IMPROVING 3G NETWORK THROUGHPUT BY NEW SERVICE AND JOINT DESIGN

BY

LI NING

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF PHILOSOPHY
IN
INFORMATION ENGINEERING

JULY 2004

Acknowledgments

I would like to express my deep and sincere gratitude to my supervisor, Professor Peter Tak-shing Yum, who has commented extensively and intensively on the contents of this thesis. His understanding and guidance have provided a good basis for the present thesis.

During the two years, I meet some friends in Chinese University of Hong Kong. We enjoy hiking. We have set up a BBS server for the students who come from mainland and are currently in CUHK. I wish to extend my warmest thanks to them who have brought me colorful life: Li Xiaoqi, Chen Xinyu, Li Qinghua, Huang Kaizhu, Meng Wei, Xue Hong, He Xinping, Xiao Lurong, and Wang Guangyu, and to the members in the Communication Technology Laboratory who have offered their help to my work: Chan Kington, Hua Cunqing, Zhang Xinyan, Zhang Lin, and Zhang Li.

I owe my loving thanks to my family and friends, for all of their encouragement and caring throughout the years.

Abstract

The challenge of third-generation wireless network is to provide multimedia service with different levels of quality of service (QoS), based on users requirements and their link budgets. We address two issues to provide end users with the necessary QoS.

The first is to support a new service, called delayable service, in wireless network in concurrence with traditional services, such as real-time service and best-effort service. In the new service, customers specify the extent of delay that is tolerable at the time of request so that the system can schedule the transmission according to the on-going traffic. It gives flexibility to both customers and system operators. Considering the bursty nature of real-time traffic, we proposed to rearrange the accepted delayable packets to release similar number of channels in the following slots. More real-time connections can be admitted consequently.

The second is to design transmission schedule that meets the target QoS requirements, given the channel fading information. Heuristic algorithms are proposed to reduce the computational complex of the optimal scheduling protocols. Users are divided into groups based on the idea that users in the same group experience similar channel conditions and adopt the same transmission schemes. With the assumption that just one data user is admitted for transmission in each slot, we select the user based on channel conditions in PHY layer and queue information in MAC layer jointly. The user is allocated with the maximum available capacity.

通過新服務類型與聯合優化方法
提高第三代無線網路的系統吞吐量

李寧

香港中文大學資訊工程學系

哲學碩士論文摘要

基於每個用戶各自的需求及所在網路環境，第三代無線網路的一個挑戰性目標在於為不同的用戶提供具有不同服務品質的多媒體資訊。本論文將著重討論其中兩項技術，從而為用戶提供所需要的服務品質。

首先，在支援一些傳統服務類型的基礎上，如實時服務、盡力服務，我們將可延遲服務引入到無線網路中。使用這種服務類型的用戶可以在申請服務的同時指明他們可以接受的延時，這樣系統可以根據網路流量的狀況安排傳輸，這給用戶和系統運營商都帶來了方便。考慮到即時服務流量的突發性，我們提出重新安排已接受的可延遲包，以便在隨後的時隙中釋放出同樣的通道，從而接受更多的即時服務請求。

其次，為了實現需要的服務品質，並提高通道的使用效率，我們聯合優化各種傳輸配置，並提出一些啟發性的演算法以降低優化的計算複雜度。基於用戶的通道品質，我們將用戶分成若干組，我們認為處於同一組的用戶經歷相似的通道品質，並採用相同的傳輸配置。在假定每個時隙只有一個資料用戶被允許傳輸的情況下，我們基於這些資料用戶在物理層的通道品質和他們在接入控制層的佇列資訊，選擇其中的一個用戶安排傳輸，這個用戶被分配可能的最大通道容量。

LIST OF CONTENTS

LIST OF FIGURES

Chapter 1 Introduction

Wireless traffic continues to increase at a steady rate. For example, there are 1.3 billion cellular subscribers worldwide by the end of 2003, and 1.7 billion expected by the end of 2007 [1]. On March 31^{st} 2004, NTT DoDoMo announced the number of subscribers to its 3G service has reached three million, just two months after hitting the two million mark [1]. Such a trend will continue as wireless services shift from voice to packet data, and users become more accustomed to conducting wireless business and multimedia applications. This course of evolution will be similar to what happened to the Internet: from a limited application environment to an integral part of the average person's life. The future of wireless networks is not just in voice, but also in the integration of voice, data, and multimedia. Multimedia applications will be supported with diverse quality-of-service (QoS) requirements, accommodating the needs of individual customers. Because the wireless access channels quality varies with time, there is a need to develop resource management and control schemes for wireless channels that provide QoS guarantees for heterogeneous traffic. The QoS control schemes also need to be simple to implement and manage.

In this thesis, we address two issues for 3G networks, new service and joint design. The service called delayable service is introduced in wireless networks, which was first proposed by Yum and Chen [2] in cable networks. In delayable service, customers specify the extent of delay that is tolerable at the time of request so that the system can transport the traffic to the customer any time before the customer's specified due-time. Thus a particular customer may make a request that an email or short message should reach his friend before a specified due-time. Delayable service is attractive to the network operators because it can often be transported through the network at non-busy hours. It will also be welcomed by the customers as they can enjoy service before due-time but pay less. Delayable service is also attractive in wireless networks because current wireless data traffic is bursty and bandwidth has low utilization most of the time. Diverse user requirements are also accommodated. A traffic scheduling algorithm is therefore required to support delayable service in concurrence with traditional wireless service, such as real-time service and best-effort service.

The radio link can be characterized by a time-varying multipath fading channel. It causes the link quality to vary with time. When the transmitter is provided with channel fading information, the transmission schemes can be adapted to it, allowing the channel to be used more efficiently. We propose to allocate channels combined with error-control coding schemes [3] and power control scheme. The transmission power can be used to compensate deep fades of channels. But this causes large interference to other users, leading to capacity reduction. The error-control coding scheme corrects some transmission errors at the cost of some redundant checking bits. On the other hand, with the error-correcting capability, interference is more tolerable. Therefore more coded data or more users can be admitted for transmission, which may compensate the coding overhead. To make full use of resource we jointly design these parameters to achieve a higher system throughput, subjected to the required decoded bit error rates. Queue information in MAC layer can also be considered to make the tradeoff between fair allocation among flows and system throughputs.

1.1 Research Background

The third-generation mobile communication system has been under active research and development in the past decade. The first issue to decide on is, of course, the air interface. After much effort by the various technical groups at International Telecommunication Union (ITU), a family of air interface standards is agreed upon. The universal terrestrial radio access (UTRA) is mainly a joint European-Japanese contribution. UTRA consists of two parts, the Frequency Division Duplex (FDD) part, choosing wideband code division multiple access (WCDMA) as air interface, is for wide-area coverage and paired spectrum allocation; the Time Division Duplex (TDD) part, choosing Time Division CDMA (TD-CDMA) as air interface, is for local-area coverage and unpaired spectrum allocation.

In UTRA, a traffic channel is identified by an orthogonal variable spreading factor (OVSF) code and OVSF codes can support multirate transmissions for different users. According to 3GPP technical specifications [4], [5], multicode transmission and singlecode transmission are both possible to support multirate multimedia applications. In general, however, single-code transmission is preferred since it requires a single decoder at the mobile terminal regardless to the rate transmitted from a base station [6]. We focus on the OVSF scheme in this thesis. The OVSF codes can be generated in the form of tree structure [7]. An OVSF code in the tree is assigned to a transmission request only when the code can fulfill the requested rate and is orthogonal to other codes already assigned. However, there is a problem called code blocking, first raised by Minn and Siu [8], which leads to a high blocking probability for high rate transmission requests. Great interest has been drawn recently to solve this problem in the OVSF codes assignments.

Yum and Chen [2] proposed the delayable service in cable networks and proposed to overflow the accepted delayable traffic onto the channels allocated to real-time service without violating the blocking requirement of real-time service. Simulation results showed that it can greatly increase the channel utilization without affecting service requirements. Passas et al [9] designed an interesting scheduling algorithm used in the media access control (MAC) protocol for the wireless ATM networks. The scheduling algorithm focused on satisfying delay constraints of the various connections. It was based on the intuitive idea that, in order to minimize the fraction of expired ATM cells, each ATM cell should be initially scheduled as close to its deadline as possible. This idea also accommodates the nature of delayable traffic.

Adaptive transmission, which requires accurate channel estimation at the receiver and a reliable feedback path between the receiver and transmitter, was first proposed in the late 1960's [10]. Interest in these techniques was short lived, perhaps due to hardware constraints, lack of good channel estimation techniques, and/or systems focusing on point-to-point radio links without transmitter feedback. The fact that these issues are less constraining in current systems, coupled with the growing demand for spectrally efficient communication, has revived interest in adaptive modulation methods. For single user narrowband wireless communications, adapting the transmit rate and power to the channel fade variations is a well-known strategy to enhance the average throughput. Adaptive variation of transmission power was considered in [10], following adaptive variation of data rate [11], constellation size [12], and adaptive coding scheme [13], all for narrow-band systems. Multirate DS-CDMA remains a relatively unexplored area of research. "Adaptation" in the context of CDMA systems has been mostly synonymous with power control. Wireless networks of the past have been designed primarily for voice traffic. Due to the delay intolerant nature of voice, rate adaptation is not desirable. Thus, in the past, the goal for system design has been to provide constant rate communications, using power adaptation to compensate for channel fades and propagation path loss. However, wireless data is viewed as a vast source of revenue for future wireless systems. The delay tolerant nature of data traffic allows rate adaptation. This leads to a significant shift in the objectives governing the design of wireless systems. Instead of traditional power adaptation schemes that maintain a constant rate, the focus is now on joint rate and power adaptation schemes that maximize the total throughput. CDMA schemes lend themselves to rate adaptation in a simple manner by using multiple codes, multiple processing gains, or multirate modulations. The throughput gains with multirate CDMA schemes have been studied in [14]-[20]. Wasserman and Oh [14] considered optimal (throughput maximizing) dynamic spreading gain control with perfect power control. They also considered optimal joint rate and power adaptation subject to a peak transmit power constraint and a maximum interference constraint [15]. Adaptive code rates with multiple orthogonal codes were considered in [16]. Hashem and Sousa [17] showed that limiting the increase in power to compensate for multipath fading, and getting the extra gain required by reducing the transmission rate, can increase the total throughput by about 231% for flat Rayleigh fading. Kim and Lee [18] showed the power gains achieved by the same scheme, and also considered truncated rate adaptation.

