the depletion rates R_m, which in turn is a consequence of the randomness in the α process — just as it is for the original rate allocation problem. It is thus α, and only α, that is the source of randomness in the system. Hence, we can think of the access-rate control problem as a computation which accepts the statistics of α as input, and produces policies $u_1, \ldots, u_M$ as output, policy u_m being the function which relates $x(\cdot)$, $\alpha(\cdot)$ to R_m

$$R_m(n) = u_m(x(n), \alpha(n)), \quad n = 0, 1, 2, \ldots.$$

The solution to the access control problem is a vector $u = (u_1, \ldots, u_M)$ of policies u_m, one for each terminal. Our objective is to determine the optimal u.

The evolution of the backlog process, $x(\cdot)$, is described by the following recursion. Define

$$\Delta_m(n) \triangleq (R_m(n) - r_m)\Delta T, \qquad m = 1, \ldots, M \text{ and } n = 0, 1, 2, \ldots,$$

where $\Delta T = 1$ is the frame length. Then

$$x_m(n+1) = \begin{cases} \max\{0, x_m(n) - \Delta_m(n)\}, & \Delta_m(n) > 0 \\ x_m(n), & \Delta_m(n) = 0 \\ \min\{B_m, x_m(n) - \Delta_m(n)\}, & \Delta_m(n) < 0. \end{cases} \tag{4.31}$$

In particular, given $x(n) = x$ and strategy u, it follows that $x(n+1) = y$ if and only if for each m

$$\begin{aligned} u_m(x, \alpha(n)) &= x_m - y_m + r_m, && 0 < y_m < B_m \\ u_m(x, \alpha(n)) &\geq x_m + r_m, && y_m = 0 \\ u_m(x, \alpha(n)) &\leq x_m - B_m + r_m, && y_m = B_m. \end{aligned} \tag{4.32}$$

Equation (4.32) provides the basis for computing the transition probabilities in $x(\cdot)$ associated with a particular strategy $u(\cdot)$. To complete the formulation of the optimization problem, we have to specify these transition probabilities and the cost structure.

4.6.3 Cost structure for the backlogs model

For any strategy u subject to the restrictions described in Equation (4.32), the cost $c_u(n)$ of operating the system over the n-th frame should be a scalar function of the backlogs $\{x_m(n)\}_m$, the channel parameters $\{\alpha_m(n)\}_m$ and u. For example, we can have that $c_u(n)$ is the total volume of data lost to overflow during the n-th frame, that is,

$$c_u(n) = \sum_m c_{u,m}(n), \text{ where } c_{u,m}(n) \triangleq \max\{0, x_m(n) - \Delta_m(n) - B_m\}.$$

Alternatively, we can have that $c_u(\cdot)$ is defined as above, in terms of components $c_{u,m}(n)$, except that now $c_{u,m}(n)$ is defined by

$$c_{u,m}(n) \triangleq \begin{cases} 1 & \text{if } x_m(n) - \Delta_m(n) > B_m \\ 0 & \text{otherwise.} \end{cases}$$

In this case, $c_u(n)$ is a measure of the probability of overflow during the n-th frame.

In the context of the optimization of u, the only attribute of $c_u(\cdot)$ that actually matters is the conditional average

$$\bar{c}_u(x) \triangleq \mathbf{E}[c_u(n) \,|\, x(n) = x],$$

for all possible values x of the backlog vector $x(n)$. Observe that the expectation here is with respect to the distribution of $\alpha(n)$, which by assumption is independent of n.

4.6.4 Transition probabilities for the backlogs model

For the backlogs approximate model, the transition probabilities, $p_{ij}(u)$, from state i to state j under control u, are developed from the conditions given in Equation (4.32). For each buffer m, we have that:

$$p_{ij}(u) \triangleq P[x_m(n+1) = j | x_m(n) = i, u],$$

which is the probability that the backlog at the start of the next frame is j, given that the backlog at the start of the present frame is i and that policy u prevails there. The relations recorded in Equation (4.32), one such set for each m, are both necessary and sufficient in order that $x(t+1) = j$ when $x(t) = i$. It follows that $p_{ij}(u)$ is just the probability (relative to the distribution of $\alpha(n)$) that Equation (4.32) holds for all m. In symbols,

$$p_{ij}(u) = \Pr\{A_1 \cap \cdots \cap A_M\},$$

where Pr refers to the distribution of α and A_m (depending on the vector x and coordinate y_m) is the set of values for $\alpha(n)$ that satisfy Equation (4.32).

What is the probability of buffer overflow? The cost structure for the backlogs model will steer the system so as to minimize the frequency of buffer overflow. For buffer m, let $L_m(n)$ denote the backlog that is in excess of B_m at interval n, then:

$$L_m(n) = \begin{cases} (x_m(n) + \Delta_m(n) - B_m)^+, & r_m > R_m(n) \\ 0, & r_m \leq R_m(n) \end{cases}$$

where $(\cdot)^+ \triangleq \max\{0, \cdot\}$. The probability of buffer overflow, $p_{of}(u)$, is then given by:

$$p_{of}(u) \triangleq P[L_m(n) = l, l > 0] = \begin{cases} P[\Delta_m(n) = B_m - x_m(n) + l], & r_m > R_m(n) \\ 0, & \text{otherwise.} \end{cases} \tag{4.33}$$

Figure 4.4 is a schematic diagram of the backlog process. The transition probabilities $p_{ij}(u)$ and overflow probabilities, $p_{of}(u)$, are both functions of *only* the α process.

Armed with the cost structure and transition probabilities, we can now proceed to solve the optimal rate allocation problem using dynamic programming.

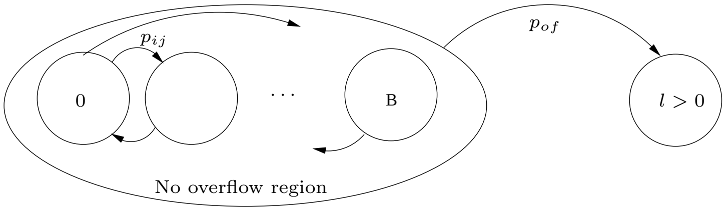


Fig. 4.4 Schematic diagram of system with buffer overflow scenario

4.6.5 Dynamic programming formulation for the backlogs model

Once again, we have an optimization problem that can be solved using dynamic programming techniques. In this case, the strategy u is to be constructed so as to minimize the steady-state average rate, expressed, say, in dollars per frame, at which cost accrues in the operation of the system. The steady-state in question is that of the backlog process $x(\cdot)$, which under the assumptions given above is a *Markov chain*. If we write π_u for the steady-state distribution of $x(\cdot)$ under u, then the objective function is in fact

$$\mathbf{E}_{\pi_u}\bar{c}(X) = \sum_x \pi_u(x)\bar{c}_u(x), \tag{4.34}$$

where X denotes a random variable whose distribution is identified in π_u, the subscript attached to the expectation operator.

It should be noted that there are two (nested) averaging operations involved in evaluation of the steady-state average cost rate: one with respect to the α_m's, used to produce $\bar{c}_u(x)$, and the other with respect to π_u. At each step n, the controller observes $x(n)$, is informed of $\alpha(n)$, combines both to construct $R_m(n)$ (from which point, till the end of that frame, there is no further randomness) and then subsides till $(n+1)$. The distribution π_u is determined by the transition probabilities $p_{ij}(u)$ described above.

The DP formulation of the optimization problem is slightly complicated by the fact that the policies u are functions of α as well as of x. One could have incorporated α into the state description, creating an augmented state variable $\tilde{x}(n) \triangleq (x(n), \alpha(n))$; the individual actions available to the controller in each state $\tilde{x} = (x, \alpha)$ would then have been ordinary real numbers $\tilde{u}(\tilde{x}) \triangleq u(x, \alpha)$. The path we chose instead, in which $x(\cdot)$ alone defines the state, is made possible by our assumption that $\alpha(\cdot)$ is memoryless from frame to frame. It entails that the action taken in each state x is in fact a whole function — the function

$u(x,\cdot)$ of α. There is thus a tradeoff between the dimensionality of the state-space and the dimensionality of the action space: the smaller the one, the bigger the other.

The decision to go with the smaller state space (and larger action space) was motivated by consideration of the *policy-iteration* DP algorithm, which is one approach to the solution of our optimal control problem. Each iteration has two parts: one part solves a system of simultaneous linear equations to obtain an approximation to the differential value (cost) function, while the other effects a minimization over the action space. The number of linear equations in question is equal to the dimensionality of the state space, meaning that the complexity of the solution step is polynomial in that parameter. The minimization step is easier. The optimization of $u(x,\cdot)$ can be done *separately* for each value of the argument α, so that complexity in the worst case is merely linear in the number of levels to which α is quantized for purposes of computation.

Applying dynamic programming techniques with discount parameter, ν, we obtain the optimal value function for each state $x \in X$ (the state space) and all α ($\alpha_m \in [0,1]$) as

$$\begin{aligned}
V^n(x) &= \min_{u(x,:)} \left[\bar{c}_{u(x,:)}(x) + \nu \sum_{y \in S} P_{u(x,:)}(x,y) V^{n-1}(y) \right] \\
&= \min_{u(x,:)} \left[\int_0^1 c_{u(x,\alpha)}(x) dF(\alpha) + \nu \sum_{y \in S} \int_0^1 P_{u(x,\alpha)}(x,y) V^{n-1}(y) dF(\alpha) \right] \qquad (4.35) \\
&= \min_{u(x,:)} \int_0^1 dF(\alpha) \left[c_{u(x,\alpha)}(x) + \nu \sum_{y \in S} P_{u(x,\alpha)}(x,y) V^{n-1}(y) \right].
\end{aligned}$$

The fact that the α are *iid* entails that the state is defined by only x. Consequently, as noted above, the optimization can be conducted as a set of independent scalar minimizations — one for each (x,α) pair separately instead of conducting a vector optimization as a function over all α, $(x,:)$. Hence, we can rewrite Equation (4.35) as

$$\int_0^1 f(\alpha) d\alpha \min_{u(x,\alpha)} \left[c_{u(x,\alpha)}(x) + \nu \sum_{y \in S} P_{u(x,\alpha)}(x,y) V^{n-1}(y) \right].$$

Moreover, from the buffer evolution conditions (Equation (4.32)), for any given u and

x, there is only one possible value of y. Hence, for each (x, α), we obtain the following simplified minimization step:

$$\min_R \left\{ [c_R | x, \alpha, R] + \nu V^{n-1}(x + r_0 - R) \right\}.$$

If we consider the α-space to be quantized and suppose that there are q values of α, we then obtain the optimal value function as

$$V^n(x) = \sum_{i=1}^{q} F_\alpha(i) \min_R \left\{ [c_R | x, \alpha_i, R] + \nu V^{n-1}(x + r_0 - R) \right\}. \tag{4.36}$$

While we have made approximations to the original access rate scheduling problem in order to scale down the solution complexity, nevertheless, we are still able to gain insight into the structure of optimal rate allocations for the time-elastic access-control problem. Such insight can be used, for example, in the design of other approximation or heuristic policies in order to yield sub-optimal or near optimal solutions for the original problem.

4.7 Optimal rate allocation policy (Backlogs approximate model)

In this section, we develop an optimal rate allocation policy for the backlogs approximate model. We apply the cost structure, transition probabilities and dynamic programming formulation discussed above. Using the policies obtained, we evaluate the steady state performance of a long-horizon policy relative to that of a short-horizon policy.

Consider a 2-user system, each with a buffer of finite size B and constant input rate r_0. The buffer output rate is governed by the system control u, given the system state and channel conditions. For each user m, $x_m = \{0, 1, 2, \ldots, B\}$, and similarly the control, $u_m = \{0, 1, 2, \ldots, B\}$. Figure 4.5 depicts the 2-user system.

The system control is such that $R(n) \triangleq (R_1(n), R_2(n)) = u(x(n), \alpha(n))$, subject to $P(R(n), \alpha(n)) \leq P_{max}$, the maximum transmission power. The attenuation coefficient, α, are assumed *iid*. We further assume that at the beginning of each frame n, the attenuation coefficient $\alpha_m(n)$ and the state $x_m(n)$ for each user m is known. Because the system state

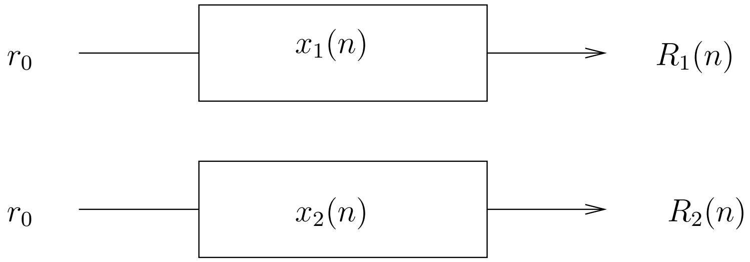


Fig. 4.5 Two-user backlogs approximate model

is defined only in terms of the buffer backlogs, the control taken at each state x is in fact a function $u(x,\cdot)$ of α. For each user m (see Equation (4.32)) we have that

$$x_m(n), \alpha_m(n) \xrightarrow{u} x_m(n) + r_0 - u(x_m(n), \alpha_m(n)).$$

The objective is to minimize the steady-state average rate at which cost accrues in the operation of the system. The system cost, $c_u(x)$, comprises a weighted combination of the buffer cost $B_c(x,\alpha,u)$ (the cost of buffer overflow events or of the data lost) and a power cost $P_c(x,\alpha,u)$. The power cost reflects our interest in controlling the long-run average energy per bit, and is defined as the total power required by both users. Our rate allocation objective is then

$$\min_u \left[(1-\sigma)B_c(x,\alpha,u) + \sigma P_c(x,\alpha,u)\right], \tag{4.37}$$

where σ is a *tradeoff coefficient* and is a design parameter.

Using well-known results, as shown in Section 4.4.3, we can obtain the average-cost optimal solution as a special case of the solution to the ν-discounted cost problem via the following approximation:

$$J_u^*(i) \approx (1-\nu) \inf_u \left[\sum_{n=0}^{\infty} \nu^n E\left[\bar{c}(x_n,u_n)|x_0=i\right]\right], \ \forall i \in X. \tag{4.38}$$

Note that here the necessary and sufficient conditions to ensure the existence of the average long-run cost are (1) a uniformly bounded cost function, $\bar{c}(x_n,u_n)$, for all n, and (2) a discount factor, $\nu \to 1$. In this work, the state and control spaces are both finite, we take $\nu \to 1$, and, for all n, the one-stage cost, $\bar{c}(x_n,u_n) \triangleq f(B, P_{max}, \alpha) < \infty$ — where B is the buffer size and P_{max} is the maximum transmission power. At each iteration, we optimize the following (Equation (4.36)):

$$J_n(i) = \sum_{i=1}^{q} F_\alpha(i) \min_R \left\{ [c_R | x, \alpha_i, R] + \nu J_{n-1}(i + r_0 - R) \right\}, \ \forall i \in X.$$

In the sequel we present the structure of the optimal access control policy.

4.7.1 Structure of the optimal policy

From the buffer evolution conditions in Equation (4.32), we can make some observations about the structure of the optimal policy. For instance, in the case of $y_m = 0$, we can assert that the optimal control is $u_m(x, \alpha(n)) = x_m + r_m$. This is because any other feasible choice makes no difference to the system state — it will remain at $y_m = 0$. Further, larger values of $u_m(x, \alpha(n))$ would result in increasing interference within the system due to the associated increase in power required to support the rate allocation.

More generally, let us consider the performance of longer-horizon policies and a short-horizon policy in a system with time-varying channel conditions. Here, a longer-horizon policy is one that selects a control based on its cost implications on both the current and subsequent states while a short-horizon policy selects a control based on its impact on the current state only.

In the following subsections we present the one-stage (short-horizon) and optimal (long-horizon) controls for a simple 2-user network. Each user has buffer size $B = 4$ and an input rate $r_0 = 4$ units. A binary-valued channel is considered, where the attenuation coefficient for a good channel is 0.8 and the coefficient for a bad channel is 0.3. We consider a cost tradeoff coefficient (Equation (4.37)) between the buffer costs and power costs of 0.4. From the controls obtained we observe that, given current channel conditions, the one-stage control will only try to transmit as much data as needed to prevent buffer overflow. Conversely, in view of the uncertainty of future channel conditions, the optimal policy will try to transmit as much data as is feasible.

One-stage control

One-stage or short-horizon control is based on information about the current state only, independent of the impact such a choice makes on the subsequent states in the process.

Table 4.1 shows the one-stage control for the case of $x_m = \{0, 1, 2, 3, 4\}$ and $\alpha_m = \{0.3, 0.8\}$. The buffer cost applied is that of buffer overflow,

$$c(x_n, u_n) = \sum_m I_{\{x_m(n)+r_0-R_m(n)>B\}},$$

where $I_{\{.\}}$ is the indicator function.

It is observed that the control policy is such that the buffer is maintained at full capacity, B, for all combinations of states and channel conditions. We see that for each (x, α), the output rate is *at most* equivalent to the buffer state, x, regardless of the prevailing channel conditions. That is, the system will expend only as much feasible power as is required to minimize buffer overflow at the current state.

Optimal control

In contrast to one-stage control, optimal (long-horizon) control is based on the current state and the impact any selected control would have on subsequent states.

Table 4.2 shows the selected control for the optimal policy. In contrast to the results obtained for short-horizon control, we observe that the optimal control seeks to transmit more than simply the current backlog, x. For example, consider the situation for buffer states 1-5. Except for the high-attenuation case of $\alpha = (0.3, 0.3)$, at least one user gets to transmit the equivalent of all incoming data to the user, $r_0 = 4$, regardless of the user's buffer state, x. Secondly, we observe that transmission is weighted in favor of the user with a good channel (higher α). Again looking at states 1-5, we see that the user with $\alpha_m = 0.8$, that in the myopic case had not transmitted, is now permitted to transmit and at full rate ($r_0 = 4,$ in this case).

At the beginning of this section on a structure for the optimal policy, we had asserted an optimal policy for the case of $x_m(n+1) = y_m = 0$ for each user m, where from the buffer evolution conditions we had that

$$u_m(x, \alpha(n)) \geq x_m + r_m, \qquad (y_m = 0).$$

Table 4.1 One-stage control for given x and α ($\sigma = 0.4$)

Buffer state		α			
	(X)	(0.3,0.3)	(0.3,0.8)	(0.8,0.3)	(0.8,0.8)
1	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)
2	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)
3	(0,2)	(0,2)	(0,2)	(0,2)	(0,2)
4	(0,3)	(0,3)	(0,3)	(0,3)	(0,3)
5	(0,4)	(0,4)	(0,4)	(0,4)	(0,4)
6	(1,0)	(1,0)	(1,0)	(1,0)	(1,0)
7	(1,1)	(1,1)	(1,1)	(1,1)	(1,1)
8	(1,2)	(1,2)	(1,2)	(1,2)	(1,2)
9	(1,3)	(1,3)	(1,3)	(1,3)	(1,3)
10	(1,4)	(1,0)	(1,4)	(1,0)	(1,4)
11	(2,0)	(2,0)	(2,0)	(2,0)	(2,0)
12	(2,1)	(2,1)	(2,1)	(2,1)	(2,1)
13	(2,2)	(2,2)	(2,2)	(2,2)	(2,2)
14	(2,3)	(2,0)	(2,3)	(2,3)	(2,3)
15	(2,4)	(2,0)	(2,4)	(2,0)	(2,4)
16	(3,0)	(3,0)	(3,0)	(3,0)	(3,0)
17	(3,1)	(3,1)	(3,1)	(3,1)	(3,1)
18	(3,2)	(0,2)	(3,2)	(3,2)	(3,2)
19	(3,3)	(0,3)	(0,3)	(3,0)	(3,3)
20	(3,4)	(3,0)	(0,4)	(3,0)	(3,4)
21	(4,0)	(4,0)	(4,0)	(4,0)	(4,0)
22	(4,1)	(0,1)	(0,1)	(4,1)	(4,1)
23	(4,2)	(0,2)	(0,2)	(4,2)	(4,2)
24	(4,3)	(0,3)	(0,3)	(4,0)	(4,3)
25	(4,4)	(0,4)	(0,4)	(4,0)	(4,4)

Table 4.2 Optimal (long-horizon) control for given x and α ($\sigma = 0.4$)

Buffer state			α		
	(X)	(0.3,0.3)	(0.3,0.8)	(0.8,0.3)	(0.8,0.8)
1	(0,0)	(0,4)	(0,4)	(4,0)	(4,4)
2	(0,1)	(0,1)	(0,4)	(4,1)	(4,1)
3	(0,2)	(0,2)	(0,4)	(4,2)	(4,2)
4	(0,3)	(0,3)	(0,4)	(4,0)	(4,3)
5	(0,4)	(0,4)	(0,4)	(4,0)	(4,4)
6	(1,0)	(1,0)	(1,4)	(4,0)	(1,4)
7	(1,1)	(1,1)	(1,4)	(4,1)	(1,4)
8	(1,2)	(1,2)	(1,4)	(4,2)	(4,2)
9	(1,3)	(1,3)	(1,3)	(4,0)	(4,3)
10	(1,4)	(1,0)	(1,4)	(4,0)	(1,4)
11	(2,0)	(2,0)	(2,4)	(4,0)	(2,4)
12	(2,1)	(2,1)	(2,4)	(4,1)	(2,4)
13	(2,2)	(2,2)	(2,2)	(2,2)	(4,2)
14	(2,3)	(2,0)	(2,3)	(4,0)	(2,3)
15	(2,4)	(2,0)	(2,4)	(4,0)	(2,4)
16	(3,0)	(3,0)	(0,4)	(4,0)	(3,4)
17	(3,1)	(3,1)	(0,4)	(3,1)	(3,4)
18	(3,2)	(0,2)	(0,4)	(3,2)	(3,2)
19	(3,3)	(0,3)	(0,4)	(4,0)	(3,3)
20	(3,4)	(3,0)	(0,4)	(4,0)	(3,4)
21	(4,0)	(4,0)	(0,4)	(4,0)	(4,4)
22	(4,1)	(0,1)	(0,4)	(4,1)	(4,1)
23	(4,2)	(0,2)	(0,4)	(4,2)	(4,2)
24	(4,3)	(0,3)	(0,4)	(4,0)	(4,3)
25	(4,4)	(0,4)	(0,4)	(4,0)	(4,4)

The assertion was that the optimal policy should be, $u_m(x, \alpha(n)) = x_m + r_m$. What is the prevailing situation? Table 4.3 and Table 4.4 show the selected controls for two different buffer sizes, $B \in \{4, 6\}$, and two different tradeoff coefficients, $\sigma \in \{0.2, 0.4\}$. Note that the premium on power costs increases with increasing σ. The $[\alpha_1 \ldots \alpha_4]$ are $[(0.3, 0.3)\ (0.3, 0.8)\ (0.8, 0.3)\ (0.8, 0.8)]$, respectively. Indeed, given input rate, $r_0 = 4$, we note that the highest output rate selected is equivalent to 4 units for buffer state $x_m = 0$ and equivalent to 5 units for buffer state $x_m = 1$. This holds true even for the case of $B = 6$ where output rates greater than 5 units are possible.

Table 4.3 Selected controls for states (0,0) and (0,1) for $\sigma = 0.2$

Buffer	Buffer	*One-stage control*				*Optimal control*			
size, B	state, x	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
4	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(3,4)	(3,4)	(4,3)	(4,4)
	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(3,4)	(3,4)	(4,3)	(4,4)
6	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,4)	(0,4)	(4,0)	(2,4)
	(0,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,5)	(0,5)	(4,1)	(2,5)

Table 4.4 Selected controls for states (0,0) and (0,1) for $\sigma = 0.4$

Buffer	Buffer	*One-stage control*				*Optimal control*			
size, B	state, x	α_1	α_2	α_3	α_4	α_1	α_2	α_3	α_4
4	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,4)	(0,4)	(4,0)	(4,4)
	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,1)	(0,4)	(4,1)	(4,1)
6	(0,0)	(0,0)	(0,0)	(0,0)	(0,0)	(0,2)	(0,4)	(4,0)	(2,4)
	(0,1)	(0,0)	(0,0)	(0,0)	(0,0)	(0,3)	(0,5)	(4,0)	(2,5)

We conclude the discussion on an optimal rate allocation policy for the 2-user backlogs model by evaluating the steady state performance of the optimal policy relative to the one-stage policy and by comparing the the long-run average system performance due to the optimal policy to that obtained for the $K = 1,\ \infty$ cases in Section 4.3.

4.7.2 Steady-state performance of one-stage and optimal policies

We compare the steady-state performance of an optimal access-control policy to a one-stage (short-horizon) policy. For buffer size, B, in the range 4 - 9, attenuation coefficient, $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1\}$, and tradeoff coefficient, σ, in the range 0.1 - 0.8, we generate an optimal control table similar to Table 4.2. Again, the one-stage cost function is given by $(1-\sigma)B_c + \sigma P_c$, where B_c denotes the buffer cost, in this case taken as the probability of buffer overflow, and P_c is the total power requirement. We then obtain the steady-state average-cost, $\bar{V} = \sum \pi^{(u)}(x) V_u^*(x)$, where $V_u^*(x)$ is the optimal cost and $\pi^{(u)}(x)$ is the steady state distribution of the buffer states, $x \in X$, given control u, σ and B.

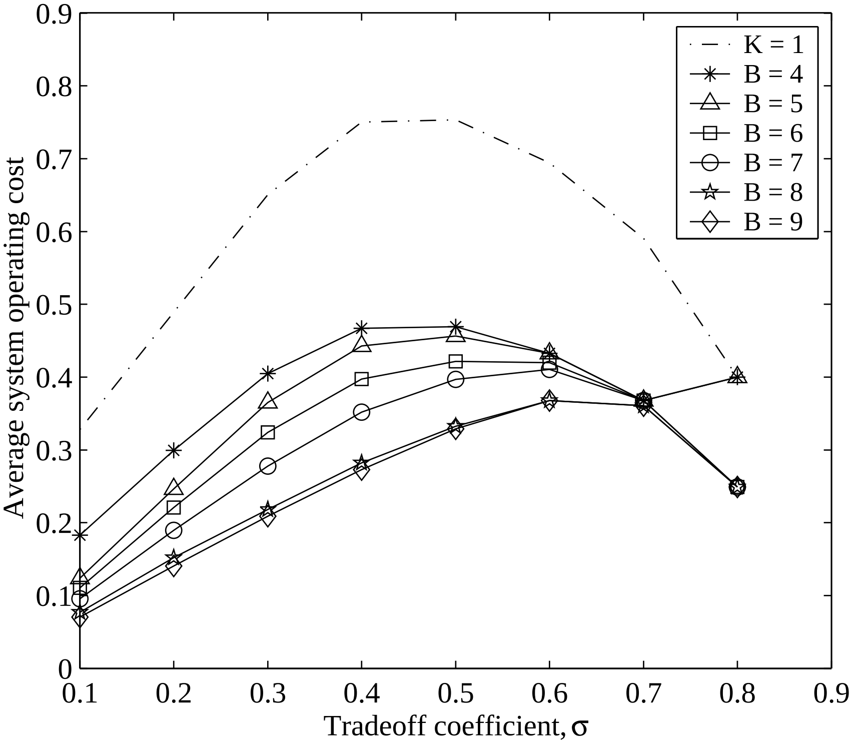


Fig. 4.6 Average operating cost versus tradeoff coefficient, σ, for $B = 4, \cdots, 9$ and for the short-horizon case $K = 1$

Figure 4.6 presents the average system operating cost for the various combinations of buffer size, B, and tradeoff coefficient, σ. We note that the one-stage horizon result is independent of B and only varies with σ. Secondly, we note that, for both the short-horizon and optimal cases, a higher cost is associated with the mid-range tradeoff coefficients. This is because here neither the premium on the power or buffer costs dominates. On the extreme ends of the σ-scale, on the low-end, the buffer costs are minimized or eliminated at the expense of power costs and the converse is true at the high-end scale of σ. Nevertheless, the optimal cost decreases with increasing buffer size, B.

Based on the optimal cost values, we can then obtain the achievable region for power requirements versus buffer overflow as shown in Figure 4.7. For each buffer size, B, each point on the curve is a tangent of the line $(P_c^*(\sigma), B_c^*(\sigma))$ for a given tradeoff coefficient, σ. We observe that the power required to achieve a given probability of buffer overflow decreases with increasing buffer size B. In the case of $B = 20$, only at the extreme high tradeoff coefficients (premium on power) is the probability of buffer overflow greater than 0. Even then, it is less than 0.05 and so barely registers in Figure 4.7.

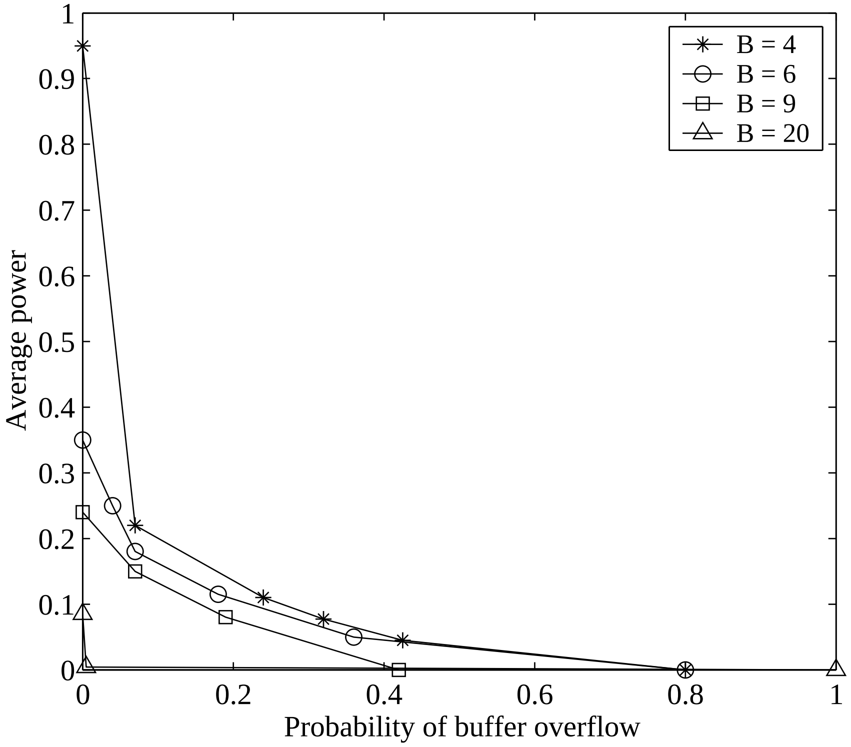


Fig. 4.7 Average power versus probability of buffer overflow for $B = 4, \cdots, 9$

In summary, time-elastic control (enabled by increasing buffer size) minimizes the steady-state average rate at which cost accrues in the operation of the system. In the following section, we compare the long-run average performance to that obtained for $K = 1$ and $K = \infty$ with enforced idleness.

4.7.3 Long-run average system performance for optimal policy

We compare the performance of the optimal access-control scheme with that of the two benchmark cases considered in Section 4.3, that is, $K = 1$ and $K = \infty$. For a system with prescribed average-rate requirement of r and equivalent buffer length, K, at each interval, the rate per terminal is r for $K = 1$ whereas for $K = \infty$ it is Mr for the single user that is permitted to transmit.

We consider the performance of the access-control policy over $N = 10^4$ time frames for 2 users with constant input rate, $r_0 = 4$. The users experience channels with homogeneous Rayleigh distribution, and for each user the attenuation coefficient, α, is *iid* from frame to frame. We apply the following long-run performance measures:

- long-run average total power: $\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\sum_{m=1}^{M} P_m(n)$
- long-run average total throughput: $\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\sum_{m=1}^{M} R_m(n)$
- long-run average energy per bit: $\dfrac{1}{M}\displaystyle\sum_{m=1}^{M}\left[\frac{\sum_n P_m(n)}{\sum_n R_m(n)}\right]$
- probability of buffer overflow: $\dfrac{1}{N}\displaystyle\sum_{n=1}^{N} I_{\{x(n)+r_0-R(n)>B\}}$

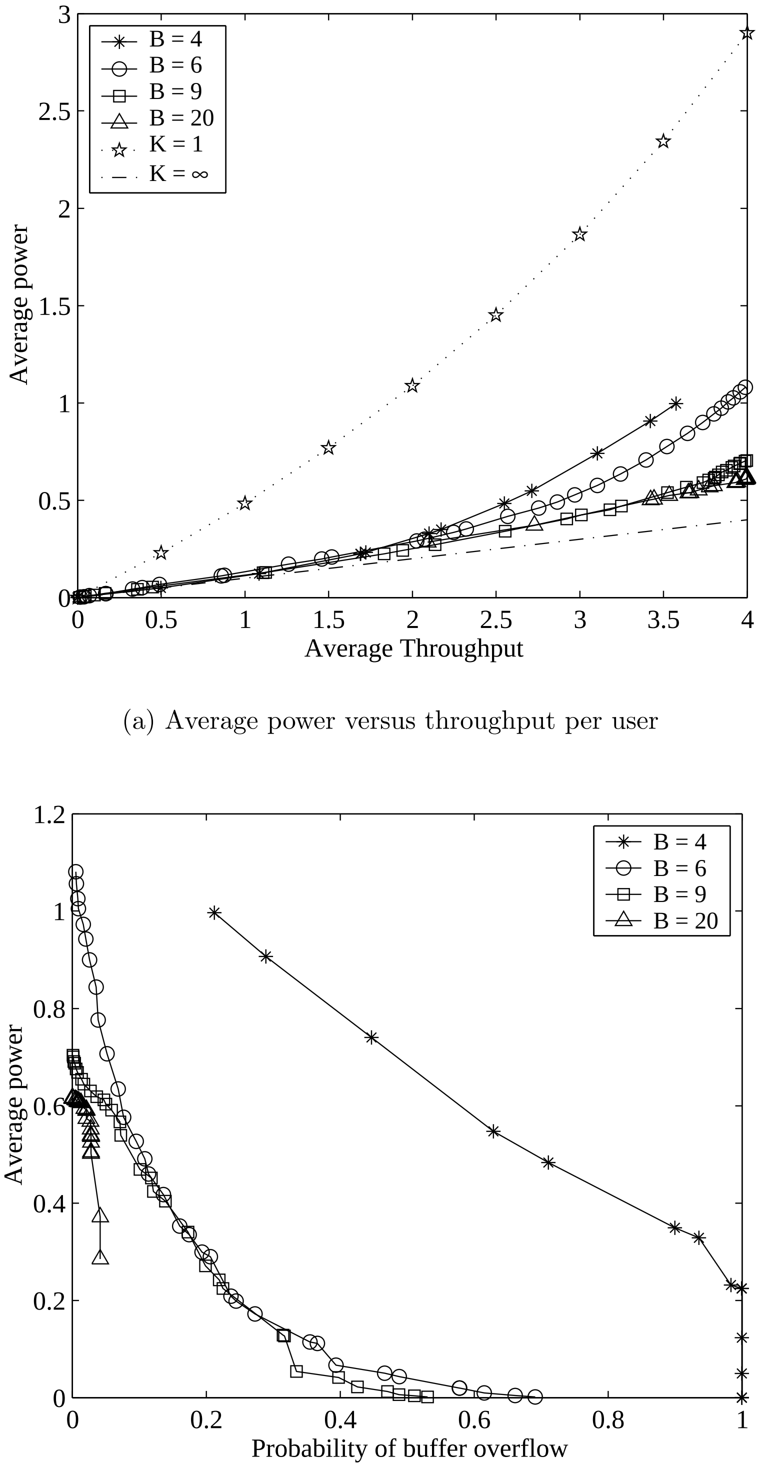
The long-run average power and throughput per user are obtained by dividing the average total measures above by the number of users, M.