1.2 Contributions of the Thesis

In Chapter 2, we review the property of OVSF codes, which will be used as platform in the following chapters.

In Chapter 3, we study the support of delayable service in wireless network in concurrence with traditional services, such as real-time service and best-effort. Delayable packets are scheduled as close to their deadlines as possible. Some channels are reserved for real-time service based on the measurement of its traffic load. The leftover bandwidth is nominated to delayable service. On the occurrence of real-time burst, we propose to overflow real-time connections to the channels nominally assigned to delayable service. It is feasible if 1) the accepted delayable traffic can be rearranged to release similar amount of bandwidth in the following slots; 2) although the bandwidth in some slots cannot be released, the overflowed real-time connections will terminate before these slots. An algorithm is designed to rearrange the accepted delayable traffic and to determine the amount of bandwidth that can be lent to real-time service. We also derive the blocking probability of delayable service. Numerical results show that the bandwidth utilization is increased without affecting service requirements and the blocking probability of real-time service is decreased with the rearrangement of delayable traffic.

When the channel can be estimated, the transmission schemes can be adapted relative to the channel characteristics. However it is too complex to optimize many parameters simultaneously in practice. Therefore, we make some assumptions to reduce the complexity and to find the sub-optimal solutions. We first assume that power adaptation is applied only to reduce the near-far effect in CDMA systems, based on the disadvantage that much of the adapted power is used in compensation for deep fades and translates into large interference to other users [18]. Therefore we combine the number of active users and the error-control coding schemes. A scheduling algorithm is designed consequently. We then assume that all voice users can be accommodated and only one data user is admitted for transmission in each slot. The problem therefore transforms to maximize the data packets transmitted in each slot with respect to transmission power and error-control coding schemes. Based the assumption that voice users with similar channel conditions adopt the same parameters, we reduce the computational complexity further. A heuristic scheduling algorithm is consequently proposed. For data users, we allocate resource based on their channel conditions in PHY layer and queue information in MAC layer jointly. It is the tradeoff between system throughput and fair allocation among flows.

1.3 Organization of the Thesis

The rest of the thesis is organized as follows. We describe the properties of OVSF codes and survey the assignment strategies in Chapter 2. In Chapter 3, we support delayable service in concurrence with traditional services, such as real-time service and best-effort service, in wireless networks and extend the scheduling algorithm to adapt the traffic burst. In Chapter 4, we assign transmission parameters with joint information and propose heuristic algorithms to reduce the computational complexity. This thesis is concluded in Chapter 5.

Chapter 2 Properties of OVSF Codes

In UTRA, a traffic channel is identified by an orthogonal variable spreading factor (OVSF) code and OVSF codes can support multirate transmissions for different users. According to 3GPP technical specifications [4], [5], multicode transmission and singlecode transmission are both possible to support multirate multimedia applications. In general, however, single-code transmission is preferred since it requires a single decoder at the mobile terminal regardless to the rate transmitted from a base station [6]. We focus on the OVSF scheme in this thesis. The OVSF codes can be generated in the form of tree structure [7]. An OVSF code in the tree is assigned to a transmission request only when the code can fulfill the requested rate and is orthogonal to other codes already assigned. However, there is a problem called code blocking, first raised by Minn and Siu [8], which leads to a high blocking probability for high rate transmission requests. Great interest has been drawn recently to solve this problem in the OVSF codes assignments.

2.1 Tree-Structured Generation of OVSF Codes

Spectrum spreading is achieved in DS-CDMA by mapping each data bit $\in \{1, -1\}$ into an assigned code sequence. The length of the code sequence per data bit is called the Spreading Factor or the bandwidth-expansion factor [35], and it is denoted by N. For example, in IS-95 and IMT-2000 WCDMA standards, the spreading factors are 64 and 256, respectively [36], [37]. An important class of orthogonal code sequence is the family of Walsh codes. Figure 2.1 (a) illustrates the Walsh encoded sequence for data bits $\{+1, -1\}$ using Walsh code $\{+1, -1, -1, +1\}$ in a DS-CDMA system. The spreading factor N is 4 in this case. Comparing Figure 2.1 (a) and (b), the larger spreading factor, the lower data rate with the same chip rate.

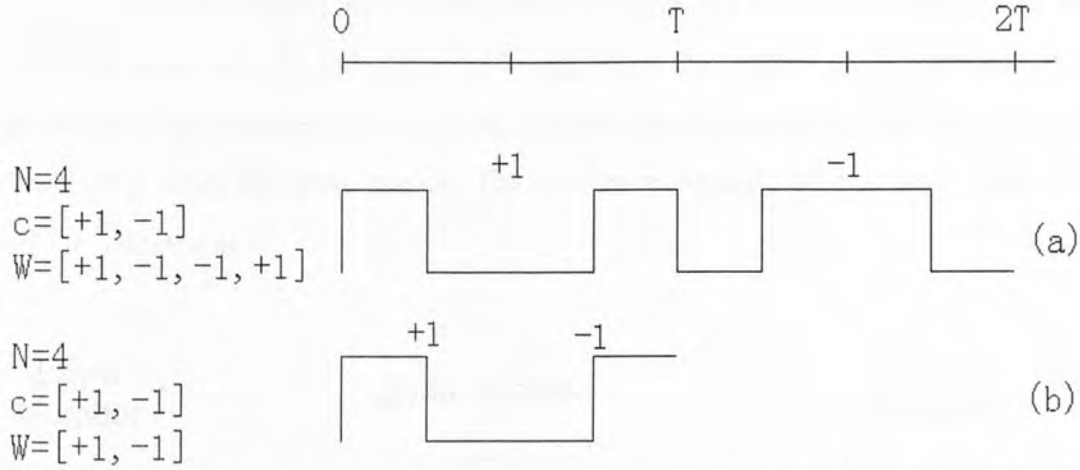

Figure 2.1 Walsh encoded waveforms for data sequence $[+1, -1]$ using Walsh code $[+1, -1, -1, +1]$, $N = 4$, and Walsh code $[+1, -1]$, $N = 2$

Let C_N denote the set of N binary spreading codes of N-chip length, $\{C_N(n)\}_{n=1}^{N}$, where $C_N(n)$ is the row vector of N elements and $N = 2^K$ (K is a positive integer); it is generated from $C_{N/2}$ as

$$C_N = \begin{bmatrix} C_N(1) \\ C_N(2) \\ C_N(3) \\ C_N(4) \\ \vdots \\ C_N(N-1) \\ C_N(N) \end{bmatrix} = \begin{bmatrix} C_{N/2}(1)C_{N/2}(1) \\ C_{N/2}(1)\overline{C}_{N/2}(1) \\ C_{N/2}(2)C_{N/2}(2) \\ C_{N/2}(2)\overline{C}_{N/2}(2) \\ \vdots \\ C_{N/2}(N/2)C_{N/2}(N/2) \\ C_{N/2}(N/2)\overline{C}_{N/2}(N/2) \end{bmatrix}, \tag{2.1}$$

where $C_{N/2}(n)$ is the row vector of $N/2$ elements and $\overline{C}_{N/2}(n)$ is the binary complement of $C_{N/2}(n)$. As a result, tree-structured spreading codes are generated recursively as shown in Figure 2.2. Starting from $C_1(1) = 1$, a set of 2^k spreading codes are generated at the kth layer ($k = 1, 2, \ldots, K$) from the top. The code length of the kth layer is 2^k chips and can be used for the code channels transmitting data at 2^{K-k} times the lowest rate. The data rate a code can support is called its capacity. The codes in the lower layer in Figure 2.2 have the larger spreading factor and the lower capacity. Let the capacity of the leaf codes (in layer K) be R. The capacity of the codes in layer $(K-1)$, $(K-2)$, ..., 1 and 0 are $2R$, $4R$, ..., $2^{K-1}R$ and $2^K R$ respectively, as shown in Figure 2.2. We also note that the capacity of OVSF codes is a power of 2 times the basic rate R. The maximum capacity of a K-layer code tree is $2^K R$, denoted as C.

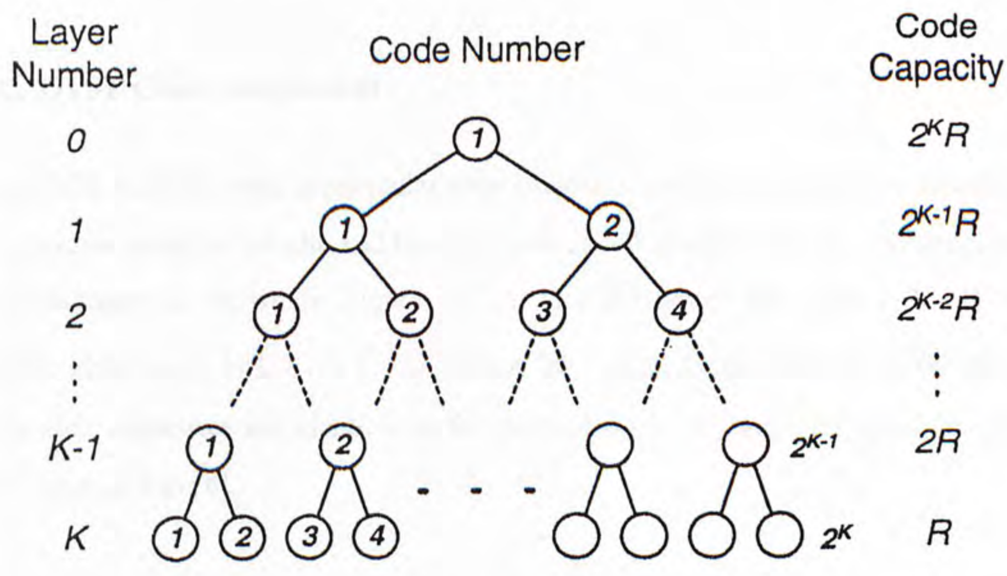

Figure 2.2 A K-layer code tree

Layer k has 2^k codes and they are sequentially labeled from left to right, starting from one. The m th code in layer k is referred to as code (k, m). The total capacity of all the codes in each layer is $2^K R$, irrespective of the layer number. For a typical code (k, m) ($k \geq 1$), its ancestor codes are the codes on the path from (k, m) to the root code $(0, 1)$. On the other hand, the descendant codes of (k, m) ($k \leq K-1$) are the codes in the branch under (k, m).