Figure 4.8(a) presents the long-run average power versus throughput per user. We observe that at lower throughput values there is no clear benefit of varying the buffer sizes among $B = 4, 6, 9, 20$. However, beyond a throughput value of $0.5r_0$, we clearly note a higher power requirement with decreasing buffer size. We also note that for $B = 4$ a throughput value well less than the input rate, $r_0 = 4$, is the maximum achievable. Finally, we observe that the performance of the access-control scheme approaches that of $K = \infty$ with increasing B.

In Figure 4.8(b), we observe a similar pattern is obtained as in steady state case (Figure 4.7) for the average user power versus the probability of buffer overflow — the smaller the buffer size, the larger the transmission power required to achieve a given constraint on probability of buffer overflow. However in Figure 4.8(b) we have (i) a significant distinction of the higher power requirement for $B = 4$, (ii) a readily observable probability of buffer overflow for $B = 20$ — even though still markedly smaller than that for the other buffer sizes, and (iii) at best, the probability of buffer overflow for $B = 4$ is 0.2 and, conversely, at worst that for $B = 20$ is 0.05.

In Figure 4.9, we display the performance with varying tradeoff coefficient, σ. We focus on the region $\sigma \in [0.1, 0.7]$ which, in Figure 4.6, has a similar cost variation for the buffer sizes, $B = 4, 6, 9$. In Figure 4.9(a), in general, we have that below about $\sigma = 0.4$, the larger the buffer size, the lower the power requirement. This trend is reversed beyond $\sigma = 0.4$. As σ increases, so does the premium on power costs. With increasing buffer size and hence a reduced buffer cost contribution, the higher the B the better able the system is to maintain a greater than zero throughput without excessive overall system costs. With low B, probability of buffer overflow is high and so the overall system costs are reduced by reducing the power costs since a buffer overflow is likely to occur anyway. In Figure 4.9(b), we observe that for buffer sizes, $B_1 > B_2$, the long run average throughput $R(B_1) \geq R(B_2)$ for all $B_1, B_2 \in \{4, 6, 9, 20\}$. We also see evidence of similar behavior between $B = 20$ and the infinite buffer case for $\sigma \leq 0.5$ — in this case, $B = 20$ maintains a long-run average throughput of r_0.

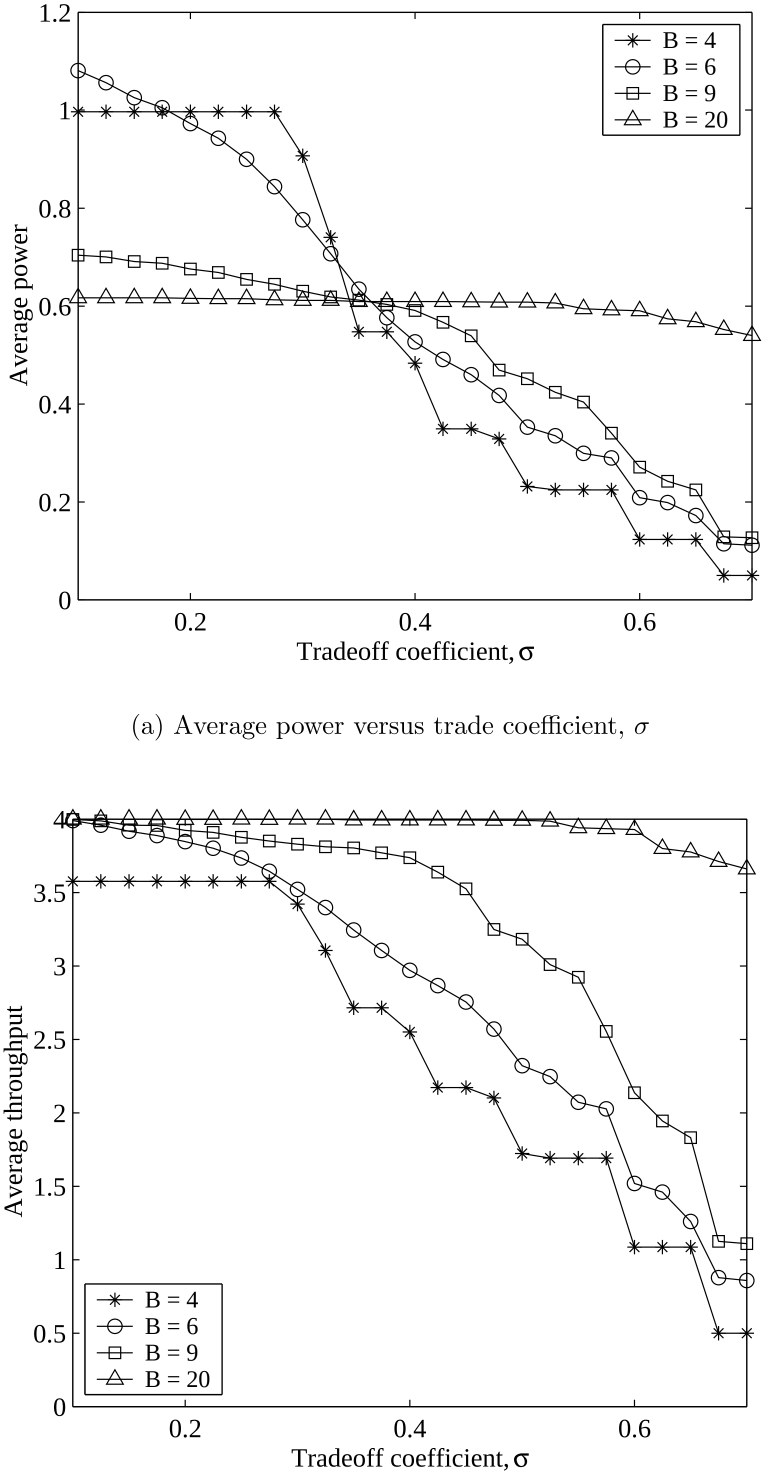
In summary, the performance of the optimal access-control approaches that of the optimal $K = \infty$ as the buffer size, B, increases. Significant similarity in performance is obtained here in the case of $B = 20$. Furthermore, a comparison of Figure 4.8(a) and Figure 4.8(b) reveals that while the $K = 1$ scenario does entail a much higher power requirement than any of the access-control options, together with the $K = \infty$ scenario, it does not suffer any loss of data. Note that if the buffer size is obtained as a measure of the user delay tolerance, then the probability of buffer overflow results are indicative of the probability that the user delay bounds will be violated.



(a) Average power versus throughput per user

(b) Average power versus probability of buffer overflow

Fig. 4.8 Long-run average performance per user



(a) Average power versus trade coefficient, σ

(b) Average throughput versus trade coefficient, σ

Fig. 4.9 Long-run average performance vs trade coefficient per user

4.8 Applying lessons learned: Rate allocation policy for original Markov model

In the preceding sections, we used a queueing model to build a Markovian approximation to the original rate allocation problem. We now apply the controls derived from the approximate model to the original problem.

Consider the two-user case, where the system state is a function of the rate history (as described in Section 4.4). Each user is considered to have a fictitious buffer that tracks the backlog resulting from the difference between the traffic applied (taken as $r_0 = \rho^*$ at each frame) and the traffic drained from the buffer over a rate averaging window, K. As before, the random attenuation coefficients for the two users are *iid* Rayleigh from frame to frame and independent from user to user.

The performance evaluation is by simulation. At the beginning of the simulation, the fictitious buffers are empty. In each time frame, the rate control is selected (using results of Section 4.7.2 and including the case of $B = 20$) according to the buffer backlogs and the values of the attenuation coefficients. Figure 4.10 presents the simulation results for the two-user system run for $N = 10^4$ time frames. Each user had an average-rate target $\rho^* = 4$. We observe that the power required to achieve a given average rate diminishes with increasing length of the averaging window, the trend here similar to what was observed for the backlogs model in Figure 4.8(a).

4.9 Concluding Remarks

In this chapter, we have developed an optimal rate allocation that exploits the difference in timescale between the user QoS requirements and the network operation. While we have considered a simple case of a 2-user network, the results obtained do highlight the case for time-elastic access rate scheduling in terms of minimizing the cost of operation for a system supporting delay-tolerant users.

In the previous chapter, we observed that adaptive beam configuration expands the achievable rate region attainable via a uniform beam configuration. From this chapter, we conclude that given a beam configuration, the cost of operation is diminished by suitable time-elastic rate scheduling. In the following chapter, we investigate heuristic and optimal

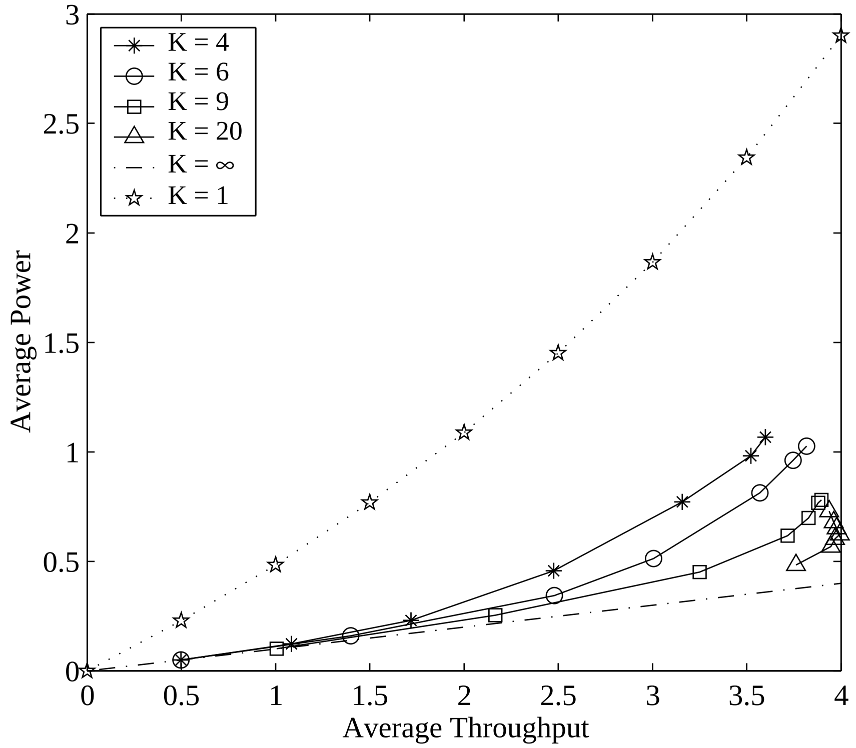


Fig. 4.10 Long-run average power vs throughput per user

solutions for the general M-user system based on the results obtained in this chapter.

Chapter 5

Access Rate Scheduling: Multi-user Scenario

5.1 Introduction

The rate control obtained in Chapter 4 addresses the simple case of a 2-terminal network. This network was in fact an approximation made so that we could gain insights into the characteristics of the optimal rate allocation policy for a system with M terminals. We now focus attention on the general case of M terminals. As before, there is an average-rate constraint, where the averaging is over a window of finite prescribed length. We are interested in the impact of rate scheduling on the energy per bit required to meet the SIR targets. We use the queueing (backlogs) model to compute the controls.

While dynamic programming techniques do yield an optimal solution, the complexity involved — due to increasing state space — makes such techniques an unattractive choice when $M > 2$. Note that a reduction in state space is possible when the terminal population is homogeneous in terms of the target rate, r, the buffer size, B, the SIR target, γ, and the distribution of the attenuation coefficient, α. We shall consider this special case in Section 5.3, where we develop an optimal access control model for homogeneous populations.

In the general case, dynamic programming (DP) remains computationally intensive. Myopic strategies, while typically sub-optimal, are usually easier to compute. An example of a myopic strategy is one that derives the power allocation such that, at each interval, the throughput is maximized subject to prevailing channel conditions and SIR constraints [130, 187]. It is shown that system throughput is maximized by allowing only one user in each time slot. It is also considered that, with time, poor channel conditions will clear and enable most users to transmit under good channel conditions. However, because of the myopic-based considerations, it is acknowledged that some users may suffer from fading as a few users may get high rates in consecutive time slots. To redress this, a variable time-slot structure is considered and user transmission scheduled in terms of the time required to transmit a fixed size packet [187]. The user with the shortest time transmits first and, again, it is considered likely that channel conditions would have improved by the time the last ranked user is scheduled to transmit.

An alternative myopic-based technique is one that builds upon DP results from a simplified network. Such a myopic strategy can result in near-optimal policies that require less intensive computation for a generalized network — and is the approach we consider. Works such as [67, 69] take a similar approach in developing 'practical' joint rate and power allocation policies based on user backlogs.

In the following section we present two myopic-based policies that are based on insights from the optimal rate allocation policy obtained in Chapter 4.

5.2 Heuristic multi-user rate allocation policies

From the optimal rate allocation policy for an $M = 2$ network, we observe that higher output rates are allocated to terminals with larger backlog, x, and higher α — see Table 4.2, for example. A myopic rate weighting that is a product of the buffer occupancy ratio and the channel conditions would support such a policy. We proceed to develop heuristic policies for a multi-user scenario that apply such a myopic-based rate weighting mechanism to guide the rate allocation strategy for the network.

A rate weighting mechanism similar to the product weighting considered here is presented in [188] where a rate and power allocation policy is developed for a satellite downlink that transmits to M locations over time-varying channels. At each interval, the allocated rate is based on a concave rate-power allocation curve reflecting diminishing returns in transmission rate with increasing power. The power allocation policy has the dual objectives to (a) provide more power to buffers with high data rates in order to maximize the throughput, and (b) provide more power to buffers with large backlog in order to minimize excessive backlogs. To achieve these objectives, while maximizing the system throughput at each interval, the allocation is conducted as an optimization problem that seeks to maximize the product of the backlog and data rate for each buffer subject to constraints on the total power. It is shown that this allocation policy is stable in terms of maximizing throughput and maintaining acceptable backlog levels even while the allocation considers only the current state information.

We propose a rate weighting, θ_m, for each user, $m = 1, \ldots, M$, such that

$$\theta_m = \left(\frac{\alpha_m}{\alpha_{max}}\right)^{w_1} \left(\frac{x_m}{B}\right)^{w_2}, \tag{5.1}$$

where α_{max} is the largest attenuation coefficient at a given time slot, $w_1 = \ln(M)$ and $w_2 = \ln(B)$. The idea is that the percentage of overall rate given to a particular terminal might reasonably depend on its value of α relative to the maximum of all the α's and its value of backlog relative to worst possible backlog; the power law form selected gives particular emphasis to values of α and x that are near maximum. Because the values of M and B can potentially be quite large, w_1 and w_2 are chosen as natural log functions of M and B respectively since $\ln(x)$ tends to infinity slower than any power of x, that is,

$$\lim_{x\to\infty} \frac{\ln(x)}{x^y} = 0, \quad \forall\, y > 0.$$

This allows the value of θ_m to vary gradually with increasing M and B. Given the weighting, we obtain the rate allocations for each user, m, as

$$R_m = \frac{\theta_m}{\Theta}\mathtt{R}, \quad \Theta \triangleq \sum_m \theta_m, \tag{5.2}$$

where $\mathtt{R}$ is a scalar quantity such that $\sum_m R_m = \mathtt{R}$. With this approach, the unknown term is a scalar quantity which should simplify the computation complexity. Based on the benchmark policies of $K = 1$ and $K = \infty$ in Section 4.3, we shall consider that $\mathtt{R} \leq Mr_0$, where r_0 is the buffer input rate for each of the M buffers.