An OVSF code in the tree is assigned to a transmission request only when the code can fulfill the requested rate and is orthogonal to other codes already assigned. It can be understood from formula (2.1) that generated codes of the same layer constitute a set of Walsh functions and they are orthogonal. Furthermore, any two codes of different layers are also orthogonal except for the case that one of the two codes is an ancestor code of the other [20]. However, there is a problem called code blocking, first raised by Minn and Siu [8], which leads to a high blocking probability for high rate transmission requests. Great interest has been drawn recently to solve this problem in the OVSF codes assignments.

2.2 OVSF Codes Assignment

In OVSF-CDMA, once a particular code is used, simultaneous use of its descendant or ascendant codes is not allowed because their encoded sequences are indistinguishable. An example is shown in Figure 2.1 (a) and (b) where the code $\{+1,-1\}$ and its descendant code $\{+1,-1,-1,+1\}$ cannot be assigned simultaneously because their encoded sequences are identical in the interval $0 < t < T$. Another example is shown as Figure 2.3 in [8].

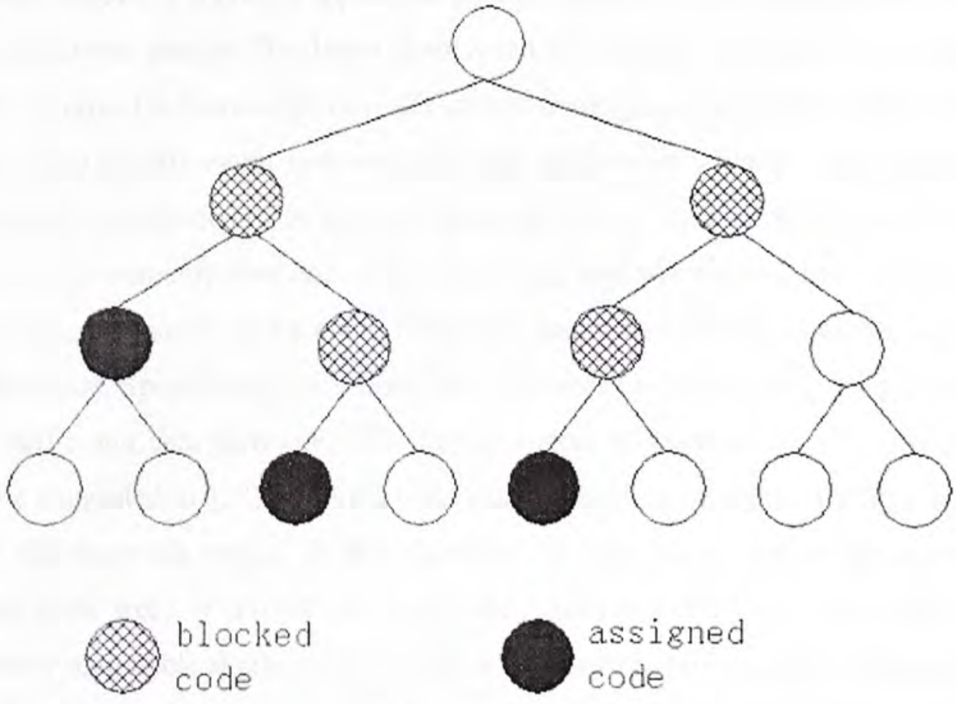

Figure 2.3 OVSF code blocking

Code $(2,1)$, $(3,3)$, and $(3,5)$ are already assigned and the capacity used is $4R$. Assuming an ideal code-limited (single-cell) scenario, the system can support a maximum capacity of $8R$, since there are 8 leaves and each leaf support R rate. Hence the unused capacity is $(8-4)=4R$. However, codes $(1,1)$, $(1,2)$, $(2,2)$, and $(2,3)$ are blocked by their respective descendant codes. (Note that code $(2,4)$ is available for assignment.) As a result, a new call requesting $4R$ is blocked although there is enough capacity left. The code blocking is defined as the condition that a new call cannot be supported although the system has excess capacity to support the rate requirement of the call [8]. It is important to note that code blocking only occurs in higher layer codes. It is also easy to see that the higher the layer number of a code, the larger the code blocking probability.

Many allocation strategies have been presented in the literature to address the code blocking issue. There are two types of code assignment schemes: nonrearrangeable and rearrangeable. In [21], several code assignment schemes were proposed. The static (nonrearrangeable) approach applies the first-fit scheme (for the bin packing problem) in the algorithm design. The dynamic approach is based on a tree partitioning method, which requires the knowledge of traffic composition (percentage of different data-rate users). Two priority-based rearrangeable code assignment schemes were proposed in [22] and [23], respectively, to accommodate both the real time traffic (circuit-switched, e.g., voice communication and video streaming) and the nonreal time traffic (best-effort, e.g., file transfer and e-mail). Obviously, real time traffic has a higher priority to obtain a code. Specifically, the scheme in [23] makes use of the bursty property of real time traffic and can, therefore, offer higher system utilization. The code assignment scheme suggested in [22] performs code reassignment for real time traffic classes on every call departure instant. At these instants, the "right-most" call in the same layer (of the code tree) is moved to occupy the "just-released" code. As a result, the remaining assignable single-code capacity is maximized. Subsequently, Chen et al [24] extended this scheme by partitioning the codes into two groups based on code capacity. When a code is released, the "right-most" or the "left-most" (according to the group the code belongs to) call in the same layer will be rearranged to the "just-released" code. After that, the ongoing calls in the lower layers (if any) are rearranged similarly, layer by layer. Furthermore, the region division assignment (RDA) scheme presented in [19] divides the code tree into multiple mutually exclusive regions with each region dedicates to a particular transmission data rate. When a new call cannot be accommodated in the corresponding region, a suitable code in other regions is borrowed and assigned to the new call. In [8], Minn and Siu proposed a rearrangeable assignment scheme whereby the number of OVSF codes that must be rearranged to support a new call is minimized. According to the authors, the main challenge of using this scheme lies in the searching effort of the "minimum-cost" branch. The "cost of a branch" is defined in [8] as the minimum number of code rearrangements necessary to reassign all occupied codes in the branch to other branches so that the branch is left empty. Another concept named flexibility index is defined in [20] to measure the capability of an assignable code set in supporting multirate traffic. Based on this concept, the authors proposed two computational efficient single-code assignment schemes, namely compact assignment (CA) and rearrangeable compact assignment (RCA). Both schemes leave the system as flexible as possible after each code assignment.

Inspired by the idea of region division, we propose to reserve codes according to the blocking probabilities of each supported class. Let the arrival of requests be Poisson process and the holding time be exponential distribution. The traffic load of class-k is ρ_k and the acceptable blocking probability is B_k. The capacity reserved for class-k is given by Erlang B

$$C_k = \min\left\{ c \left| \frac{\rho_k^c / c!}{\sum_{i=0}^{c} \rho_k^i / i!} \leq B_k \right. \right\}. \tag{2.2}$$

If $\sum_k C_k < C$, the leftover capacity is shared among classes or equally reserved to each supported class. Given acceptable blocking probability, Figure 2.4 illustrates the capacity reserved verse different traffic load. With the measure of traffic load, we can update the reservation volumes of each class periodically. The class whose blocking probability is not satisfied by reservation will be given higher priority to borrow.

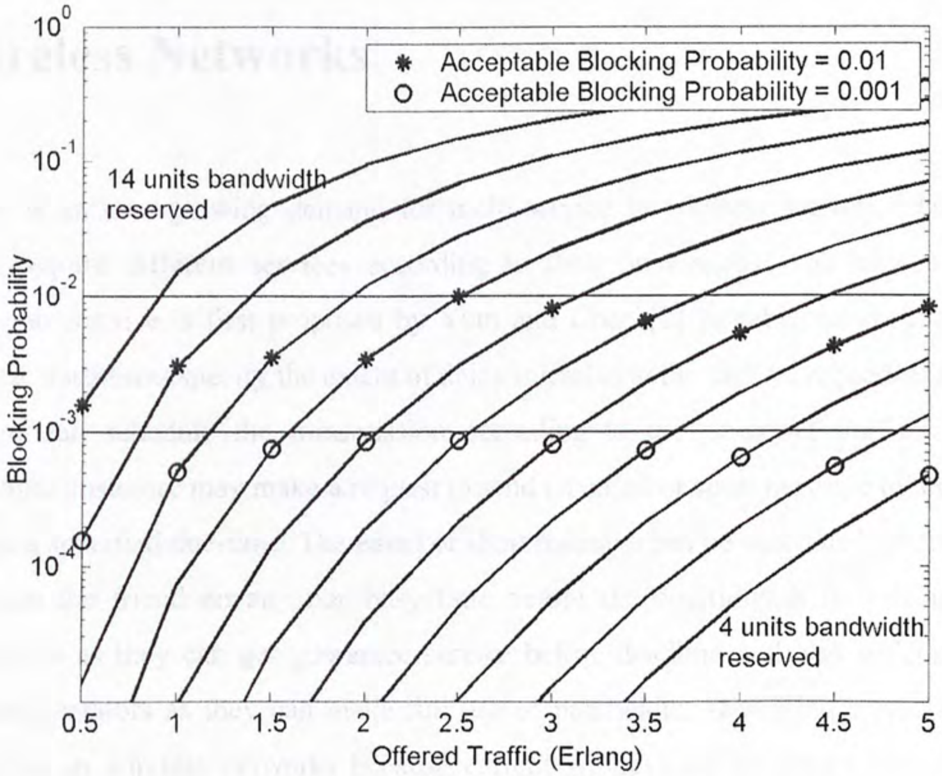

Figure 2.4 Reserved bandwidth versus offered traffic load

Chapter 3 Support Delayable Traffic in Wireless Networks

There is an ever-growing demand for multi-service in wireless networks. Different users require different services according to their own needs. The service called delayable service is first proposed by Yum and Chen [2] in cable networks. In this service, customers specify the extent of delay tolerable at the time of request so that the system can schedule the transmission according to the on-going traffic. Thus a particular customer may make a request to send an email or short message to his friend before a specified due-time. The email or short message can be scheduled for transport to reach the friend on any non-busy time before the due-time. It is welcomed by customers as they can get guarantee service before deadline and also welcomed by system operators as they can make full use of bandwidth. Delayable service is also attractive in wireless networks because current wireless traffic occurs as burst and bandwidth has low utilization in other period. A traffic scheduling algorithm is therefore required to support this service in concurrence with traditional wireless service, such as real-time service and best-effort service.