In the sequel, we consider two algorithms that are based on this rate weighting scheme.

5.2.1 Cost-based multi-user rate allocation

In this scheme, the selection of $\mathtt{R}$ is so as to minimize the system operating cost in terms of buffer overflow and required transmission power. We apply a cost function similar to that used for the optimal rate allocation policy (in Section 4.7):

$$\min_{\mathtt{R}} \sum_{m} (1-\sigma) I_{\{x_m + r_0 - a_m \mathtt{R} > B)\}} + \sigma P_m(\mathtt{R}), \tag{5.3}$$

where $0 \leq \mathtt{R} \leq M r_0$, $a_m \triangleq \theta_m / \Theta$ and $P_m(\mathtt{R})$ is the power required by terminal m.

From Equation (5.3), we observe that power costs are decreased by reducing $\mathtt{R}$ and buffer costs are eliminated when

$$x_m + r_0 - B \leq a_m \mathtt{R}, \quad \forall\ m.$$

Summing both sides, we get

$$\mathtt{R} \geq \sum_{m=1}^{M} x_m - M B',$$

where $B' = B - r_0$. As B increases, the required $\mathtt{R}$ decreases along with the power required. Given the channel conditions, one could then determine the minimum $\mathtt{R}$ so as to minimize the probability of buffer overflow. This would be a fine scenario if channel conditions were time invariant. Since we are dealing with time-varying channels, the question is how to select a minimum $\mathtt{R}$ so as to "allow" for potential bad times ahead whereby the selected $\mathtt{R}$ is a way of "buying insurance" for future scenarios. In the optimal case, the controls for each (x, α) are selected based on all possible future scenarios. However, this is what makes the optimal control algorithm complex and time-consuming.

Ideally, we would like that for each terminal m, we have $R_m = x_m + r_0$, where x_m is the backlog for terminal m. Solving $\mathtt{R}^{(m)} = \min\{R_m / a_m, M r_0\}$ for all terminals will result in a set of $[\mathtt{R}^{(1)}, \ldots, \mathtt{R}^{(M)}]$ optimized for each terminal. Let $\mathcal{V}$ denote such a set. For each element j of $\mathcal{V}$, we obtain the corresponding one-stage cost of operation as:

$$C_{\mathcal{V}}(j) = \sum_{m=1}^{M} (1-\sigma) I_{\{x_m + r_0 - R_m > B)\}} + \sigma P_m,$$

where for each terminal m, the rate and power allocation is given by:

$$\begin{aligned} R_m &= \min\{a_m \mathtt{R}^{(j)}, x_m + r_0\}, \\ P_m &= f(R_1, \ldots, R_M, \alpha). \end{aligned}$$

The next step is to select a single system-wide $\mathtt{R}^*$ from the set $\mathcal{V}$. To obtain this, we select $\mathtt{R}^*$ such that $\mathtt{R}^* = \arg\min\{C_{\mathcal{V}}\}$. Figure 5.1 presents a flowchart of the cost-based multi-user rate allocation algorithm.

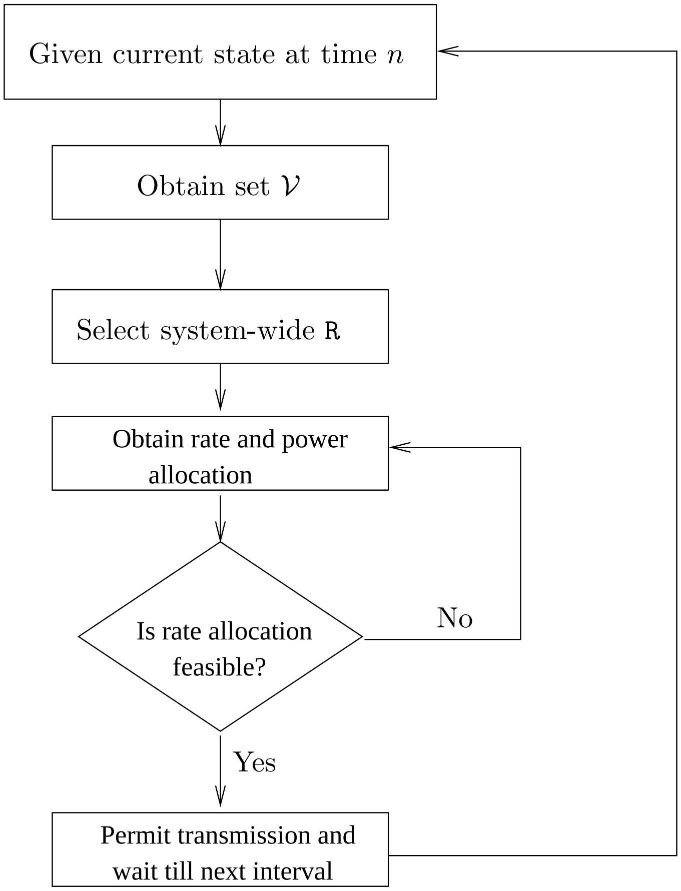

Given current state at time n
Obtain set $\mathcal{V}$
Select system-wide $\mathtt{R}$
Obtain rate and power allocation
Is rate allocation feasible?
No
Yes
Permit transmission and wait till next interval

Fig. 5.1 Flowchart of the cost-based multi-user rate allocation algorithm

Note that in a power constrained system, $\mathtt{R}^*$ may not necessarily yield a feasible rate allocation where $P_m \leq p_{max}$, $\forall m$. Hence, in such cases, an alternative feasible allocation must be sought. One option could be to only permit transmission for the terminal with the largest, a_m. Let $a_{max} \triangleq \max\{a_1, \ldots, a_M\}$, then:

$$R_m = \begin{cases} \min\left\{\frac{W}{\gamma} p_{max}\alpha_m, x_m + r_0\right\}, & a_m = a_{max} \\ 0, & a_m \neq a_{max} \end{cases}$$

Based on Figure 5.1, an experiment was conducted for $M = \{5, 10, 15, 20, 25, 30\}$ users and with varying buffer sizes, $B = \{4, 9, 20\}$. Each experiment was conducted over 10^4 frames with $W = 8 \times 10^6$, SIR $= 10$ and $r_0 = 4$. Figure 5.2 compares the results obtained to the optimal $K = \infty$ scenario. In general, the larger the number of users the smaller the average throughput.

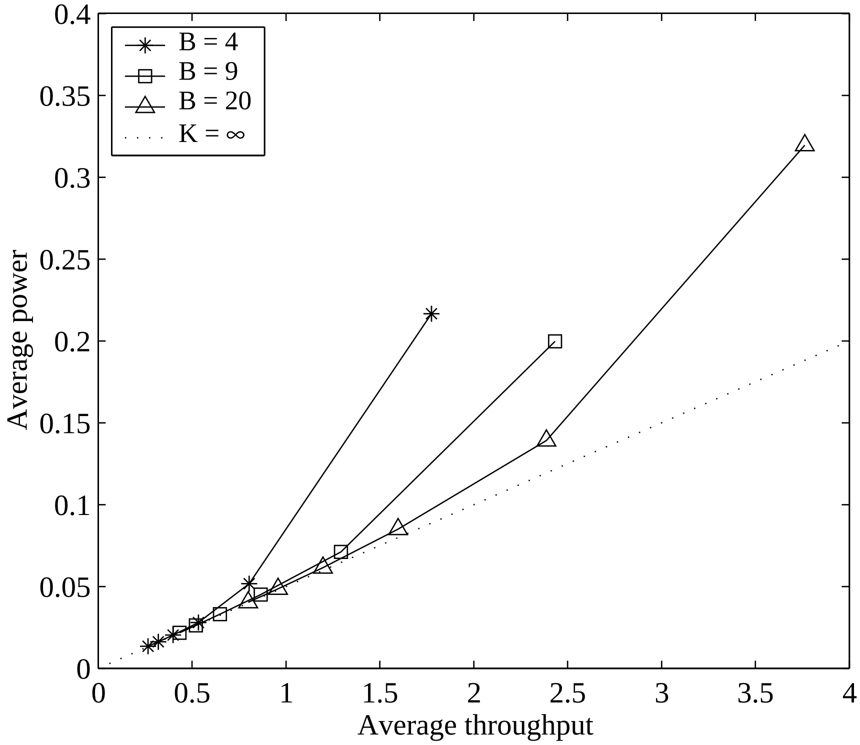


Fig. 5.2 Cost-based algorithm: Long-run average power versus average throughput per user for $M = [5, 10, 15, 20, 25, 30]$

From the figure we notice that even though the average power range remains consistent with that obtained in the $M = 2$ case (Figure 4.8(a)), the throughput is greatly limited for the small buffer sizes, $B = 4, 9$. This implies that the algorithm is better suited for traffic with relatively long delay tolerances. Nevertheless, the gains of time-elastic control are clearly evident as we observe that use of a larger buffer size results in reduced power required for the same throughput. In the next section, we consider an alternative heuristic algorithm that seeks to enhance the throughput achieved with small buffer sizes.

5.2.2 Sequential-selection multi-user rate allocation

In this case, the desired operating cost is zero and users are selected in sequential manner as long as they do not result in an operating cost that exceeds zero. The $\{a_m\}$'s represent the priorities by which users are considered for transmission — for each terminal m, the higher the value of a_m, the higher the priority of the terminal.

In addition, we amend the definition of the power cost so that instead of tracking the sum of powers required, we track the sum of power violations. The amended cost function is then given by:

$$\min_{\mathtt{R}} \sum_{m} (1-\sigma) I_{\{x_m + r_0 - a_m \mathtt{R} > B)\}} + \sigma I_{\{P_m(\mathtt{R}) > p_{max}\}}. \tag{5.4}$$

As a result of the amendment, the policy encourages transmission of more than simply what is needed to prevent overflow — since power costs are incurred only when $P_m(\mathtt{R}) > p_{max}$.

Figure 5.3 presents a flowchart of the sequential-selection multi-user rate allocation algorithm which proceeds as follows. At each interval, the $\{a_m\}$'s are arranged in order of decreasing value. Let $z = 1, \ldots, M$ denote the rankings. Beginning with the user, $z = 1$, we have that $R_{z=1} = \min\{x_{z=1} + r_0, Mr_0\}$. If $P_{z=1} \geq p_{max}$, no additional users are considered and the user $z = 1$ is allocated

$$R_{z=1} = \left(\frac{W}{\gamma} p_{max} \alpha_{z=1} \right).$$

In the case that $P_{z=1} < p_{max}$, transmission of additional users is considered — one additional user at a time as long as the operating cost does not exceed zero. A primary

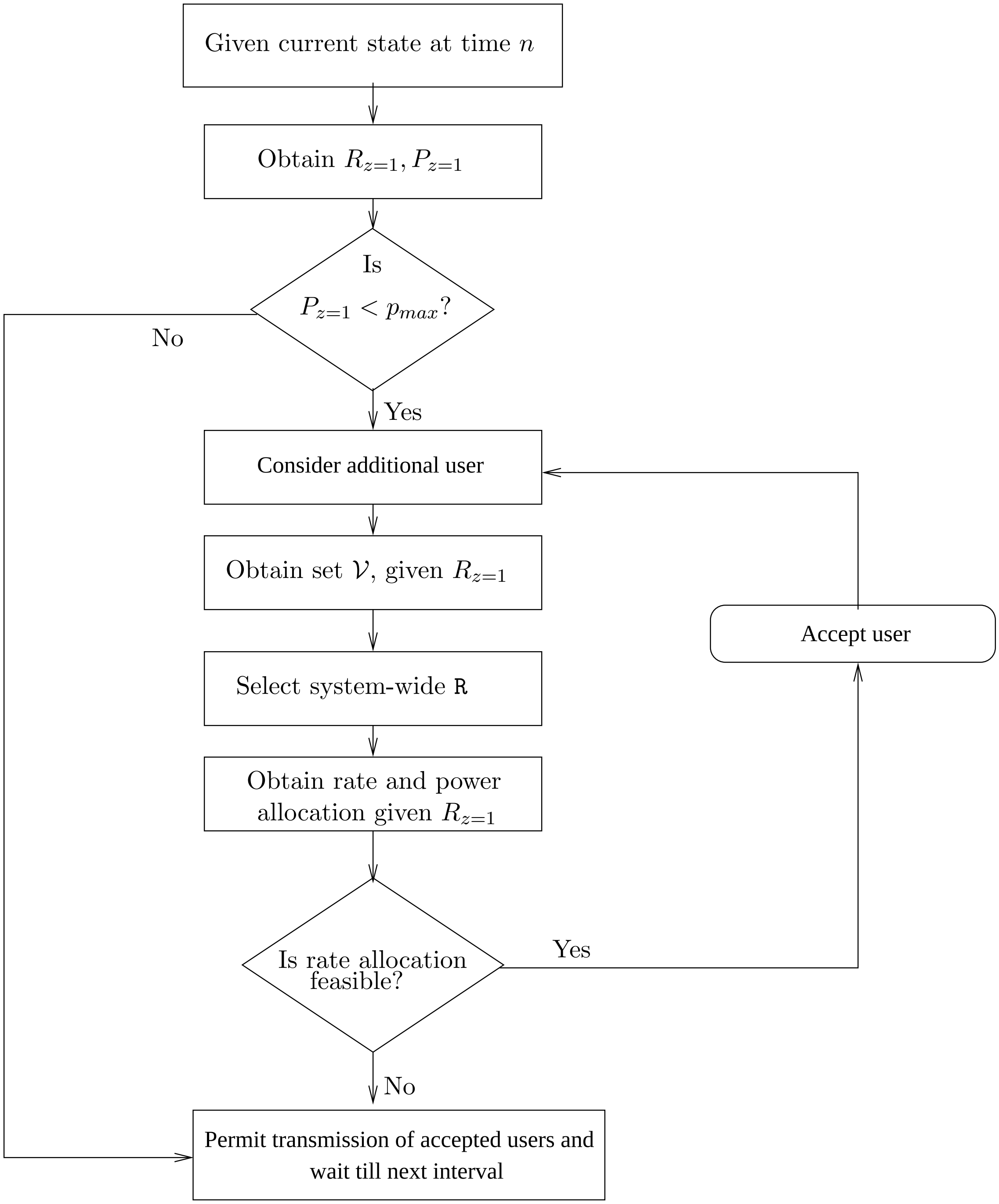

Given current state at time n
Obtain $R_{z=1}, P_{z=1}$
Is $P_{z=1} < p_{max}$?
No
Yes
Consider additional user
Obtain set $\mathcal{V}$, given $R_{z=1}$
Select system-wide R
Obtain rate and power allocation given $R_{z=1}$
Is rate allocation feasible?
Yes
No
Accept user
Permit transmission of accepted users and wait till next interval

Fig. 5.3 Flowchart of the sequential-selection multi-user rate allocation algorithm

difference with this algorithm is that the set $\mathcal{V}$ is compiled by considering the transmission permission of a single user at a time. For $z \geq 2$, as in Section 5.2.1, we have that $\mathtt{R}^{(z)} = \min\{R_z/a_z, Mr_0\}$, $R_z = x_z + r_0$. As each additional user is considered, the cost of operation, $C_{\mathcal{V}}(z)$, is determined with $R_{z=1}$ fixed as initially obtained. If $C_{\mathcal{V}}(z) \neq 0$, only users that had previously been permitted will get to transmit and no additional users are considered. Let z^* denote the largest rank permitted to transmit, then $\mathtt{R}^* = \mathtt{R}^{(z^*)}$.

Based on Figure 5.3, an experiment was conducted for $M = \{5, 10, 15, 20, 25, 30\}$ users and with varying buffer sizes, $B = \{4, 9, 20\}$. Each experiment was conducted over 10^4 frames with $W = 8 \times 10^6$, $SIR = 10$ and $r_0 = 4$. Figure 5.4 compares the results obtained to that for the optimal $K = \infty$ scenario. Again, we note the inverse relationship between the energy required for transmission and the buffer size.