In multi-service systems, some bandwidth is reserved for real-time service, but usually only a fraction of them is busy. Yum and Chen [2] proposed to overflow the accepted delayable traffic onto the channels allocated to real-time service without violating the blocking requirement of real-time service. We consider the problem in the other way around. Bandwidth is reserved for real-time service based on the measurement of real-time traffic load. The leftover resource is nominated to delayable service and the accepted delayable traffic should not exceed such a boundary. It is based on the intuitive idea that, in order to maximize the fraction of packets that are transmitted before their deadlines, each delayable packet is initially scheduled for transmission as close to its deadline as possible. On the other hand, delayable traffic is pre-scheduled and the leftover bandwidth is not equal in the following slots.

Since the real-time traffic has random arrival time and random service duration, precise prediction of real-time traffic load is impossible. When a real-time burst occurs, more channels will be required by real-time service. However, it is difficult because the bandwidth in some following slots has been occupied by delayable traffic and it is unbearable for real-time connections to be interrupted. We propose to smooth the scheduled delayable traffic to release similar amount of bandwidth in the following slots. Real-time traffic is therefore overflowed onto the bandwidth nominated to delayable service. More uniform the delayable traffic, more real-time connections can be accepted. The difficulty to smooth delayable traffic is that delayable traffic is initially scheduled at its deadline so there is just one direction to move the packets. We define ascending slot series and prove that smoothing all of the ascending slot series in the following slots results in the uniformity of overall delayable traffic.

3.1 System Model

Let R be the minimum capacity of codes in a K-layer OVSF code tree. The capacity of codes in the tree is a power of two times R. Therefore R is also referred to as the Bandwidth Unit of OVSF codes. The number of BU of a code in layer k is 2^{K-k} and the maximum BU of the tree is 2^K. If the required bit rate of a real time call is X, the number of BU needed is X/R. We assume that an incoming call can be served by a single code. We represent the request of best-effort service or delayable service as a packet. Such a packet is divided into one-slot-long and one-BU-capacity "segments" that are reassembled at the receiving end. In other words, a segment can be transmitted by a code with one BU capacity in one slot. Let T_s denote the length of a slot. If the size of a packet is S, the number of segments it contains is expressed as $S/(R \cdot T_s)$. The number of segments scheduled in a slot corresponds to the number of BU should be allocated to the user in the slot. This number should also be a power of two.

For the purpose of the algorithm, time slots are numbered. The slot in which the target request arrives is numbered as zero. The slots following or preceding that slot are numbered starting from 1 and -1 respectively. We agree that "forward" is the direction towards the positive slots and "backward" is the direction towards the negative slots.

First, we derive the number of BU should be reserved for real-time class, which satisfies the blocking probability of real-time class. Suppose real-time class supports I types of rate. A rate-i call requires b_{ri} units BU to set up connection and offers the mean traffic load ρ_{ri}. Let x_i denote the number of ongoing rate-i calls and $\vec{x} = (x_1, x_2, \ldots, x_I)$. The steady state probability of on-going real-time calls is denoted as $P(\vec{x})$. If C_r units BU are reserved for real-time class, $P(\vec{x})$ is derived in [25] to be of the product form.

$$P(\vec{x}) = \left[\sum_{\vec{x} \in \Omega} \prod_{j=1}^{I} \frac{1}{x_j!} (\rho_{rj})^{x_j} \right]^{-1} \cdot \prod_{j=1}^{I} \frac{1}{x_j!} (\rho_{rj})^{x_j}, \tag{3.1}$$

where $\Omega = \left\{ \vec{x} \middle| \sum_{j=1}^{I} x_j b_{rj} \leq C_r \right\}$.

Let $B_r(i)$ denote the blocking probability of rate-i calls. It is given by

$$B_r(i) = \sum_{\vec{x} \in \Omega_i} P(\vec{x}), \tag{3.2}$$

where $\Omega_i = \left\{ \vec{x} \middle| C_r - \sum_{j=1}^{I} x_j b_{rj} \leq b_{ri} \right\}$.

The overall blocking probability of real-time class B_r is the weighted sum of $B_r(i)$'s, or

$$B_r = \left[\sum_{j=1}^{I} \rho_{rj} \right]^{-1} \cdot \sum_{j=1}^{I} \rho_{rj} B_r(j). \tag{3.3}$$

Therefore, if the required blocking probability of real-time class is B_{req}, we need reserve C_r units BU, which satisfies $B_r \leq B_{req}$.

Then, let C_d be the maximum number of BU can be used by delayable class in each slot. It is given by

$$C_d = 2^K - C_r. \tag{3.4}$$

C_d is the upper bound that the accepted delayable packets in each slot should not exceed. When a delayable packet with deadline in slot L comes in, it will be accepted if there are enough bandwidth left before its deadline with the constraint of C_d. The packet will be divided into segments and packed backward starting from slot L. It is based on the intuitive idea that, in order to maximize the fraction of packets that are transmitted before their deadlines, each packet is initially scheduled for transmission as close to its deadline as possible. This idea is first proposed by Passas *et al.* [9] in wireless ATM networks. It also accommodates the nature of delayable traffic.

Best-effort packets are scheduled after the satisfaction of real-time traffic and delayable traffic. Best-effort packets are divided into segments and queued for transmission. The queue policy is decided by system operators. A packet is removed from the queue only when all of its segments have been completely transmitted.

If best-effort packets do not exhaust the resources in a slot, delayable segments scheduled in the following slots will be moved backward based on the idea that codes never stay idle as long as there are segments requesting transmission. The backward procedure will begin with the segments scheduled in the nearest slots.

3.2 Scheduling Algorithm with Burst Adaptation

When a real-time burst occurs, overflowing these real-time connections can reduce their blocking probability. It is feasible if 1) the accepted delayable traffic can be rearranged to release similar amount of bandwidth in the following slots; 2) the overflowed real-time connections will terminate in time.

Suppose the maximum delay of real-time class is n slots and most of real-time connections will terminate in L slots. If we could rearrange enough bandwidth in the following $(n+L)$ slots, the real-time connections will be overflowed on arrival. We assume $n=1$ for ease of analysis. Suppose the duration of a real-time connection is geometrically distributed with parameter p which is the probability that the connection will end up in a slot no matter how long it has held. The probability connections don't terminate before slot L is given by

$$P_L = 1 - \sum_{k=1}^{L} p(1-p)^{k-1} = (1-p)^L . \tag{3.5}$$

Delayable traffic is smoothed from slot one to slot L to release bandwidth for real-time service. Therefore, real-time connections should terminate before slot L to return bandwidth, otherwise they will be dropped. We refer to the probability real-time connections dropped with as dropping probability. Given the acceptable dropping probability Y, we derive the number of L as

$$L = \log_{(1-p)} Y . \tag{3.6}$$

Other distributions of holding time are also applicable if slot L can be derived.

The scheduled delayable traffic is rolling and they can be moved just one direction. We now consider the problem to determine the amount of bandwidth to be released. For two slots, say slot i and slot j ($i<j$), when we could not move some segments backward from slot j to slot i to reduce the difference of the number of segments in the two slots, we regard that the distribution of traffic in the two slots is uniform. Let d_k denote the number of delayable segments accepted in slot k. We define a slot series from slot i to slot j as an ascending slot series when the number of segments in the slots satisfies $d_i \le \mathrm{K} \le d_k \le \mathrm{K} \le d_j$ and $(d_j - d_i) > 1$, $i \le k < j$. For an ascending slot series, after the traffic is smoothed, the difference of the number of segments in any two slots will be less than or equal to one. Let d'_k denote the number of segments in slot k after the traffic is smoothed. Obviously, for an ascending slot series, no segments in the slots will be moved forward when the traffic is smoothed. We prove a proposition before describing the algorithm to smooth delayable traffic.

Proposition 1: In the area from slot one to slot L, if there is no ascending slot series, no segments can be moved backward to make the traffic more uniform.

Proof: We prove this by contradiction.
Suppose there exist slot i and slot j where $i < j$ and $(d_j - d_i) > 1$. If $j = i + 1$, slot i and slot j form an ascending series. It contradicts the premises. If $j > i + 1$, we consider the segments scheduled in slot $(i+1)$ to slot $(j-1)$. $d_k \le d_i + 1$ holds for $k = (i+1), \ldots, (j-1)$. Since $d_j > d_i + 1$, there must be an ascending series in the area from slot i to slot j. It contradicts the premises and the assumption is wrong. This completes our proof.

The proposition indicates the way to smooth scheduled delayable traffic in the area from slot one to slot L. Starting from slot one, we find the first ascending slot series and smooth the number of segments in its slots. This may result in a new ascending slot series. All the slots in the area form several ascending series. Supported by Proportion 1, smoothing all these series must result in the uniformity of segments distribution. We detail the procedures to realize above algorithm as follows. Specially, Procedure A smoothes an ascending slot series. Procedure B finds the first ascending slot series since slot one. And Procedure C is the main procedure to smooth the segments distribution in slot one to slot L.

Procudure A

INPUT: d_k, $k = 1, \ldots, L$.

OUTPUT: d_k, $k = 1, \ldots, L$.

1. Repeat the following:
   1.1. Call Procedure *B* to find the first ascending series since slot one. Let slot i denote its beginning and slot j denote its end.
   1.2. IF $(d_j - d_i) > 1$, THEN
   call Procedure *C* to smooth the series.
   ELSE break.
2. $d'_k = d_k$, $k = 1, \ldots, L$.
3. Return d'_k's.

Procedure B

INPUT: an ascending series from slot i to slot j.

OUTPUT: d'_k, $k = i, \ldots, j$.

1. $t = 0$.
2. FOR $k = i + 1$ TO $k = j$, repeat $t = t + d_k - d_i$ and $d_k = d_k - d_i$.
3. $k = 0$.
4. WHILE $t > 0$, repeat the following:
   4.1 d_{j-k} ++.
   4.2 $k = (k+1)\%(j-i)$, where % is module operation.
   4.3 t − −.
5. $d'_k = d_k$, $k = i, \ldots, j$.
6. Return d'_k's.

Procudure C

INPUT: d_k, $k = 1, \ldots, L$ ..

OUTPUT: the first ascending series since slot one, from slot i to slot j.

1. $num\_i = num\_(k-1) = d_1$, $i = j = 1$, and $k = 2$.
2. WHILE $k \le L$, repeat the following:

   IF $d_k \ge num\_(k-1)$, THEN

   $\quad num\_(k-1) = d_k$.

   $\quad k = k+1$.

   ELSE IF $d_{k-1} > num\_i + 1$ $d_{k-1} > num\_i + 1$, THEN

   $\quad j = k - 1$ and break.

   ELSE

   $\quad i = k$.

   $\quad num\_i = num\_(k-1) = d_k$.

   $\quad k = k+1$.
3. IF $j > 1$ THEN return i and j.

   ELSE return failure.

Returned from *Procedure A*, d'_k's indicate the number of segments should be scheduled in each slot for uniform distribution. Let d'_M be the maximum among d'_k's. The maximum BU can be used by real-time class in slot one to slot L is $\left(2^K - d'_M\right)$. The upper bound for delayable class in these slots is decreased to be d'_M or a little larger than d'_M if the real-time traffic is not too heavy.

The difficulty of the algorithm with burst adaptation is that the come of burst can't be predicted and the delayable packets can't be moved forward. If the maximum delay of real-time class is not so stringent that it is more than one slot, our scheme will be more flexible. We also considered another scheme that the segments of a delayable packet were uniformly scheduled in the slots before its deadline when accepted. This makes the distribution of delayable segments more uniform and therefore it is easier to accept overflowed real-time connections. On the other hand, the segments of a packet are scheduled in more slots and the data terminals are required to keep alive for longer time. So this scheme isn't applicable in wireless communication.

3.3 Performance Analysis

For delayable class, suppose the packet size in unit of segment is geometrically distributed with mean $\overline{m}$ and the packet arrival process is modeled as a Bernoulli process with probability $\lambda_d T_s$ of having an arrival in each slot. In other words, in each slot a packet formed by a number of segments geometrically distributed with mean $\overline{m}$ arrives with probability $\lambda_d T_s$. Let M denote the number of segments arriving in each slot. Its probability is given by

$$P(M = m)\begin{cases} 1-\lambda_d T_s, & \text{if } m = 0 \\ \lambda_d T_s \dfrac{1}{\overline{m}}\left(1-\dfrac{1}{\overline{m}}\right)^{m-1}, & \text{if } m > 0 \end{cases}. \tag{3.7}$$

Let D denote the deadline of a packet in unit of slot. It is uniformly distributed in range $[U, V]$, where $1 < U < V$. The segments of a packet will be packed in slot D to slot one. When a packet with M segments arrives in slot zero, some segments have been scheduled in slot one to slot $(V-1)$. The target packet will be blocked if the capacity left in slot one to slot D is less than M. We are to derive the distribution of segments on arrival and thus obtain the conditional blocking probability depending on D and M. Suppose the upper bound for delayable class in the following slots is C_d. Let B_d denote the blocking probability of delayable class. Let N_{ij} denote the number of segments arriving in slot i ($i < 0$) and with deadline in slot j ($j > 0$). N_{ij} is larger than zero when three conditions are satisfied: 1) there is a packet arriving in slot i; 2) the packet is not blocked; and 3) the packet has deadline in slot j. The probability of

N_{ij} is given by (3.8). We note the probability does not depend on i and j, so we use N instead of N_{ij} for $i < 0$.

$$P(N_{ij} = n) = \begin{cases} (1-\lambda_d T_s) + \lambda_d T_s B_d + \lambda_d T_s (1-B_d)\dfrac{V-U}{V-U+1}, & if\ n = 0 \\ \lambda_d T_s (1-B_d)\dfrac{1}{V-U+1}\dfrac{1}{m}\left(1-\dfrac{1}{m}\right)^{n-1}, & if\ n > 0 \end{cases}. \tag{3.8}$$

The segments requiring to be scheduled in slot j include two parts: the segments with deadline in slot j and the segments moved backward. When a segment is moved backward, it will be first packed in the nearest slot and will be moved again if the nearest slot is full. Therefore, it is because slot $(j+1)$ is full that some segments are moved backward into slot j. Let N_j denote the total number of segments requiring slot j and let $N_{(j+1)j}$ denote the number of segments moved backward from slot $(j+1)$ to slot j. The probability of $N_{(j+1)j}$ is given by

$$P(N_{(j+1)j} = n) = \begin{cases} P(N_{j+1} = C_d + n), & n > 0 \\ \sum_{k=0}^{C_d} P(N_{j+1} = k), & n = 0 \end{cases}, \tag{3.9}$$

and N_j is given by

$$N_j = \begin{cases} (V-U+1)N + N_{(j+1)j}, & 1 \le j \le U-1 \\ (V-j)N + N_{(j+1)j}, & U \le j \le V-2 \\ N, & j = V-1 \end{cases}. \tag{3.10}$$

Since there are no segments requiring slot V, we have $P(N_V = 0) = 1$. Based on (3.8), (3.9), and (3.10), the distribution of N_j's ($1 \le j \le V-1$) can be derived recursively.

If $N_j > C_d$, some segments will be moved backward because of bandwidth limitation in each slot. Let $N_j^{(1)}$ denote the number of segments scheduled in slot j. Its probability is given by

$$P\left(N_j^{(1)} = n\right) = \begin{cases} P\left(N_j = n\right), & 0 \le n < C_d \\ \sum_{k=C_d}^{\infty} P\left(N_j = k\right), & n = C_d \end{cases}. \tag{3.11}$$

The target packet with size M and deadline D will be blocked if and only if $\sum_{k=1}^{D} N_k^{(1)} > C_d D - M$. This gives the conditional blocking probability. The blocking probability of delayable class is then derived by considering the distribution of packet size and its deadline.

Once a delayable packet is accepted, it is supposed to be transmitted completely before its deadline. Therefore the delay of delayable class is not concerned.

For real-time class, the blocking probability is given by (3.3) when C_r units BU are reserved.

For best-effort class, suppose the arrival process is modeled as Poisson process and the queue policy is Short-Job-First. It follows that the best-effort queue is Head-Of-the-Line [26]. The service rate depends on the traffic load of real-time class and delayable class at the time.

3.4 Simulation Results

In this section, we compare the blocking probability and system throughput when delayable service is supported or not, and when rearrangement strategy is applied or not. Real-time service is treated as voice calls, each requiring a code with one BU capacity. It arrives as Poisson process and has exponential holding time. Delayable service also has Poisson process and exponential size. The deadline of delayable packet is uniformly distributed in range 4 to 8 in unit of slot. Best-effort service is not considered in simulation. Let the system capacity be 64 units BU. Codes are reserved for real-time class based on the method stated in Section 3.1. The required blocking probability of real-time class is set to be 0.001.

We first compare the performance when delayable service is supported or not, by increasing the offered traffic load, one half from real-time service and the other half from delayable service. When delayable service is not supported, the same amount of traffic is treated as if it were real-time traffic with the arrival time set at their respective due-time. The results are shown as Figure 3.1 and Figure 3.2.