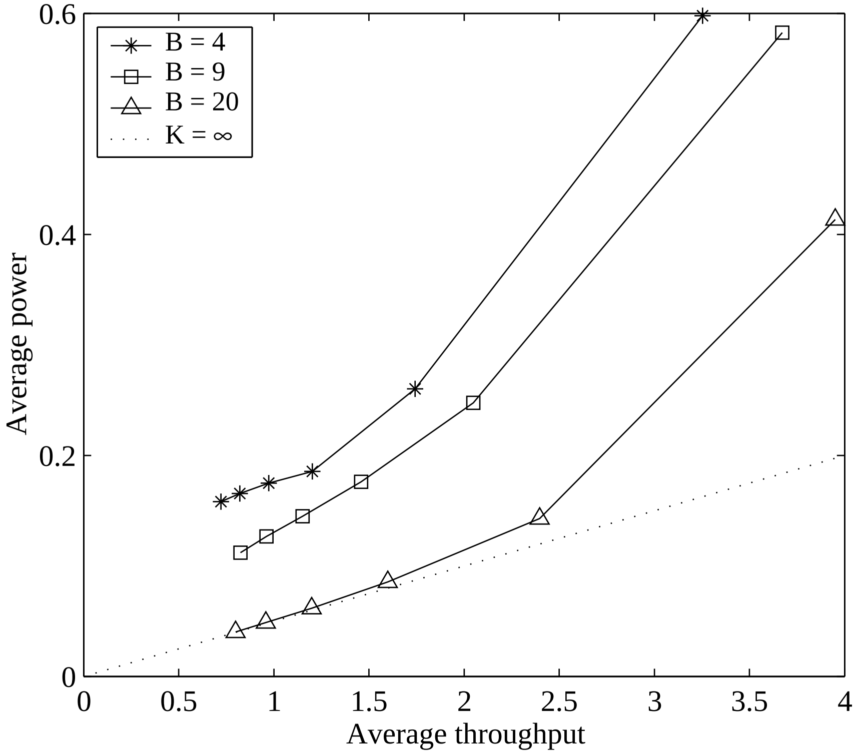


Fig. 5.4 Sequential-selection algorithm: Long-run average power versus average throughput per user for $M = [5, 10, 15, 20, 25, 30]$

From Figure 5.5, we note that the sequential-selection algorithm achieves an improvement in throughput, particularly for small buffer sizes, such as $B = 9$. This improvement is at the cost of increased average power requirements. In the case of $B = 20$, the two algorithms have similar power requirements for average rates below $0.5r_0$.

5.3 Optimal access control model for homogeneous populations

While the heuristic algorithms presented highlight the gains of time-elastic control, the optimality of the allocations remains in question. In this section we present the problem formulation for a revised dynamic programming model that is applicable in a generalized network when the source population is homogeneous. The computational savings, such as they are, flow from the reduction in the size of the state space, which in turn is achieved by suppressing source identities in the definition of the state variable. The situation is most easily described when the channel state, described in each frame by the attenuation α, is binary-valued ("Good" or "Bad"). The goal is to find a manageable approach to the computation of optimal controls when the number M of sources exceeds two.

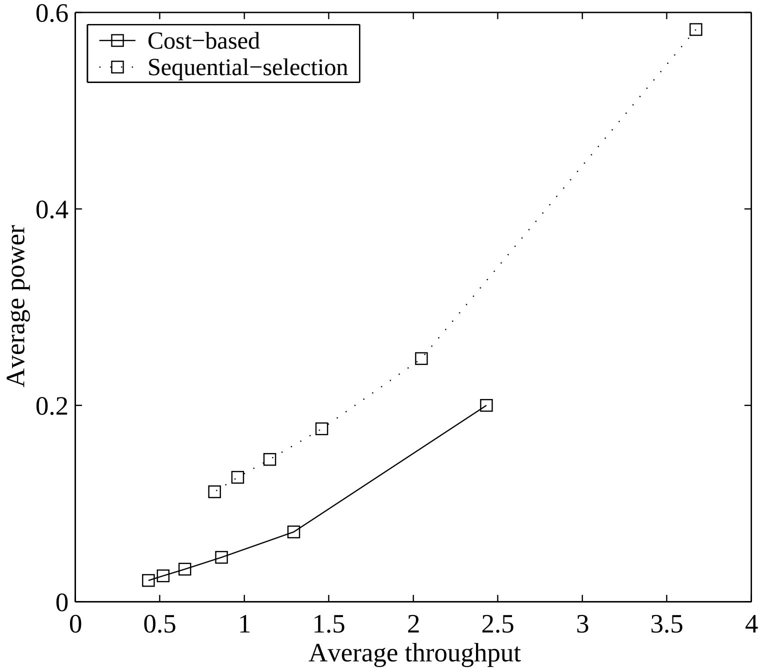
Homogeneity of the source population means that the parameters B (buffer size), γ (SIR target), p_{max} (maximum power constraint) and r_0 (data generation rate) are the same for all M sources, and that the α's (one in each frame for each source) are *iid* from source to source and from frame to frame. As before, vector $\mathbf{x} = (x_1, \ldots, x_M)$ represents the buffer backlogs at the start of a frame. Previously, we had used $\mathbf{x}$, taking $(B+1)^M$ possible values, as the state variable. The idea now is to replace $\mathbf{x}$ in that role by the (empirical) distribution of $\mathbf{x}$. The new state variable is the vector $\mu = (\mu_0, \ldots, \mu_B)$, where μ_x $(x = 0, \ldots, B)$ is the number of terminals with backlog x at the start of the frame:

$$\mu_i \triangleq \text{Card } \{m : x_m = x\}.$$

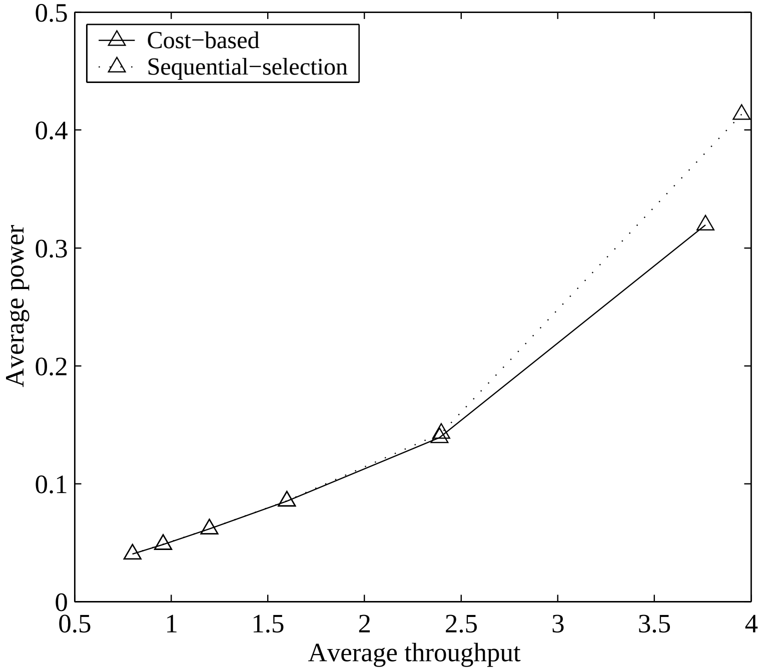
By definition, $\sum_0^B \mu_x = M$. It follows that the number of possible μ vectors, identical to the number of possible partitions of M objects into $B+1$ cells, equals

$$\binom{M+B}{M} = \frac{(M+B)\cdots(M+1)}{B!}. \tag{5.5}$$

The size of the state space, previously exponential in M, is now polynomial in M. By way of example: when $M = B = 5$, there are 252 states in the new model, *versus* 7776 states



(a) Performance comparison of cost-based and sequential-selection algorithms for $B = 9$



(b) Performance comparison of cost-based and sequential-selection algorithms for $B = 20$

Fig. 5.5 A comparison of long-run average power versus average throughput for the cost-based and the sequential-selection algorithms

in the old one. When $M = 2$ (the case already studied in detail), a buffer of size $B = 15$ yields 256 states in the old model and 135 states in the new one. The hope is that the redefinition of the state variable will bring more interesting values of M within reach of computation, at least for modest values of B.

Vector μ represents the data carried forward from frame to frame in the continual revision of the rate allocation. Within each frame that allocation depends not only on μ, but also on the way in which particular values of α are associated with particular backlogs x. In the present setting, α being binary-valued by assumption, that association can be represented by the vector $\lambda = (\lambda_0, \ldots, \lambda_B)$, where λ_x $(x = 0, \ldots, B)$ is the number of *good* channels among the μ_x channels at backlog x.

The vector pair (μ, λ) provides the data in each frame from which the rate allocation is to be constructed. Write R for any such allocation. Where μ, λ are given, R (or $R_{(\mu,\lambda)}$) is simply an M-vector of integers specifying the volume of data carried in that frame on each of the M channels; the indexing of the components of such R (an issue in that the correspondence between backlog and terminal identity is not carried in the state descriptor) is described below. In the absence of reference to specific (μ, λ), R denotes a *policy*, a *rule* which attaches an integer-valued M-vector to all possible combinations of μ and λ. The number of equations in the DP formalism equals the number, displayed in Equation (5.5), of possible values of μ. Each such equation, corresponding to a particular value of μ, includes a minimization over R that can be computationally onerous. We proceed to describe the situation in more detail.

The DP Optimality Equations, relative to the revised definition of the state variable, have the form

$$V(\mu) = \min_R \left\{ c_R(\mu) + \beta \sum_{\mu'} p_R(\mu' \,|\, \mu) V(\mu') \right\}, \quad \text{(all states } \mu\text{).}$$

The minimization is over *policies* R. The notation is otherwise standard: $V(\cdot)$ is the optimal value function (now a function of μ), β is the discount factor and c_R is the one-step average cost, a function both of R and of μ. The transition probabilities $p_R(\mu' \,|\, \mu)$ characterize the Markovian evolution of μ under policy R. Bearing in mind that the randomness in the system is due exclusively to λ, we can reformulate the Optimality Equations so that the RHS averaging is with respect to λ rather than to μ':

$$V(\mu) = \min_R \sum_\lambda p(\lambda \,|\, \mu) \left(c_R(\mu, \lambda) + \beta V(\mu'_R(\mu, \lambda))\right).$$

We have used $c_R(\mu, \lambda)$ for the one-step cost given R, μ and λ, and $\mu'_R(\mu, \lambda)$ to denote the state which *follows* μ under the combined influence of λ and R. The conditional distribution of λ given μ is *independent* of R and simply described:

- The variables $\lambda_0, \ldots, \lambda_B$ are statistically independent;
- The variable λ_x $(x = 0, \ldots, B)$ has the BINOMIAL(μ_x, ϵ) distribution, where ϵ is the probability that any particular channel is a good one.

Notice that policy R can be optimized separately for each (μ, λ) pair. It follows that the Optimality Equations can be recast one more time:

$$V(\mu) = \sum_\lambda p(\lambda \,|\, \mu) \min_R \left\{c_R(\mu, \lambda) + \beta V(\mu'_R(\mu))\right\}. \tag{5.6}$$

This new DP formulation is distinguished from the previous one (Section 4.4) in two respects: first, the minimization, previously over *policies* R, is now over M-vectors, which are simpler; second, the argument of $V(\cdot)$ on the RHS [1] is *independent* of λ. The R which achieves the minimum in Equation (5.6) is the optimal $R_{(\mu,\lambda)}$.

Remark. *The index λ on the RHS of Equation (5.6) is a vector of dimension $B+1$. Its range, given μ, has cardinality $(\mu_0+1)\cdots(\mu_B+1)$. Since $\sum_x \mu_x = M$, the latter is bounded above by*

$$\left(1 + \frac{M}{B+1}\right)^{B+1}.$$

This follows directly from

$$\left(1 + \frac{u_1+u_2}{2}\right)^2 > (1+u_1)(1+u_2) \textit{ for all } u_1, u_2, u_1 \neq u_2$$

[1] An admitted abuse of notation: $\mu'_R(\mu)$ denotes the state that ensues when M-vector R is applied in state μ.

Thus, for example, when $M = 10$, $B = 4$, the sum in the Optimality Equation for any μ has at most 243 terms. For the same values of M and B there are 1001 possible states μ. A single cycle of the Value Iteration Algorithm in that case thus includes at most 244,000 instances of the minimization step prescribed in the RHS of the Optimality Equations.

It remains to describe the parameterization of $R_{(\mu,\lambda)}$. Assume, then, that μ, λ are fixed. The associated rate allocation, which for the moment we simply write R, suppressing the subscripts for convenience, is an M-vector. Inasmuch as source identities are irrelevant here, we are free to choose the source indexing. Think of the M sources as organized into two disjoint blocks. The sources in the first block, indexed $1, \ldots, \Lambda$ where $\Lambda \triangleq \lambda_0 + \cdots + \lambda_B$, are those assigned a *good* channel. Those in the second block, indexed $\Lambda + 1, \ldots, M$, are those assigned a *bad* channel. Within each block the sources are enumerated in *descending* order of backlog x. Vector R inherits that same two-fold partition, one block for the good channels and one for the bad. We regard as self-evident, but have not proved, the following:

- The rate allocated to a source with larger backlog is no less than the rate allocated to a source with smaller backlog if the *channel conditions* (good or bad) are the same for both sources.
- The rate allocated to a source with a good channel is no less than the rate allocated to a source with a bad channel if the *backlogs* are the same for both sources.

These two principles can be used to narrow the search for an optimal allocation policy. The first one, in light of the scheme selected for indexing sources, implies that the rate vector R is *non-increasing* within each block. Each of the M coordinates of R takes one of $B + 1$ possible values, corresponding to the $B + 1$ possible values of backlog at the end of the frame. It follows that the minimization in Equation (5.6) can be effected, in the worst case, by testing at most

$$\binom{\Lambda + B}{B}\binom{M - \Lambda + B}{B} \tag{5.7}$$

distinct R vectors. The bound is arrived at by noting that the number of non-increasing functions with domain of cardinality a and range of cardinality b is equal to the number of ways in which a objects can be separated into b bins without ordering [189]; that is,

$$\binom{a+b-1}{a}.$$

The bound in Equation (5.7), specialized to the case $M = 10$, $B = 4$, is never greater than

$$\binom{5+4}{4}^2 \approx 16,000.$$

Following on from the two principles we can also state that the rate allocated to a source with a good channel and large backlog is no less than the rate allocated to a source with a bad channel and small backlog. However, the structure of the converse scenario — good channel and small backlog versus bad channel and large backlog — remains to be investigated.

5.3.1 Steady-state performance of myopic and optimal policies

In this section we compare the steady-state performance of the optimal access-control policy to a myopic policy. For each M and B, we obtain the steady-state average-cost as

$$\bar{V} = \sum_{\mu} \pi^{(u)}(\mu) V_u^*(\mu),$$

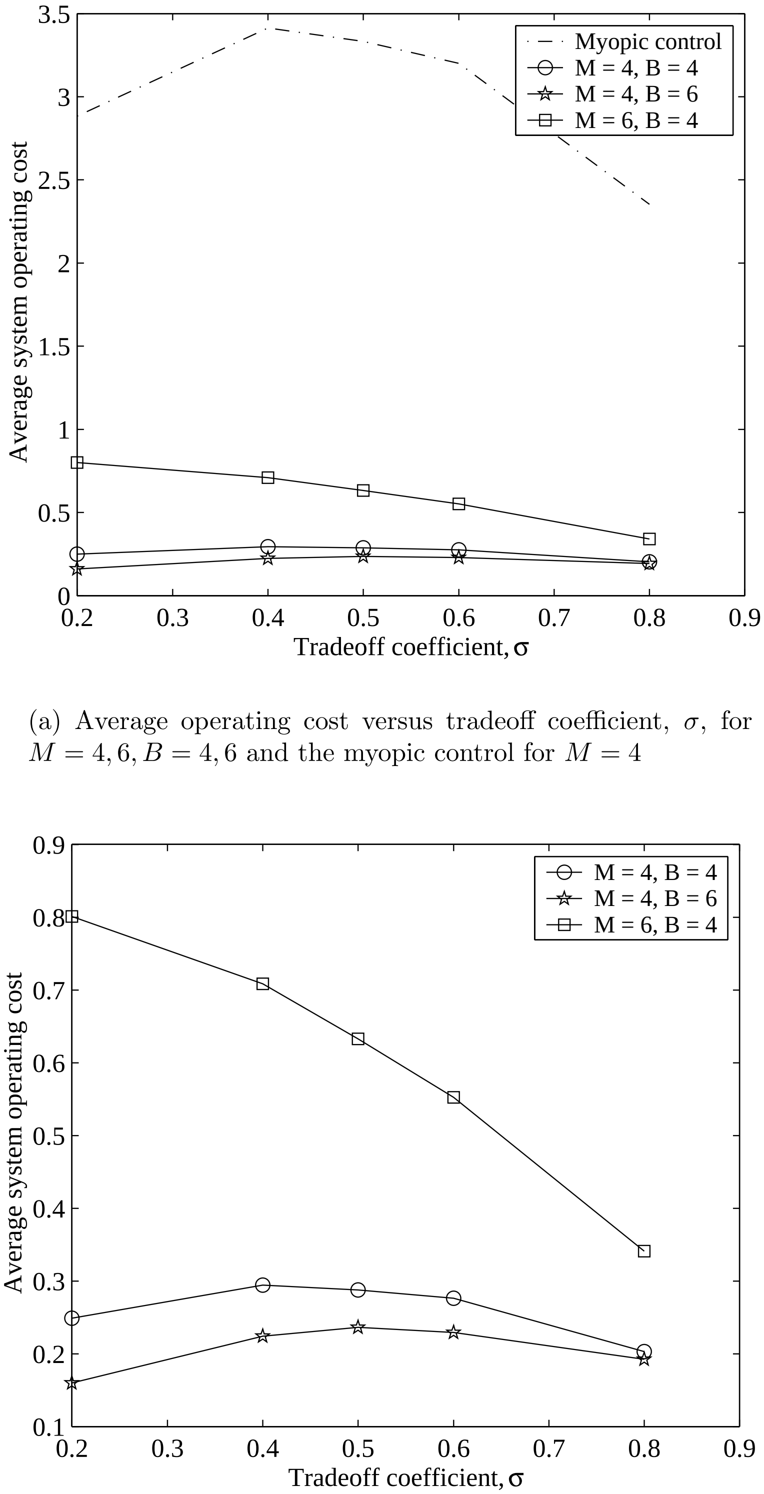
where $V_u^*(\mu)$ is the optimal cost as defined in Equation (5.6) and $\pi^{(u)}(\mu)$ is the steady state distribution of partitions, μ, given control u. The one-stage cost function is given by $(1-\sigma)B_c + \sigma P_c$, where σ is the tradeoff coefficient, B_c denotes the buffer cost, in this case taken as the data volume lost, and P_c is the total power requirement.

Figure 5.6(a) presents the average system operating cost for various combinations of $M \in \{4, 6\}$, $B \in \{4, 6\}$, and $\sigma \in \{0.2, 0.4, 0.5, 0.6, 0.8\}$. A 2-channel system was considered whereby the attenuation coefficient of a good channel was $\alpha_G = 0.8$ and of a bad channel was $\alpha_B = 0.3$. From the results obtained, we note that in all cases, the optimal costs are less than the costs of myopic operation. The myopic policy shown here is for (μ, λ) when $M = 4$. As obtained in Section 4.7.2, the myopic control is independent of buffer size B. Secondly, we note that the operating cost increases with the number of users, M, but decreases with increasing buffer size, B, as emphasized in Figure 5.6(b). In summary, as noted in the simple two-user case in Section 4.7.2, we observe that time-elastic control (enabled by increasing buffer size) minimizes the steady-state average rate at which cost accrues in the operation of the system.

5.4 Concluding Remarks

Based on insights from the optimal rate allocation policy given by the backlogs approximate model, we have developed heuristic algorithms for the multi-user scenario of $M > 2$. The algorithms select a control based on the current state only but use a rate weighting that is a product of the buffer occupancy ratio and the channel conditions. This was motivated by the optimal rate allocation policy in which it was observed that higher output rates are allocated to terminals with high buffer occupancy, x, and high attenuation coefficients, α. Both algorithms revealed a clear case for time-elastic access control which yields less average power for the same throughput.

However, the optimality of the heuristic algorithms remains in question and hence we have developed a dynamic programming model that is applicable for homogeneous source populations. This is certainly a special case and therefore further work is required to translate the insights from this case to a more general scenario of inhomogeneous populations.



(a) Average operating cost versus tradeoff coefficient, σ, for $M = 4, 6, B = 4, 6$ and the myopic control for $M = 4$

(b) Average optimal cost versus tradeoff coefficient, σ, for $M = 4, 6, B = 4, 6$

Fig. 5.6 Average optimal cost versus tradeoff coefficient for varying M and B

Chapter 6

Conclusion

6.1 Thesis Summary

We studied resource management in the context of the uplink of a multibeam CDMA-based GEO satellite network, focusing on two specific problems:

1. Beam management - Given a two-dimensional distribution of users over the satellite coverage area, we were interested in beam management techniques that yield roughly equal numbers (or mean numbers) of users per beam in cases where the user quality-of-service requirements are homogeneous. We were also interested in the impact of such techniques on the achievable rate region, compared to the achievable rate region for beam management techniques that ignore the geographical distribution of users.
2. Rate allocation - We were interested in the tradeoff between transmission rate and transmission power that is achieved by exploiting temporal elasticity in user quality-of-service requirements.

The following summarizes the results. Section 6.2 presents directions for future work.

6.1.1 Beam Management

The user distribution was assumed governed by inhomogeneous spatial Poisson. The problem was to generate the beam pattern over the geographic area. Our approach to a solution was via nonlinear programming. The coverage area was quantized into a grid of user locations, and an objective function was specified to determine the cost associated with a given allocation of terminals among the beams.

We developed two beam allocation algorithms. The key elements of these algorithms were a beam allocation function and the Hooke and Jeeves search method. In the first algorithm, the beam allocation function distributes terminals such that a terminal is assigned to the beam whose center is the shortest Euclidean distance away from the terminal. The Hooke and Jeeves method is used to search across the coverage area in search of an improving direction with respect to objective function. The second algorithm is built on the same principle. However, in addition, vector quantization techniques are applied at each beam allocation stage to minimize the separation distance between terminals and their beam centers. The performance of the algorithms developed is measured in terms of a Beam Variability Factor (BVF). The BVF is defined as the ratio of the standard deviation to the mean of the vector of mean users per beam. The more evenly balanced the beam configuration, the smaller the BVF. The two algorithms outperformed the uniform beam allocation. The second algorithm (incorporating beam size adjustment) provided only marginal performance improvement relative to the first algorithm.

We compared the performance of a uniform and an adaptive allocation in terms of the achievable rate region and the minimum power required to achieve a given throughput. Relative to the uniform beam allocation, adaptive beam allocation was observed to provide increased system throughput. The adaptive scenario required a higher power allocation because more terminals are permitted to transmit. It should be noted that the allocation algorithms developed can be adapted to a variety of network performance criteria by changing the form of the objective function.

6.1.2 Rate Allocation

We were interested in the tradeoff between transmission rate and transmission power when exploiting temporal elasticity in user quality-of-service requirements. Instantaneous transmission power was subject to an upper bound; and instantaneous transmission rates were assumed continuously variable and subject to network control. The objective was to allocate rates to users in each frame so as to minimize power or energy per bit while meeting the average-rate and SIR constraints; where the averaging in the definition of average-rate is over a sliding window of prescribed finite length. The power savings achievable by exploiting the temporal elasticity in the specification of the rate targets were confirmed through analysis of two extremal cases (short averaging window, infinite averaging window).

We developed a Markovian model of the multi-user access system so as to bring the problem within the purview of Markov Decision Theory and Dynamic Programming (DP). However, even with all the tools necessary to solve for the optimal policy that minimizes the long-run average cost of the system, the associated computational complexity was significant. To explore the structure of optimal policies, a simple two-user network was considered. The original model, in which state is represented by the whole rate history over one averaging window, was approximated by a Markovian queueing model in which buffer backlogs are used to capture the rate history.

The structure of the optimal policy revealed that higher rates were allocated to those terminals with high buffer occupancy and clearer channels. Based on this, two heuristic policies were designed to guide the rate allocation process for a general multi-user network. In this case, the control was based on the current system state and the prevailing channel conditions. Performance evaluation showed the same trends as obtained for the optimal two-user scenario, again confirming time-elastic control does yield savings in the long-run average energy per bit per terminal.

We also considered a general multi-user scenario in which terminal parameters and QoS requirements are homogeneous. The channel conditions remained *iid* from terminal to terminal and from frame to frame. In this case, we showed that by suppressing source identities within the state definition, thereby reducing the state space, we can reduce DP complexity and complete the calculation of optimal controls.

In general, increasing the buffer size, thereby increasing the time elasticity of the access-rate control, reduces the average transmission power required to support a given rate allocation.

6.2 Further directions

We conclude with possible directions for further work on the two problems we studied.

Beam management

- The user distribution was assumed inhomogeneous spatial Poisson. Future work might look at other distributions, such as the generalized Gaussian distributions considered in [115, 190], to determine sensitivity of the solution to detailed statistics.
- Future work might consider the joint impact of geographic user distribution and propagation environments on gains in system capacity attributed to beam shaping.
- Future work might consider the case where the number of beams is also variable, and its impact on adaptive beam management.

Rate allocation

- Inherent in Equation (3.7), relating SIR to rate and power allocations, are the CDMA access signaling format and the single-user receiver architecture assumed in this thesis. If the receivers are upgraded to include multi-user detection capability, then the formulas will change, but the general formulation of the resource management problem will not. Further work might thus extend the problem considered here to systems with alternative receiver structures, and alternative signaling formats, such as TDMA.
- Future work might evaluate the role of interbeam interference and propagation delay in satellite resource management.
- We have assumed that channel conditions can be estimated on a frame-by-frame basis. Invariably, we can expect that there will be errors in the rate control. Future work might evaluate the rate allocation strategy under conditions of imperfect rate or power control.
- We have also assumed that attenuation coefficients, α, are *iid*. Future work might consider Markov-dependent attenuation coefficients within each channel, which are more realistic.
- Future work might develop better heuristics for large M number of users.
- Future work might consider the rate allocation problem in non-geostationary satellite networks, such as LEO satellite networks.

References

[1] B. A. Pontano, “Satellite Communications: Services, Systems & Technologies,” *IEEE International Microwave Symposium Digest*, vol. 1, pp. 1–4, 1998.

[2] A. Jamalipour and T. Tung, “The Role of Satellites in Global IT: Trends and Implications,” *IEEE Personal Communications*, pp. 5–11, June 2001.

[3] A. Iera, A. Molinaro, and S. Marano, “IP with QoS Guarantees via GEO Satellite Channels: Performance Issues,” *IEEE Personal Communications*, pp. 14–19, June 2001.

[4] P. J. Garland, F. Hayes, and P. Takats, “An Overall Architecture for a North American Multimedia Satcom System,” in *Proceedings of the Second Ka Band Utilization Conference and International Workshop on SCG11*, (Florence, Italy), pp. 35–42, Sept. 1996.

[5] G. Losquadro, “EUROSKYWAY: Satellite System for Interactive Multimedia Services,” in *Proceedings of the Second Ka Band Utilization Conference and International Workshop on SCG11*, (Florence, Italy), pp. 13–20, Sept. 1996.

[6] G. Losquadro, A. Aerospazio, and R. E. Sheriff, “Requirements of Multiregional Mobile Broadband Satellite Networks,” *IEEE Personal Communications*, pp. 26–30, Apr. 1998.

[7] W. Wu, E. F. Miller, W. L. Pritchard, and R. L. Pickholtz, “Mobile Satellite Communications,” *Proceedings of the IEEE*, pp. 1431–1448, Sept. 1994.

[8] A. Iera, A. Molinaro, P. Pace, and S. Marano, “Multimedia Traffic in Broadband Satellite Networks,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 428–432, 2003.

[9] Q. Liu and J. Li, “Multiple Access in Broadband Satellite Networks,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 417–421, 2003.

[10] F. Gargione, T. Iida, F. Valdoni, and F. Vatalaro, "Services, Technologies and Systems at Ka Band and Beyond - A Survey," *IEEE Journal on Selected Areas in Communications*, pp. 133–144, Feb. 1999.

[11] F. Yegenoglu, R. Alexander, and D. Gokhale, "An IP Transport and Routing Architecture for Next-Generation Satellite Networks," *IEEE Network*, pp. 32–38, Sept. 2000.

[12] C. G. F. Valadon, G. A. Verelst, P. Taaghol, R. Tafazolli, and B. G. Evans, "Code-Division Multiple Access for Provision of Mobile Multimedia Services with a Geostationary Regenerative Payload," *IEEE Journal on Selected Areas in Communications*, pp. 223–237, Feb. 1999.

[13] J. Farserotu and R. Prasad, "Broadband wide-area networking via IP/ATM over SATCOM," *IEEE Journal on Selected Areas in Communications*, pp. 270–285, Feb. 1999.

[14] Ray E. Sheriff and Y. Fun Hu, *Mobile Satellite Communication Networks*. John Wiley & Sons, Ltd, 2001.

[15] E. Lutz, M. Werner, and A. Jahn, *Satellite Systems for Personal and Broadband Communications*. Springer-Verlag, 2000.

[16] J. M. Gómez, *Satellite Broadcast Systems Engineering*. Artech House, Inc., 2002.

[17] J. V. Evans, "Communication Satellite Systems for High-Speed Internet Access," *IEEE Antennas and Propagation Magazine*, vol. 43, pp. 11–22, Oct. 2001.

[18] T. Le-Ngoc, V. Leung, P. Takats, and P. Garland, "Interactive Multimedia Satellite Access Communications," *IEEE Communications Magazine*, vol. 41, pp. 78–85, July 2003.

[19] B. Fan, R. Tafazolli, and B. Evans, "Connection Management for Broadband Mobile Satellite Systems," *IEE Proceedings on Communications*, vol. 42, pp. 298–303, Aug. 2003.

[20] G. Akkor, M. Hadjitheodosiou, and J. S. Baras, *IP Multicast via Satellite: A Survey*. The Center for Satellite and Hybrid Communication Networks, Jan. 2003. CSHCN TR 2003-1.

[21] L. Bella and B. Barani, "An Experimental On-board Processing Satellite System providing ISDN Services," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, pp. 1144–1148, 1990.

[22] International Telecommunication Union (ITU), *Handbook on Satellite Communications*. John Wiley & Sons, Inc., 2002. Third Edition.

[23] F. Manshadi and V. Jamnejad, "Technology Challenges and Trade-Offs in Application of Multiple-Beam Antennas to NASA's Personal Access Satellite System," in *IEEE Aerospace Applications Conference Digest*, vol. 3, pp. 3/1–3/11, 1991.

[24] G. Washington, H.-S. Yoon, M. Angelino, and W. H. Theunissen, "Design, Modeling and Optimization of Mechanically Reconfigurable Aperture Antennas," *IEEE Transactions on Antennas and Propagation*, vol. 50, pp. 628–637, May 2002.

[25] M. Shimada, S. Yoshimoto, Y. Suzuki, and T. Iida, "Antenna-Beam Allocation for a Satellite-Fed Millimeter-Wave Personal Communication System," *Electronics and Communications in Japan, Part 1: Communications*, pp. 70–77, Dec. 1988.

[26] R. H. Roy, "An Overview of Smart Antenna Technology: The Next Wave in Wireless Communication," in *Proceedings of the IEEE Aerospace Conference*, pp. 339–345, 1998.

[27] A. U. Bhobe and P. L. Perini, "An Overview of Smart Antenna Technology for Wireless Communications," in *Proceedings of the IEEE Aerospace Conference*, pp. 2/875–2/883, 2001.

[28] B. D. V. Veen and K. M. Buckley, "Beamforming: A Versatile Approach to Spatial Filtering," *IEEE ASSP Magazine (IEEE Signal Processing Magazine)*, vol. 5, pp. 4–24, Apr. 1988.

[29] S.-I. Jeon, Y.-W. Kim, and D.-G. Oh, "A New Active Phased Array Antenna for Mobile Direct Broadcasting Satellite Reception," *IEEE Transactions on Broadcasting*, vol. 49, pp. 34–40, Mar. 2000.

[30] E. Lier and D. S. Purdy, "Techniques to Maximize Communication Traffic Capacity in Multi-Beam Satellite Active Phased Array Antennas for Non-Uniform Traffic Model," in *Proceedings of the IEEE International Conference on Phased Array Systems and Technology*, pp. 505–508, May 2000.

[31] A. Miura, S. Yamamoto, N. Obara, H. Saito, T. Takahashi, H. Wakana, and M. Tanaka, "Development of a Ka-band Active Phased Array Antenna for Mobile SATCOM Stations," in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 814–818, 1999.

[32] A. I. Zaghloul and O. Kilie, "Use of Active Phased Arrays for Multiple-Beam Cellular Communications Systems," in *Proceedings of the National Radio Science Conference*, (Cairo, Egypt), pp. 1–12, 2002.

[33] A. Dreher, N. Niklasch, F. Klefenz, and A. Schroth, “Antenna and Receiver System with Digital Beamforming for Satellite Navigation and Communications,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 51, pp. 1815–1821, July 2003.

[34] L. C. Stange, H. Pawlak, A. Dreher, S. Holzwarth, A. F. Jacob, O. Litschke, and M. Theil, “Components of a Highly Integrated DBF Terminal Antenna for Mobile Ka-Band Satellite Communications,” in *IEEE MTT-S International Microwave Symposium Digest*, pp. 583–586, 2003.

[35] R. Miura, T. Tanaka, I. Chiba, A. Horie, and Y. Karasawa, “Beamforming Experiment with a DBF Multibeam Antenna in a Mobile Satellite Environment,” *IEEE Transactions on Antennas and Propagation*, vol. 45, pp. 707–714, Apr. 1997.