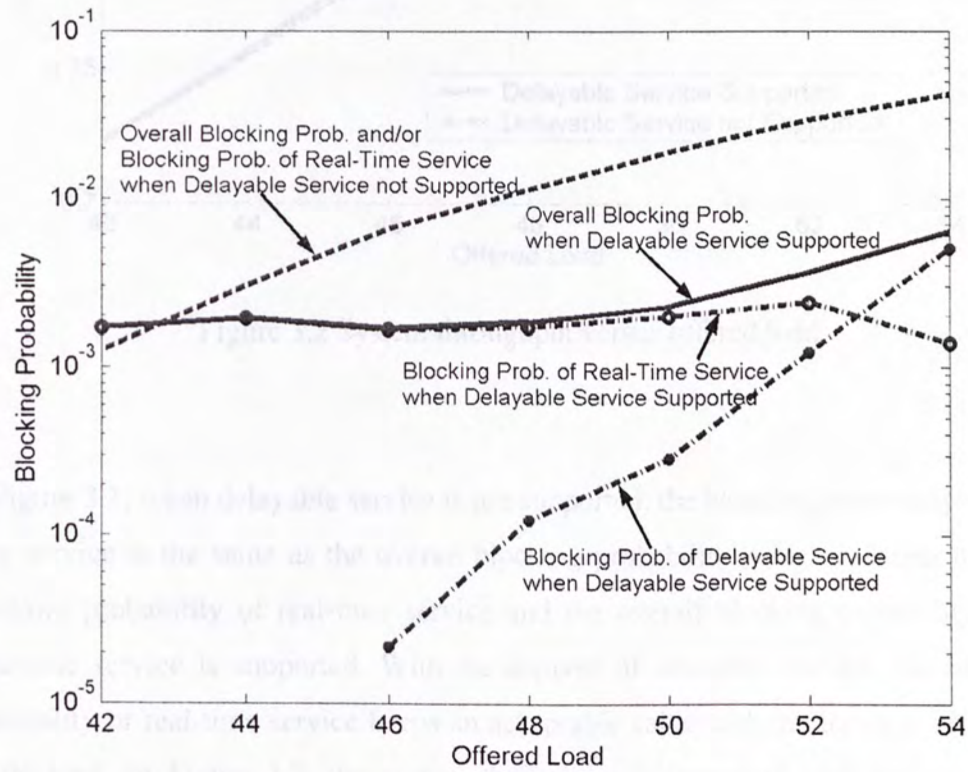

Figure 3.1 Blocking probability versus offered load

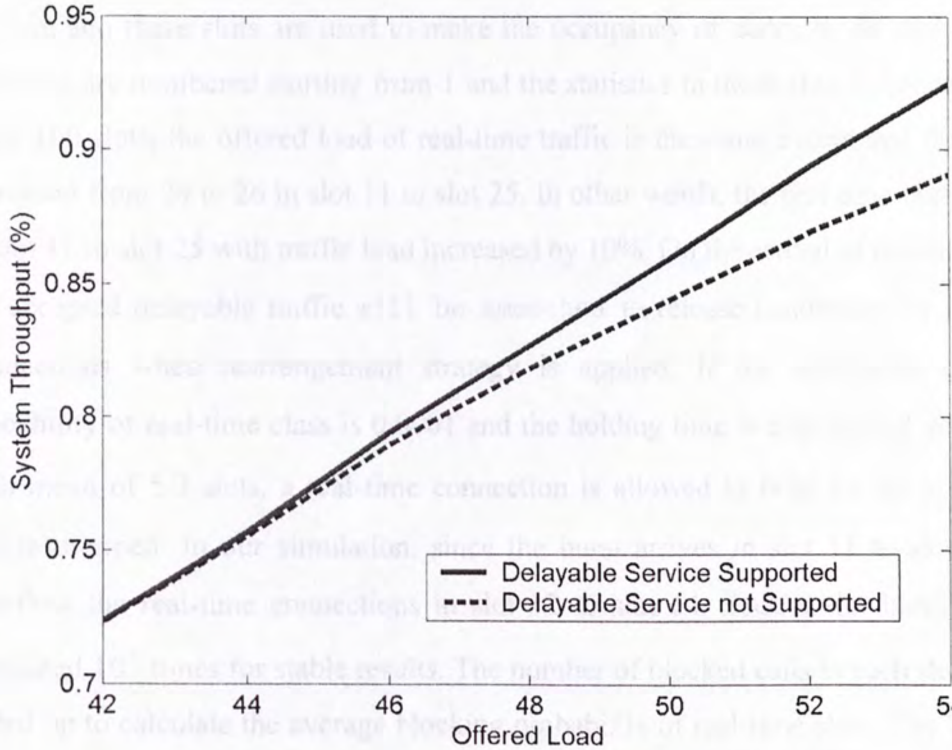

Figure 3.2 System throughput versus offered load

In Figure 3.1, when delayable service is not supported, the blocking probability of real-time service is the same as the overall blocking probability, which is larger than the blocking probability of real-time service and the overall blocking probability when delayable service is supported. With the support of delayable service, the blocking probability of real-time service keeps an acceptable value with the increase of offered traffic load. In Figure 3.2, the system throughput is improved with the support of delayable service, compared with that achieved where delayable service is not supported. The improvement is up to 5% when the blocking probability of real-time service is still acceptable.

We then compare the performance when rearrangement strategy is applied or not. We set the size of a slot block to be 500 slots. The offered traffic load in the first 400 slots is set to be 48, one half from real-time service and the other half from delayable service, and these slots are used to make the occupancy of codes be steady. The last 100 slots are numbered starting from 1 and the statistics in these slots is concerned. In these 100 slots, the offered load of real-time traffic is the same except that the load is increased from 24 to 26 in slot 11 to slot 25. In other words, the real-time burst occurs in slot 11 to slot 25 with traffic load increased by 10%. On the arrival of real-time burst, the accepted delayable traffic will be smoothed to release bandwidth for real-time connections when rearrangement strategy is applied. If the acceptable dropping probability of real-time class is 0.0001 and the holding time is exponential distributed with mean of 5.0 slots, a real-time connection is allowed to hold for up to 39 slots before dropped. In our simulation, since the burst arrives in slot 11 to slot 25, we overflow the real-time connections in slot 12 to slot 64. Such a slot block will be simulated 10^5 times for stable results. The number of blocked calls in each slot will be added up to calculate the average blocking probability of real-time class. The blocking probability of real-time service is shown in Figure 3.3 and the system throughput is shown in Figure 3.4.

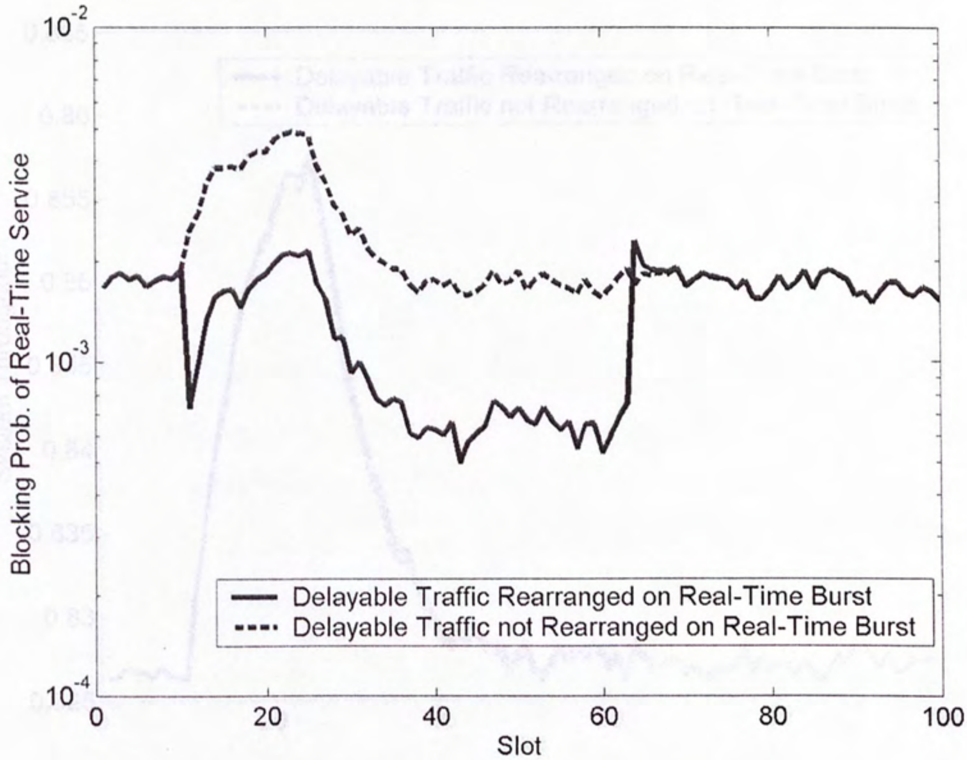

Figure 3.3 Blocking probability of real-time service on real-time burst

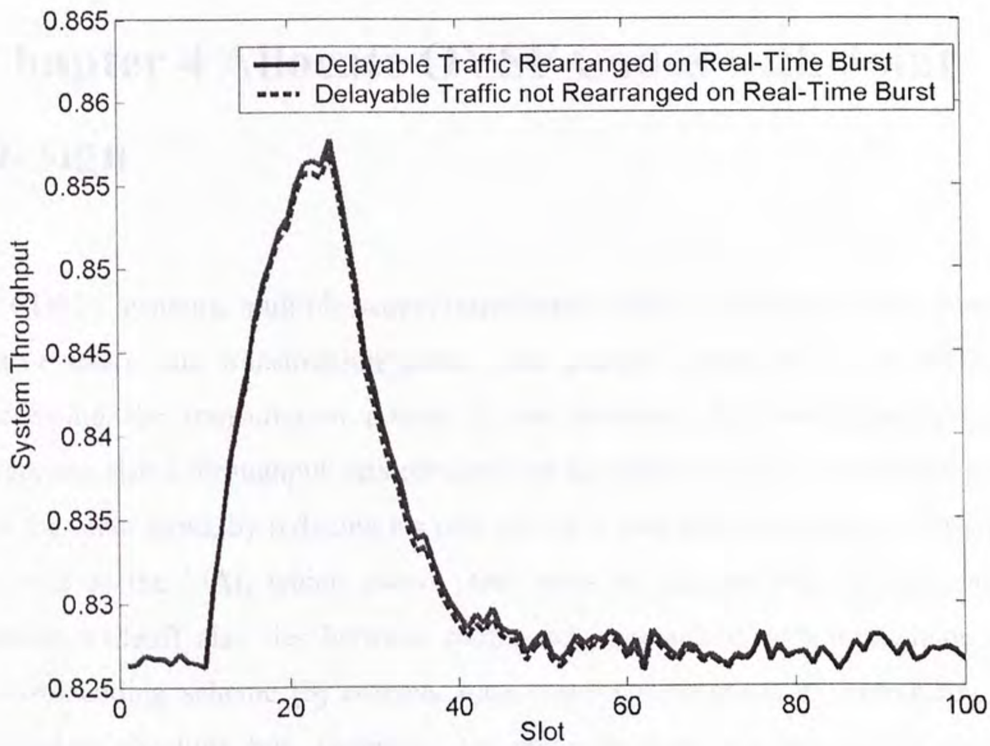

Figure 3.4 System throughput on real-time burst

In Figure 3.3, we note that the blocking probability of real-time service keeps an acceptable value even when the real-time burst occurs in slot 11 to slot 25, by rearranging the accepted delayable traffic. We also note that the blocking probability is much lower after the burst in slot 25 to slot 64. It is because that more bandwidth is still used by real-time class. After slot 64, delayable class retrieves its bandwidth and most connections of the burst have terminated, so the blocking probability in the two cases becomes the same. In Figure 3.4, we note the system throughput is similar achieved in two cases. It shows that the system throughput is not descended while the blocking probability of real-time service is maintained. It is because that some bandwidth is borrowed from delayable service to serve the real-time burst.

Chapter 4 Allocate OVSF Codes with Joint Design

In CDMA systems, multiple access interference (MAI) is affected by the number of active users, the transmission power, the channel coding rate, and so on. By increasing the transmission power, a user increases his own throughput while everyone else's throughput gets reduced due to additional MAI contributed by him. On the other hand, by reducing his own power, a user decreases his own throughput, as well as the MAI, which allows other users to increase their throughputs. The similar tradeoff also lies between coding schemes and system throughput. Error-control coding scheme [3] corrects some transmission errors by introducing some redundant checking bits. Generally, the more checking bits transmitted, the larger error-correcting capability achieved, but the fewer information bits transmitted in a block. On the other hand, with the error-correcting capability, more coded data or more users can be admitted for transmission, which may compensate the low coding efficiency. To make full use of resource we assign all these parameters jointly to achieve higher system throughput, subjected to the required decoded bit error rates. Queue information in MAC layer can also be considered to make the tradeoff between fair allocation among flows and system throughputs.

OVSF codes are allocated to achieve the highest system throughput with the constraint of acceptable bit error rate (BER) performance. We make some assumptions to find the effects of different factors. In Section 4.1, we consider the number of active users and their error-control coding schemes, assuming the power adaptation is only applied in near-far problem. In Section 4.2, we combine the power adaptation and the error-control coding scheme, assuming voice traffic can be accommodated and only one data user is admitted in each slot. Information of data queue is also considered to select the best user for transmission.