[36] E. Brookner, “Phased Arrays for the New Millennium,” in *Proceedings of the IEEE International Conference on Phased Array Systems and Technology*, pp. 3–19, 2000.

[37] D. H. Martin, *Communication Satellites.* The Aerospace Press, 2000. Fourth Edition.

[38] B. Barani, “Satellite Communications in the European Union R & D Programmes: An Overview,” in *Proceedings of the IEE Colloquium on EU’s Initiatives in Satellite Communications*, pp. 1/1–1/8, May 1997.

[39] A. Grami and K. Gordon, “Next-Generation Ka-Band Satellite Concept To Extend the Reach of Canada’s Broadband Infrastructure,” in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, vol. 4, pp. 2754–2758, 2001.

[40] J. E. Allnutt and F. Haidara, “Ku–band Diurnal Fade characteristics and Fade Event Duration Data From Three, Two-year, Earth–space Radiometric Experiments in Equatorial Africa,” *International Journal of Satellite Communications*, vol. 18, pp. 161–183, May/June 2000.

[41] I. Minei and R. Cohen, “High-Speed Internet Access Through Unidirectional Geostationary Satellite Channels,” *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 345–359, Feb. 1999.

[42] J. S. Baras, S. Corson, S. Papademetriou, I. Secka, and N. Suphasindhu, “Fast Asymmetric Internet over Wireless Satellite-Terrestrial Networks,” in *Proceedings of the IEEE Military Communications Conference (MILCOM)*, pp. 372–377, 1997.

[43] M. Allman, C. Hayes, and S. Ostermann, “An Evaluation of TCP with Larger Initial Windows,” *ACM Computer Communication Review*, vol. 28, pp. 41–52, July 1998.

[44] C. Partridge and T. J. Shepard, “TCP/IP Performance over Satellite Links,” *IEEE Network*, vol. 11, pp. 44–49, Sep/Oct 1997.

[45] D. Boudreau, G. Caire, G. E. Corazza, R. D. Gaudenzi, G. Gallinaro, M. Luglio, R. Lyons, J. Romero-García, A. Vernucci, and H. Widmer, "Wide-Band CDMA for the UMTS/IMT-2000 Satellite Component," *IEEE Transactions on Vehicular Technology*, vol. 51, pp. 306–331, Mar. 2002.

[46] D. I. Kim, E. Hossain, and V. K. Bhargava, "Downlink Joint Rate and Power Allocation in Cellular WCDMA Systems," *IEEE Transactions on Wireless Communications*, vol. 2, pp. 69–80, Jan. 2003.

[47] R. Vannithamby and E. S. Sousa, "An Optimum Rate/Power Allocation Scheme for Downlink in Hybrid CDMA/TDMA Cellular System," in *Proceedings of the IEEE Vehicular Technology Conference*, vol. 4, pp. 1734–1738, 2000.

[48] A. Bedekar, S. Borst, K. Ramanan, P. Whiting, and E. Yeh, "Downlink Scheduling in CDMA Data Networks," in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, pp. 2653–2657, 1999.

[49] V. A. Siris, "Resource Control for Elastic Traffic in CDMA Networks," in *Proceedings of the annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, (Atlanta, Georgia, USA), pp. 193–204, 2002.

[50] N. Joshi, S. R. Kadaba, S. Patel, and G. S. Sundaram, "Downlink Scheduling in CDMA Data Networks," in *Proceedings of the annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, pp. 179–190, 2000.

[51] A. K. Yesufu, "Possible use of Satellites in Rural Telecommunications in Africa," in *IEE Second International Conference on Rural Telecommunications*, pp. 156–159, 1990.

[52] B. Franklin, "Progress Report on Regional Interconnection by the Commonwealth Telecommunications Organisation (CTO)," in *Proceedings of the* 12^{th} *East African Regulatory Postal Telecommunications Organisations (EARPTO) Meeting*, (Nairobi, Kenya), 2003.

[53] G. Tedros, "The Global Communication Needs of the New Nations of Africa," *IEEE Transactions on Communications*, vol. 12, pp. 34–37, Sept. 1964.

[54] P. O. Okundi, "Pan–African Telecommunication Network: A Case for Telecommunications in the Development of Africa," *IEEE Transactions on Communications*, vol. 24, pp. 749–755, July 1976.

[55] A. O. Taylor, "Appropriate Telecommunication Technology for Rural Africa," in *IEE International Conference on Rural Telecommunications*, pp. 7–10, 1988.

[56] N. Nageshar, R. Sewsunker, and S. H. Mneney, "Economic Analysis of a Satellite–terrestrial Telephony System for Rural SADC," in *IEEE Africon Conference in Africa (AFRICON)*, pp. 401–406, 2002.

[57] Uganda Communications Commission (UCC), "Bridging the Digital Divide - Uganda's Experience," in *Proceedings of the* 12^{th} *East African Regulatory Postal Telecommunications Organisations (EARPTO) Meeting*, (Nairobi, Kenya), 2003.

[58] Uganda Communications Commission (UCC), *Rural Communications Development Policy for Uganda*. Uganda Communications Commission, 2001.

[59] NEPAD Secretariat, "Bringing the benefits of ICT to Africa," *The Weekly electronic Newsletter of the NEPAD Secretariat*, vol. 6, June 2003.

[60] ALCATEL, "Alcatel Space Wins a US Dollar 150 million Turnkey Contract to Build and Deliver in Orbit the First Pan-African RASCOM Telecommunication Satellite," in *Alcatel Press Release*, June 2003.

[61] A. K. Parekh and R. G. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case," *IEEE/ACM Transactions on Networking*, pp. 344–357, June 1993.

[62] R. Roy, R. C. Mudumbai, and S. S. Panwar, "Analysis of TCP Congestion Control using a Fluid Model," in *Proceedings of the IEEE International Conference on Communications (ICC)*, vol. 8, pp. 2396–2403, 2001.

[63] J. P. Cances, G. Maral, and B. Coulomb, "Coverage Reconfiguration for Dynamically Allocating Channels to Beams in a Multibeam Satellite System," in 15^{th} *AIAA International Communications Satellite Systems Conference*, pp. 1032–1041, 1994.

[64] A. D. Cliff and J. K. Ord, *Spatial Processes. Models & Applications.* Pion Limited, 1981.

[65] H. D. Sherali, C. M. Pendyala, and T. S. Rappaport, "Optimal Location of Transmitters for Micro-Cellular Radio Communication System Design," *IEEE Journal on Selected Areas in Communications*, pp. 662–673, May 1996.

[66] R. Jäntti and S.-L. Kim, "Transmission Rate Scheduling for the Non-real-time Data in a Cellular CDMA System," *IEEE Communications Letters*, pp. 200–202, May 2001.

[67] R. A. Berry, *Power and Delay Trade–offs in Fading Channels.* PhD thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, June 2000.

[68] B. E. Collins and R. L. Cruz, “Transmission Policies for Time Varying Channels with Average Delay Constraints,” in *Proceedings of the Allerton Conference on Communication, Control and Computing*, (Monticello, IL), pp. 1–9, 1999.

[69] N. Bambos and S. Kandukuri, “Power Controlled Multiple Access (PCMA) in Wireless Communication Networks,” in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, pp. 386–395, 2000.

[70] K. S. Rao, G. A. Morin, M. Q. Tang, S. Richard, and K. K. Chan, “Development of a 45 GHz Multiple-Beam Antenna for Military Satellite Communications,” *IEEE Transactions on Antennas and Propagation*, pp. 1036–1047, Oct. 1995.

[71] S. Egami, “A Power-Sharing Multiple-Beam Mobile Satellite in Ka Band,” *IEEE Journal on Selected Areas in Communications*, pp. 145–152, Feb. 1999.

[72] G. Arnold and T.-W. Kao, “Rain Attenuation in EHF Satellite Communications,” in *Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, pp. 153–156, 1995.

[73] N. Blefari-Melazzi and G. Reali, “Improving the Efficiency of Circuit-Switched Satellite Networks by Means of Dynamic Bandwidth Allocation Capabilities,” *IEEE Journal on Selected Areas in Communications*, pp. 2373–2384, Nov. 2000.

[74] A. Iera, A. Molinaro, S. Marano, and M. Petrone, “QoS for Multimedia Applications in Satellite Systems,” *IEEE Multimedia*, pp. 46–53, Oct. 1999.

[75] N. Blefari-Melazzi and G. Reali, “A Resource Management Scheme for Satellite Networks,” *IEEE Multimedia*, pp. 54–63, Oct. 1999.

[76] A. Iera, A. Molinaro, and S. Marano, “Call Admission Control and Resource Management Issues for Real-Time VBR Traffic in ATM-Satellite Networks,” *IEEE Journal on Selected Areas in Communications*, pp. 2393–2403, Nov. 2000.

[77] A. Iera and A. Molinaro, “Designing the Interworking of Terrestrial and Satellite IP-Based Networks,” *IEEE Communications Magazine*, vol. 40, pp. 136–144, Feb. 2002.

[78] H. Biscéré and M. Terré, “On the Choice of TDMA or CDMA for a Multimedia Satellite System,” in *Proceedings of the IEEE International Conference on Universal Personal Communications (ICUPC)*, vol. 1, pp. 641–645, 1998.

[79] R. Di Girolamo and T. Le-Ngoc, “Multi-code TDMA (MC-TDMA) for Multimedia Satellite Communications,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, vol. 2, pp. 684–688, 1997.

[80] I. Forkel, B. Wegmann, and E. Schulz, "On the Capacity of a UTRA-TDD Network with Multiple Services," in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 585–589, 2002.

[81] P. Mermelstein, A. Jalali, and H. Leib, "Integrated Services on Wireless Multiple Access Networks," in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 863–867, 1993.

[82] E. Cianca, F. Graziosi, and F. Santucci, "An Approach to Maximize the Capacity of a Multimedia CDMA Wireless System," in *Proceedings of the IEEE Vehicular Technology Conference*, vol. 2, pp. 909–913, 1998.

[83] D.-H. Kim, S.-H. Hwang, U.-Y. Pak, and K.-C. Whang, "Adaptive CDMA Scheme as a Rain Fade Countermeasure in Ka-Band Geosynchronous Satellite Communications," *IEICE Transactions on Communications*, pp. 2600–2606, Dec. 2000.

[84] X. Wu, L.-L. Yang, and L. Hanzo, "Uplink Capacity Investigations of TDD/CDMA," in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 997–1001, 2002.

[85] H. Yomo, A. Nakata, and S. Hara, "An Efficient Slot Allocation Algorithm to Accommodate Multimedia Traffic in CDMA/TDD-Based Wireless Communication Systems," in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 787–791, 2001.

[86] A. Ibrahim and S. Tohme, "CDMA/PRMA Analytical Model for Voice Users in Satellite-UMTS Systems," in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 3027–3032, 2002.

[87] G. J. R. Povey and M. Nakagawa, "A Review of Time Division Duplex - CDMA Techniques," in *Proceedings of the IEEE International Symposium on Spread Spectrum Techniques and Applications*, pp. 630–633, 1998.

[88] H. Honkasalo, K. Pehkonen, M. T. Niemi, and A. T. Leino, "WCDMA and WLAN for 3G and Beyond," *IEEE Wireless Communications*, vol. 9, pp. 14–18, Apr. 2002.

[89] F. Fitzek, A. Köpsel, A. Wolisz, M. Krishnam, and M. Reisslein, "Providing Application-Level QoS in 3G/4G Wireless Systems: A Comprehensive Framework based on Multirate CDMA," *IEEE Wireless Communications*, vol. 9, pp. 42–47, Apr. 2002.

[90] L. B. Milstein, "Wideband Code Division Multiple Access," *IEEE Journal on Selected Areas in Communications*, vol. 18, pp. 1344–1354, Aug. 2000.

[91] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjö, M. Persson, and C. Roobol, "WCDMA - The Radio Interface for Future Mobile Multimedia Communications," *IEEE Transactions on Vehicular Technology*, vol. 47, pp. 1105–1118, Nov. 1998.

[92] A. Sampath and J. M. Holtzman, "Access Control of Data in Integrated Voice/Data CDMA Systems: Benefits and Tradeoffs," *IEEE Journal on Selected Areas in Communications*, pp. 1511–1526, Oct. 1997.

[93] H. Koraitim and S. Tohmé, "GRAP: A Multiple Access Protocol for Packet Satellite Networks," in *IEEE International Conference on Communications*, pp. 196–202, 1999.

[94] D. P. Gerakoulis, W.-C. Chan, and E. Geraniotis, "Throughput Evaluation of a Satellite-Switched CDMA (SS/CDMA) Demand Assignment System," *IEEE Journal on Selected Areas in Communications*, pp. 286–302, Feb. 1999.

[95] A. Baiocchi, N. Blefari-Melazzi, M. Listanti, and C. Soprano, "Definition and Performance Analysis of a Simple, ABR-Like Congestion Control Scheme for Satellite ATM Networks with Guaranteed Loss Performance," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 303–313, Feb. 1999.

[96] B. R. Elbert, *The Satellite Communication Applications Handbook*. Artech House, 1997.

[97] W. J. Vogel and J. Goldhirsh, "Multipath Fading at L Band for Low Elevation Angle, Land Mobile Satellite Scenarios," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 197–204, Feb. 1995.

[98] E. Geraniotis and D. Gerakoulis, "Bit Error Rate Evaluation of a Spectrally Efficient CDMA Scheme for Geostationary Satellite Communications," in *Proceedings of the IEEE Fifth International Symposium on Computers and Communications*, pp. 254–258, 2000.

[99] G. Butt, B. G. Evans, and M. Parks, "Modelling the Mobile Satellite Channel for Communication System Design," in *Proceedings of the Ninth International Conference on Antennas and Propagation*, vol. 2, pp. 387–394, 1995.

[100] S. Yao and E. Geraniotis, "On the Power Control of a Multiple-Beam Mobile Satellite CDMA System," in *IEEE Military Communications Conference*, vol. 2, pp. 523–528, 1995.

[101] A. M. Monk and L. B. Milstein, "Open-Loop Power Control Error in a Land Mobile Satellite System," *IEEE Journal on Selected Areas in Communications*, pp. 205–212, Feb. 1995.

[102] E. Matricciani, “Transformation of Rain Attenuation Statistics from Fixed to Mobile Satellite Communication Systems,” *IEEE Transactions on Vehicular Technology*, pp. 565–569, Aug. 1995.

[103] E. Matricciani and S. Moretti, “Rain Attenuation Statistics for the Design of Mobile Satellite Communication Systems,” *IEEE Transactions on Vehicular Technology*, pp. 637–648, May 1998.

[104] H. Helmken and R. Henning, “Satellite Ka-Band Propagation Measurements in Florida,” in *Proceedings of the Fourth International Mobile Satellite Conference*, pp. 140–144, 1995. Session 4.

[105] H. Helmken, R. E. Henning, J. Feil, L. J. Ippolito, and C. E. Mayer, “A Three-site Comparison of Fade-duration Measurements,” *Proceedings of the IEEE*, pp. 917–925, June 1997.

[106] L. J. Ippolito Jr., *Radiowave Propagation in Satellite Communications*. Van Nostrand Reinhold Company Inc., 1986.

[107] D. G. Sweeney and C. W. Bostian, “Implementing Adaptive Power Control as a 30/20-GHz Fade Countermeasure,” *IEEE Transactions on Antennas and Propagation*, vol. 47, pp. 40–46, Jan. 1999.

[108] M. Luglio, “Fade Countermeasures in Ka Band: Application of Frequency Diversity to a Satellite System,” in *Tenth International Conference on Digital Satellite Communications*, vol. 1, pp. 143–151, 1995.

[109] M. Stojanovic and V. Chan, “Adaptive Power and Rate Control for Satellite Communications in Ka Band,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, vol. 5, pp. 2967–2972, 2002.

[110] A. Paraboni, C. Capsoni, G. Masini, J. P. V. Poiares Baptista, and C. Riva, “Dynamic Fade Restoration in Ka-band Satellite Systems,” *International Journal of Satellite Communications*, vol. 20, pp. 283–291, July/August 2002.

[111] M. Luglio, R. Mancini, C. Riva, A. Paraboni, and F. Barbaliscia, “Large-scale Site Diversity for Satellite Communication Networks,” *International Journal of Satellite Communications*, vol. 20, pp. 251–260, July/August 2002.

[112] A. F. Ismail and P. A. Watson, “Characteristics of Fading and Fade Countermeasures on a Satellite-Earth Link Operating in an Equatorial Climate, with Reference to Broadcast Applications,” *IEE Proceedings on Microwaves, Antennas and Propagation*, vol. 147, pp. 369–373, Oct. 2000.

[113] Y. Zhou, F. Chin, Y.-C. Liang, and C.-C. Ko, "Performance Comparison of Transmit Diversity and Beamforming for the Downlink of DS-CDMA System," *IEEE Transactions on Wireless Communications*, vol. 2, pp. 320–334, Mar. 2003.