4.1 Combine Number of Active Users and Error-Control Coding Scheme

4.1.1 System Model

Error-control coding scheme corrects some bit errors by introducing some redundant checking bits. We adopt linear block codes as coding scheme in this section. Linear block codes have mathematic expressions on the error-correcting capability and the coding efficiency, which highlights the effect of bit error probability and system throughput. Other coding schemes are also applicable and may demonstrate advantages in error-correcting capability or coding efficiency, such as Turbo codes proposed in [27]. For an (n, x) linear block code, x denotes the number of information bits and the coding efficiency is $r = x/n$. The error-correcting capability of a linear block code is determined by the parameter called the minimum distance, $d_{\min}$. A code with $d_{\min}$ guarantees correcting all the error patterns of $t = \lfloor (d_{\min} - 1)/2 \rfloor$ or fewer errors, where $\lfloor z \rfloor$ denotes the largest integer no greater than z. The parameter t is called the error-correcting capability of the code. For example, the linear block code $(7, 4)$ has minimum distance 3 and can correct all 1-bit error patterns. So the $(7, 4)$ code is called a 1-error-correcting code. Also a block code with t-error-correcting capability is capable of correcting many error patterns of $(t+1)$ or more. Some literatures settled for upper and/or lower bounds on $d_{\min}$ for linear block codes. One of them, referred to as the Plotkin bound, is given by [3]

$$d_{\min} \le \frac{n \cdot 2^{x-1}}{2^{x} - 1}. \tag{4.1}$$

When x is constant, the coding scheme with larger n has better error-correcting capability, but lower coding efficiency. For small codes, the best codes with larger or the largest $d_{\min}$ can be found by exhaustive search of generator matrices. A table of the tightest known bounds on $d_{\min}$ for small codes was listed in [28].

We assume that users from the same class experience the same BER performance and subscribe the same information rate, so they adopt the same coding scheme and

OVSF codes with the same capacity. Let k_i denote the number of active users in class i and N_i denote the spreading factor of their OVSF codes. If the received power is the same for all users, the BER performance experienced by a user in class i on Additive White Gaussian Noise (AWGN) channels after demodulation was derived in [13]. It is approximated by

$$P_e(N_i) = Q\left[\left(\frac{1}{2\gamma_b} + \frac{1}{3N_i}\left(\sum_j \frac{N_i}{N_j} k_j - 1\right)\right)^{-1/2}\right] \tag{4.2}$$

where $\gamma_b = E_b/N_o$ represents the signal to noise ratio per bit. According to this formula, P_e increases with the number of active users and decreases with the spreading factor of their codes.

Let p denote the probability that a symbol is received in error on a binary symmetric channel (BSC). The probability that an n-bit linear block code is decoded in error over a BSC is upper bounded by [3]

$$P_M \le \sum_{j=t+1}^{n} \binom{n}{j} p^j (1-p)^{n-j} . \tag{4.3}$$

The decoded bit error probability is approximated as [3]

$$P_B \approx \frac{1}{n} \sum_{j=t+1}^{n} j \binom{n}{j} p^j (1-p)^{n-j} . \tag{4.4}$$

We note that the decoded bit error probability is reduced if we adopt the coding scheme with more error-correcting capability. The cost is that fewer information bits are transmitted in a block.

OVSF codes support multiple rates by varying the spreading factor. The code with capacity $n \cdot R$ maps each data bit into N/n chips. It follows from (4.2) and (4.4) that in CDMA systems the decoded bit error probability of a user adopting the coding scheme $[n_i, x_i]$ with t_i-error-correcting capability on AWGN is given by

$$P_{Bi} \approx \frac{1}{n_i} \sum_{j=l_i+1}^{n_i} j \cdot \binom{n_i}{j} \cdot P_e^j \left(\frac{N}{n_i}\right) \cdot \left[1 - P_e\left(\frac{N}{n_i}\right)\right]^{n_i - j} . \tag{4.5}$$

Based on the assumption that resource is not fully utilized in most time, more checking bits are transmitted. The decoded bit error probability at the receiver may meet the BER requirement, even when more users are admitted for transmission and the multiple access interference is increased. More transmitted data may compensate the lower coding efficiency and higher system throughput may be achieved. According to the formula (4.2) and (4.4), k_i increases, P_e increases, and the decoded P_B increases when the coding schemes keep constant. It is not obvious how the decoded P_B varies with n_i, as n_i affects several factors in the formulas by different way. The numerical results are provided in simulation section.

4.1.2 Scheduling Algorithm Description

Suppose K classes of service with constant rates are supported in the system. For class i, let E_i denote the target BER requirement and $x_i \cdot R$ denote the subscribed rate, where x_i is a power of 2. The system throughput is given by

$$T = \sum_i k_i x_i R , \tag{4.6}$$

and the resource utilization is given by

$$L = \sum_i k_i n_i R . \tag{4.7}$$

Given the system capacity $C \cdot R$, we are looking for an optimal adaptation strategy to maximize the system throughput T with the constraints

$$\begin{cases} L \le C \cdot R; \\ P_{Bi} \le E_i . \end{cases} \tag{4.8}$$

Since the capacity of an OVSF code is a power of 2, we need also take this constraint into consideration. Each class searches all possible linear codes. The attempt begins with $n_i = x_i$, increases twice $n_i = 2n_i$, and ends in $n_i = C$ if all the resource can be allocated to this class. The decoding schemes together with the number of users admitted from each class are selected to allocated resource among classes.

Let us consider the case where all classes have the same BER requirement. We are to find what kind of class will be favored in allocation. Suppose additional resource is allocated to k number of users with coding scheme $[n, x]$. The interference introduced to an active user, say in class i, is increased to

$$P_e = Q\left[\left(\frac{1}{2\gamma_b} + \frac{n_i}{3N}\left(\sum_j \frac{n_j k_j}{n_i} - 1\right) + \frac{kn}{3N}\right)^{-1/2}\right]. \tag{4.9}$$

It indicates that no matter which class the additional resource is allocated to and no matter what the value of k and n is, the transmission error suffered by an active user is the same as long as the product $k \cdot n$ is the same. The same amount of resource can be allocated either to more users who adopt the coding scheme with smaller n or to fewer user who adopt the coding scheme with larger n. There is tradeoff between coding efficiency and error-correcting capability in coding schemes. The higher coding efficiency, the worse error-correcting capability. Among coding schemes with the same coding efficiency, the coding scheme with larger n transmits more information bits and possesses better error-correcting capability. Lower decoded P_B results from better error-correcting capability given the same transmission error P_e. Therefore when all classes adopt the coding schemes with the same coding efficiency, users who adopt the coding scheme with larger n achieve the same throughput but suffer lower decoded P_B. On the other hand, when all classes require the same BER performance, the class adopting the coding scheme with larger n possesses better error-correcting capability and more users in this class can be admitted for transmission. Such a class is to be favored in allocation to achieve higher system throughput. Another case is that a class has smaller information rate but lower BER requirement. More checking bits are required by this class compared with other classes with larger information rates or higher BER requirements. It is more obvious that such a class will be starved in allocation to maximize the system throughput.

Therefore a mechanism to guarantee the allocation fairness among classes is needed in both cases.

4.1.3 Simulation Results

In this section, we provide the numerical results how the decoded P_B varies with n. Figure 4.1 shows that the curves are not monotonous but zigzag given different γ_b. The capacity of an OVSF code is a power of 2 and the coding scheme takes only those points with value of a power of 2 at the n-coordinate in Figure 4.1. Therefore in OVSF system, the decoded P_B can be regarded to decrease monotonously against n when k keeps constant.

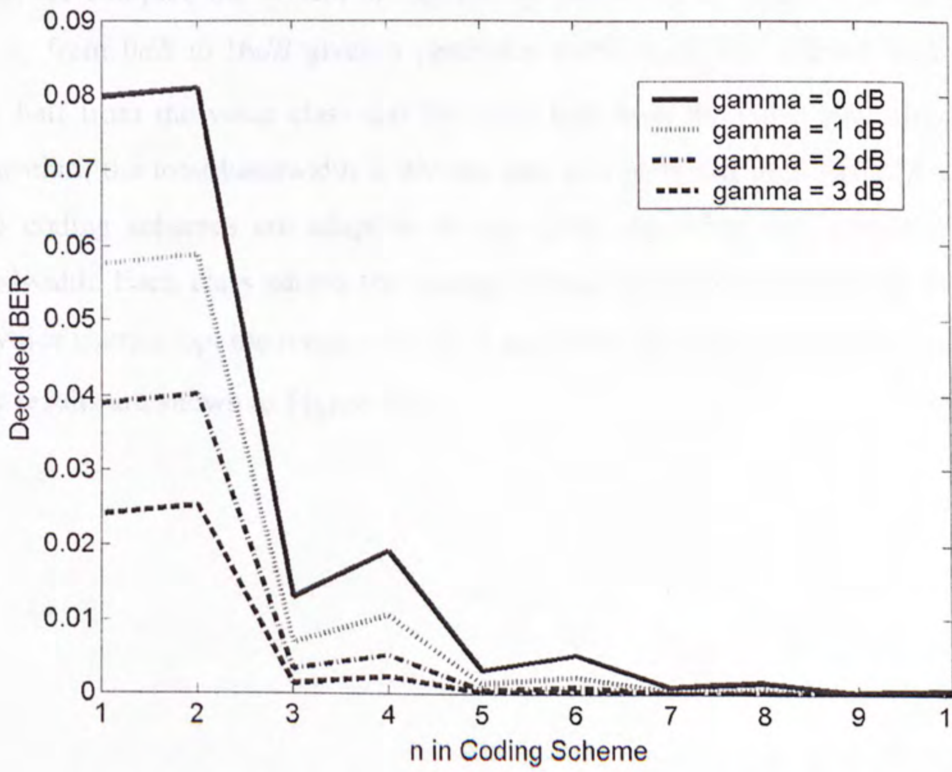

Figure 4.1 Decoded BER versus n in coding scheme

We compare the throughput performance of the proposed joint-design algorithm and those of two conventional scheduling algorithms, Coding Scheme Adaptive algorithm (CSA) and Codes Allocation Adaptive algorithm (CAA). In the CSA algorithm, the total bandwidth is partitioned among classes. Each class has exclusive use of its dedicated portion but adapts its coding scheme. In the CAA algorithm, all classes share the bandwidth but adopt a particular coding scheme each. Two classes of service are supported for simulation, voice service whose information bit rate is R and low resolution video service whose information bit rate is $4R$. The required BER performance is lower than 10^{-3} for voice users and lower than 10^{-5} for video users. The arrival of voice traffic and video traffic are both modeled as Poisson processes. The holding time of them are modeled as exponential distribution. Let the maximum spreading factor of the OVSF-code tree be 512 and the system capacity be $64R$.