[114] P. J. Garland, P. Takats, and T. Le-Ngoc, "Access protocols and On-board switch techniques for Multimedia Personal Satellite Communications," in *Tenth International Conference on Digital Satellite Communications*, vol. 1, pp. 63–71, 1995.

[115] H.-A. M. Mourad, "A Generalized Gaussian User Distribution and A Modified Traffic Power Control (MTPC) for LEO Systems," in *Proceedings of the National Radio Science Conference*, (Cairo, Egypt), pp. 415–421, 2001.

[116] G. Maral and M. Bousquet, *Satellite Communications Systems. Systems, Techniques and Technology.* John Wiley & Sons, 1993. Translated by J. C. C. Nelson. Second Edition.

[117] A. I. Zaghloul, Y. Hwang, R. M. Sorbello, and F. T. Assal, "Advances in Multibeam Communications Satellite Antennas," *Proceedings of the IEEE*, vol. 78, pp. 1214–1232, July 1990.

[118] W. L. Pritchard, H. G. Suyderhoud, and R. A. Nelson, *Satellite Communication Systems Engineering.* Prentice-Hall, 1993. Second Edition.

[119] S. Egami, "Satellite Link Requirements in Personal Satellite Communications," *Space Communications*, pp. 105–114, Dec. 1994.

[120] D. Roddy, *Satellite Communications.* McGraw-Hill, 1996. Second Edition.

[121] A. M. Vernon, M. A. Beach, and J. P. McGeehan, "Planning and Optimisation of Smart Antenna Base Stations in 3G Networks," in *Proceedings of the IEEE Colloquium on Capacity and Range Enhancement Techniques for the Third Generation Mobile Communications and Beyond*, vol. 2000/003, pp. 1/1–1/7, 2000.

[122] J. Laiho and A. Wacker, "Radio Network Planning Process and Methods for WCDMA," *Annals of Telecommunications*, vol. 56, pp. 317–331, May/June 2001.

[123] G. Losquadro, "SECOMS: Advanced Interactive Multimedia Satellite Communications for a Variety of Compact Terminals," in *Colloquium on EU's Initiatives in Satellite Communications - Mobile*, vol. 3, (Savoy Place, London), pp. 3/1–3/7, 1997.

[124] C. Valadon, P. Taaghol, B. G. Evans, and R. Tafazolli, "Link Design and Dimensioning of CDMA, the Alternative Multiple Access in SECOMS," in *Proceedings of the* 3^{rd} *Ka-band Utilization Conference*, pp. 503–510, 1997.

[125] D. Stamatelos and A. Ephremides, “Spectral Efficiency and Optimal Base Placement for Indoor Wireless Networks,” *IEEE Journal on Selected Areas in Communications*, pp. 651–661, May 1996.

[126] S.-T. Yang and A. Ephremides, “Optimal Network Design: the Base Station Placement Problem,” in *Proceedings of the IEEE* 36^{th} *Conference on Decision and Control*, pp. 2381–2386, Dec. 1997.

[127] E. Cianca, S. D. Fina, R. Lojacono, M. Ruggieri, and R. Prasad, “Downlink Capacity Analysis for DS-CDMA Satellite Systems Accounting for On-Board Power Constraints,” in *Proceedings of the IEEE Sixth International Symposium on Spread Spectrum Techniques and Applications*, vol. 1, pp. 319–324, 2000.

[128] Y. Birk and Y. Keren, “Judicious Use of Redundant Transmissions in Multichannel ALOHA Networks with Deadlines,” *IEEE Journal on Selected Areas in Communications*, pp. 257–269, Feb. 1999.

[129] D. Ayyagari and A. Ephremides, “Optimal Admission Control in Cellular DS-CDMA Systems With Multimedia Traffic,” *IEEE Transactions on Wireless Communications*, vol. 2, pp. 195–202, Jan. 2003.

[130] S. A. Jafar and A. Goldsmith, “Adaptive Multirate CDMA for Uplink Throughput Maximization,” *IEEE Transactions on Wireless Communications*, vol. 2, pp. 218–228, Mar. 2003.

[131] L. Qiang, G. Jichang, T. Jianfu, and Z. Quanmin, “A Review of Power Control in Cellular Mobile Communication Systems,” in *Proceedings of the International Symposium on Electromagnetic Compatibility*, pp. 675–680, 2002.

[132] S. A. Grandhi, R. Vijayan, D. J. Goodman, and J. Zander, “Centralized Power Control in Cellular Radio Systems,” *IEEE Transactions on Vehicular Technology*, vol. 42, pp. 466–468, Nov. 1993.

[133] Q. Wu, W.-L. Wu, and J.-P. Zhou, “Centralized Power Control in CDMA Cellular Mobile Systems,” in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 1268–1271, 1997.

[134] L. Lv, S. Zhu, and S. Dong, “Fast Convergence Distributed Power Control Algorithm for WCDMA Systems,” *IEE Proceedings on Communications*, vol. 150, pp. 134–140, Apr. 2003.

[135] D. Kim, K.-N. Chang, and S. Kim, “Efficient Distributed Power Control for Cellular Mobile Systems,” *IEEE Transactions on Vehicular Technology*, vol. 46, pp. 313–319, May 1997.

[136] C. U. Saraydar, N. B. Mandayam, and D. J. Goodman, "Efficient Power Control via Pricing in Wireless Data Networks," *IEEE Transactions on Communications*, vol. 50, pp. 291–303, Feb. 2002.

[137] J. W. Lee, R. R. Mazumdar, and N. B. Shroff, "Downlink Power Allocation for Multi-class CDMA Wireless Networks," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, pp. 1480–1489, 2002.

[138] C. W. Sung and W. S. Wong, "A Noncooperative Power Control Game for Multirate CDMA Data Networks," *IEEE Transactions on Wireless Communications*, vol. 2, pp. 186–194, Jan. 2003.

[139] X. Wenhu and G. Zhongmin, "Performance Comparison of Distributed Power Control Algorithms in Cellular Radio Systems," in *Proceedings of International Conference on Communication Technology*, pp. 977–980, 2003.

[140] S. V. Hanly, "An Algorithm for Combined Cell-Site Selection and Power Control to Maximize Cellular Spread Spectrum Capacity," *IEEE Journal on Selected Areas in Communications*, vol. 13, pp. 1332–1340, Sept. 1995.

[141] B. Gremont, M. Filip, P. Gallios, and S. Bate, "Comparative Analysis and Performance of Two Predictive Fade Detection Schemes for Ka-Band Fade Countermeasures," *IEEE Journal on Selected Areas in Communications*, pp. 180–192, Feb. 1999.

[142] A. W. Dissanayake, "Application of Open-Loop Uplink Power Control in Ka-Band Satellite Links," *Proceedings of the IEEE*, vol. 85, pp. 959–969, June 1997.

[143] A. Sampath, N. B. Mandayam, and J. M. Holtzman, "Erlang Capacity of a Power Controlled Integrated Voice and Data CDMA System," *Proceedings of the IEEE*, pp. 1557–1561, 1997.

[144] S. Ramakrishna and J. M. Holtzman, "A Scheme for Throughput Maximization in a Dual-Class CDMA System," *IEEE Journal on Selected Areas in Communications*, pp. 830–844, Aug. 1998.

[145] T.-K. Liu and J. A. Silvester, "Joint Admission/Congestion Control for Wireless CDMA Systems Supporting Integrated Services," *IEEE Journal on Selected Areas in Communications*, pp. 845–857, Aug. 1998.

[146] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power Control and Resource Management for a Multimedia CDMA Wireless System," in *IEEE Personal, Indoor and Mobile Radio Communications*, pp. 21–25, 1995.

[147] D. Kim, “Rate-Regulated Power Control for Supporting Flexible Transmission in Future CDMA Mobile Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 968–977, May 1999.

[148] S.-J. Oh and K. M. Wasserman, “Optimality of Greedy Power Control and Variable Spreading Gain in Multi-class CDMA Mobile Networks,” in *Proceedings of the annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, (Seattle, Washington, United States), pp. 102–112, Aug. 1999.

[149] P. J. M. Havinga and G. J. M. Smit, “QoS Scheduling for Energy-Efficient Wireless Communication,” in *Proceedings of the IEEE International Conference on Information Technology: Coding and Computing*, pp. 167–171, 2001.

[150] D. Goodman and N. Mandayam, “Network Assisted Power Control for Wireless Data,” in *Proceedings of the IEEE Vehicular Technology Conference*, vol. 2, pp. 1022–1026, 2001.

[151] F. Berggren, S.-L. Kim, R. Jäntti, and J. Zander, “Joint Power Control and Intracell Scheduling of DS-CDMA Nonreal Time Data,” *IEEE Journal on Selected Areas in Communications*, vol. 19, pp. 1860–1870, Oct. 2001.

[152] C. W. Sung and W. S. Wong, “Power Control and Rate Management for Wireless Multimedia CDMA Systems,” *IEEE Transactions on Communications*, pp. 1215–1226, July 2001.

[153] L. Song and N. B. Mandayam, “Hierarchical SIR and Rate Control for CDMA Data Users on the Forward Link,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, vol. 3, pp. 1514–1518, 2000.

[154] M. Elaoud and P. Ramanathan, “Adaptive Allocation of CDMA Resources for Network-level QoS Assurances,” in *Proceedings of the annual ACM/IEEE International Conference on Mobile Computing and Networking (MobiCom)*, pp. 191–199, 2000.

[155] F. Berggren and R. Jäntti, “Asymptotically Fair Scheduling on Fading Channels,” in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 1934–1938, 2002.

[156] A. B. MacKenzie and S. B. Wicker, “Game Theory in Communications: Motivation, Explanation, and Application to Power Control,” in *Proceedings of the IEEE Global Communications Conference (GLOBECOM)*, pp. 821–826, 2001.

[157] D. A. Hayes, M. Rumsewicz, and L. Andrew, “Quality of Service Driven Packet Scheduling Disciplines for Real-time applications: Looking Beyond Fairness,” in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, vol. 1, pp. 405–412, 1999.

[158] S. Lu, V. Bharghavan, and R. Srikant, "Fair Scheduling in Wireless Packet Networks," *IEEE/ACM Transactions on Networking*, pp. 473–489, Aug. 1999.

[159] V. Bharghavan, S. Lu, and T. Nandagopal, "Fair Queueing in Wireless Networks: Issues and Approaches," *IEEE Personal Communications*, pp. 44–53, Feb. 1999.

[160] T. S. Eugene-Ng, I. Stoica, and H. Zhang, "Packet Fair Queueing Algorithms for Wireless Networks with Location-Dependent Errors," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, pp. 1103–1111, 1998.

[161] P. Lin, B. Bensaou, Q. L. Ding, and K. C. Chua, "CS-WFQ: A Wireless Fair Scheduling Algorithm for Error-Prone Wireless Channels," in *Proceedings of the IEEE Ninth International Conference on Computer Communications and Networks*, pp. 276–281, 2000.

[162] C. Fragouli, V. Sivaraman, and M. B. Srivastava, "Controlled Multimedia Wireless Link Sharing via Enhanced Class-Based Queueing with Channel-State-Dependent Packet Scheduling," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, vol. 2, pp. 572–580, 1998.

[163] Y. Yi, Y. Seok, T. Kwon, Y. Choi, and J. Park, "w^2f^2q: Packet Fair Queueing in Wireless Packet Networks," in *Proceedings of the ACM International Workshop on Wireless Mobile Multimedia (WOWMOM)*, pp. 2–10, 2000.

[164] S. Lu, T. Nandagopal, and V. Bharghavan, "Design and Analysis of an Algorithm for Fair Service in Error-Prone Wireless Channels," *ACM/Baltzer Wireless Networks Journal*, vol. 6(4), pp. 323–343, Aug. 2000. (Invited Paper).

[165] L. Xu, X. Shen, and J. W. Mark, "Dynamic Bandwidth Allocation with Fair Scheduling for WCDMA Systems," *IEEE Wireless Communications*, vol. 9, pp. 26–32, Apr. 2002.

[166] R. M. Sheldon, *Introduction to Stochastic Dynamic Programming.* Academic Press, Inc., 1983.

[167] R. A. Berry and R. G. Gallager, "Communication Over Fading Channels With Delay Constraints," *IEEE Transactions on Information Theory*, vol. 48, pp. 1135–1149, May 2002.

[168] D. Zhang and K. M. Wasserman, "Transmission Schemes for Time-varying Wireless Channels with Partial State Observations," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, pp. 467–476, 2002.

[169] R. B. Dybdal, “Adaptive Control of Multiple Beam Satellite Transponders,” in *Proceedings of the IEEE Military Communications Conference (MILCOM)*, pp. 252–255, 1997.

[170] W. H. Theunissen and W. D. Burnside, “Contoured Beam Reflector Antenna for Wireless Applications,” *IEEE Transactions on Antennas and Propagation*, vol. 50, pp. 205–210, Feb. 2002.

[171] M. S. Bazaraa, H. D. Sherali, and C. M. Shetty, *Nonlinear Programming: Theory and Algorithms*. John Wiley & Sons, New York, 1993. Second Edition.

[172] S.-J. Oh, D. Zhang, and K. M. Wasserman, “Optimal Resource Allocation in Multi-service CDMA Networks,” *IEEE Transactions on Wireless Communications*, vol. 2, pp. 811–821, July 2003.

[173] W. Choi, B. S. Kang, J. C. Lee, and K. T. Lee, “Forward link Erlang capacity of 3G CDMA system,” in *Proceedings of First International Conference on 3G Mobile Communication Technologies*, pp. 213–217, 2000. IEE Conf. Publ. No. 471.

[174] J. P. Choi and V. W. S. Chan, “Adaptive Communications over Fading Satellite Channels,” in *Proceedings of the IEEE International Conference on Communications (ICC)*, pp. 2635–2639, 2001.

[175] S. Yao and E. Geraniotis, “Optimal Power Control Law for Multi-Media Multi-Rate CDMA Systems,” in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 392–396, 1996.

[176] M.-H. Chung and K.-C. Chen, “Power Allocation for Multi-Rate Multiuser Detection in Wideband CDMA Systems,” in *Proceedings of the IEEE Vehicular Technology Conference*, vol. 1, pp. 608–612, 1999.

[177] P.-W. Fu and K.-C. Chen, “Multi-Rate MC-DS-CDMA with Multiuser Detections for Wireless Multimedia Communications,” in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 1536–1540, 2002.

[178] A. Yener, R. D. Yates, and S. Ulukus, “Interference Management for CDMA Systems Through Power Control, Multiuser Detection, and Beamforming,” *IEEE Transactions on Communications*, vol. 49, pp. 1227–1239, July 2001.

[179] S. Jordan and P. P. Varaiya, “Throughput in Multiple Service, Multiple Resource Communication Networks,” *IEEE Transactions on Communications*, vol. 39, pp. 1216–1222, Aug. 1991.

[180] V. K. N. Lau and S. V. Maric, “Variable Rate Adaptive Modulation for DS-CDMA,” *IEEE Transactions on Communications*, vol. 49, pp. 1227–1239, July 2001.

[181] T. Inoue, S. Sampei, and N. Morinaga, "DS-CDMA and Adaptive Modulation Based TDMA Dual Mode Scheme for High Speed Data Transmission Service in Wireless Multimedia Communication Systems," in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 585–589, 2001.

[182] S. Abeta, H. Atarashi, and M. Sawahashi, "Broadband Packet Wireless Access Incorporating High-Speed IP Packet Transmission," in *Proceedings of the IEEE Vehicular Technology Conference*, pp. 844–848, 2002.

[183] Q. Liu, E.-H. Yang, and Z. Zhang, "Throughput Analysis of CDMA Systems Using Multiuser Receivers," *IEEE Transactions on Communications*, vol. 49, pp. 1192–1202, July 2001.

[184] Sheldon M Ross, *Introduction to Probability Models*. Academic Press, 2003. Eigth Edition.

[185] Richard M Feldman and Ciriaco Valdez-Flores, *Applied Probability and Stochastic Processes*. PWS Publishing Company, 1996.

[186] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*. Academic Press,Inc., 1976.

[187] R. Vannithamby and E. S. Sousa, "Resource Allocation and Scheduling Schemes for WCDMA Downlinks," in *Proceedings of the IEEE International Conference on Communications (ICC)*, vol. 5, pp. 1406–1410, 2001.

[188] M. J. Neely, E. Modiano, and C. E. Rohrs, "Power and Server Allocation in a Multi-Beam Satellite with Time Varying Channels," in *Proceedings of the IEEE Conference on Computer Communications (INFOCOM)*, vol. 3, pp. 1451–1460, 2002.

[189] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*. Addison-Wesley Publishing Company, 1994. Second Edition.

[190] A. Jamalipour and A. Ogawa, "Packet Admission Control in a Direct-Sequence Spread-Spectrum LEO Satellite Communications Network," *IEEE Journal on Selected Areas in Communications*, vol. 15, pp. 1649–1656, Oct. 1997.