First, we compare the system throughput by increasing the signal to noise ratio per bit γ_b from $0dB$ to $16dB$ given a particular traffic load. The offered load is $40R$, one half from the voice class and the other half from the video class. In the CSA algorithm, the total bandwidth is divided into two parts and each has $32R$ capacity. The coding schemes are adaptive. In the CAA algorithm, two classes share the bandwidth. Each class adopts the coding scheme with 1-error-correcting capability, so voice users adopt the linear code $[4, 1]$ and video users adopt the linear code $[8, 4]$. The results are shown as Figure 4.2.

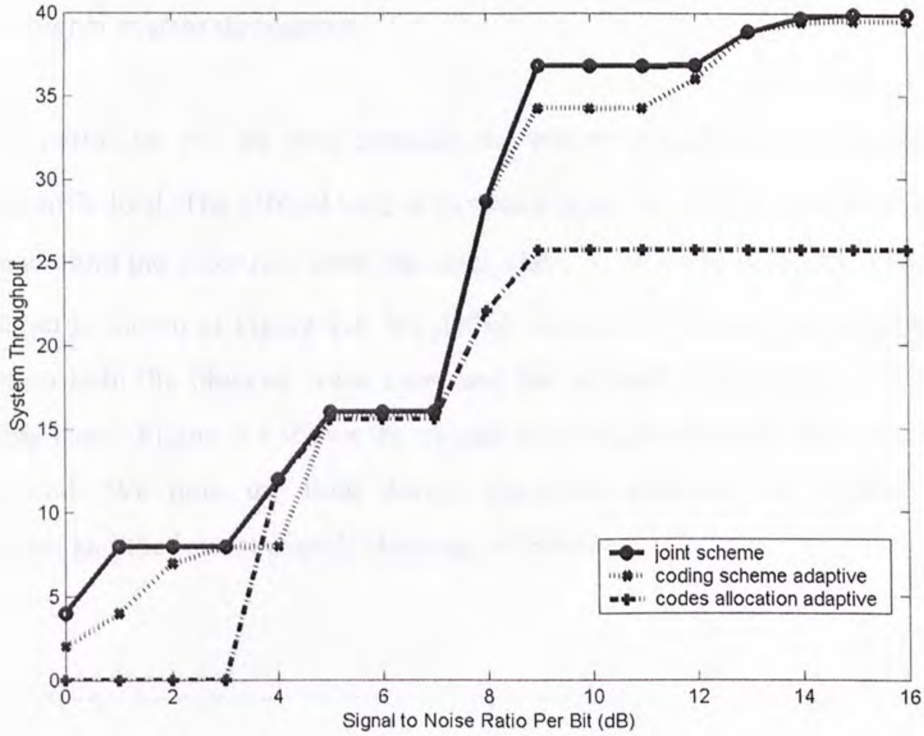

Figure 4.2 System throughput versus signal to noise ratio per bit

We note the joint design algorithm achieves the highest system throughput for all γ_b. When γ_b is low, there are many errors in transmission. For CAA users, their coding scheme cannot correct all transmission errors and the decoded bit error probability is greater than the requirement. Therefore, few users are admitted for transmission in the CAA algorithm. Video users have lower BER requirement and require coding schemes with higher error-correcting capability but lower coding efficiency. Voice users have higher BER requirement and their coding schemes are more efficient. More voice users are admitted for transmission in the joint design algorithm to improve the system throughput, comparing with those admitted in CSA algorithm due to bandwidth partition. When γ_b is high, the coding scheme with low error-correcting capability is required. For CAA users, more redundant bits are transmitted and the throughput of information bits is limited. Voice users adopt the coding scheme with few checking bits and do not exhaust their dedicated bandwidth. The leftover bandwidth is borrowed by video users in the joint design algorithm to achieve higher system throughput.

Given a particular γ_b, we then compare the system throughput by increasing the offered traffic load. The offered load is increased from $5R$ to $60R$, one half from the voice class and the other half from the video class. γ_b is set to be $10dB$. The system throughput is shown as Figure 4.3. We define the overall blocking probability as the proportion both the blocked voice users and the blocked video users over the total incoming users. Figure 4.4 shows the overall blocking probability versus the offered traffic load. We note the joint design algorithm achieves the highest system throughput and the lowest overall blocking probability.

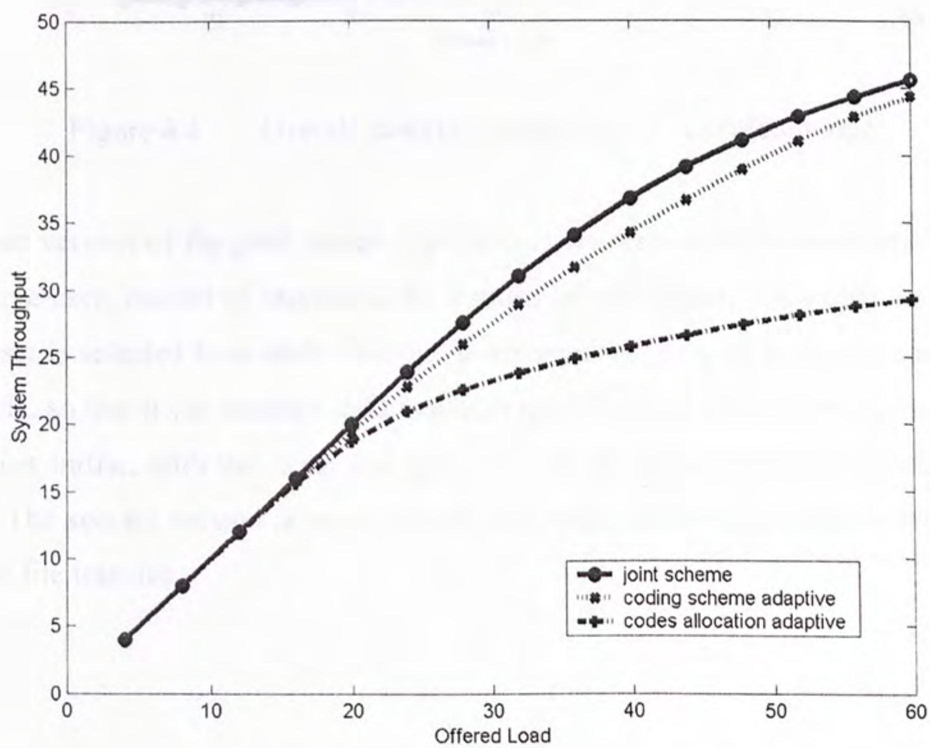

Figure 4.3 System throughput versus offered load

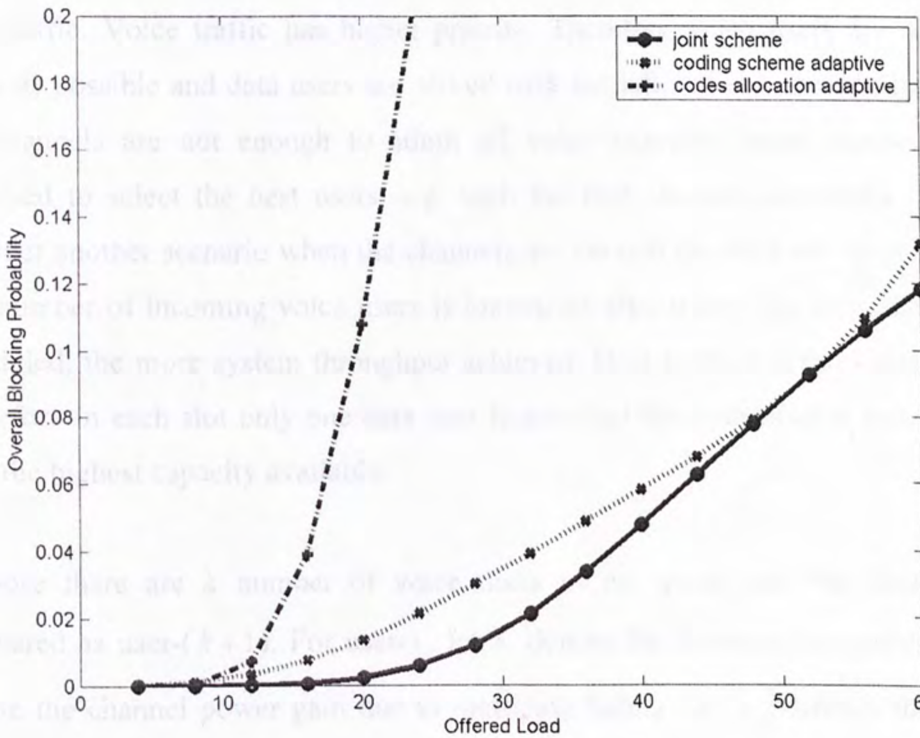

Figure 4.4 Overall blocking probability versus offered load

Another version of the joint design algorithm is to increase the transmission rate of an active user, instead of increasing the number of active users form each class. The best user is selected from each class and is assigned a code with as much capacity as possible, so that it can transmit data in a high rate. The best user can be the user with the most traffic, with the latest due time, or with the most credit tokens, etc. in its class. The second version is more suitable for bursty traffic with variable rate, such as date file transfer.

4.2 Combine Power Adaptation and Error-Control Coding Scheme

4.2.1 System Model

We consider the system where two kinds of traffic are scheduled, voice traffic and data traffic. Voice traffic has higher priority. Therefore voice users are served as many as possible and data users are served with the leftover resource. In the case if the channels are not enough to admit all voice requests, some strategies were proposed to select the best users, e.g. with the best channel conditions [30]. We consider another scenario when the channels are enough to admit all voice requests. The number of incoming voice users is known on allocation. The more data traffic scheduled, the more system throughput achieved. Data traffic tolerates some delay. Therefore in each slot only one data user is admitted for transmission using a code with the highest capacity available.

Suppose there are k number of voice users in the queue and the data user is numbered as user-($k+1$). For user-i, let S_i denote the transmission power and g_i denote the channel power gain due to multipath fading. $[n_i, x_i]$ denotes the coding scheme, where $n_i \cdot R$ corresponds to the channel coding rate and $x_i \cdot R$ corresponds to the source coding rate. The bit energy-to-equivalent noise spectral density ratio E_b/N_e on the additive white Gaussian noise (AWGN) with MAI is given by [18]

$$\frac{E_b}{N_e} = \frac{S_i \cdot g_i \cdot \dfrac{N}{n_i}}{\dfrac{2}{3}\sum_{j \neq i} S_j \cdot g_j + \dfrac{N_o}{T_c}}, \tag{4.10}$$

where N_o is the one-sided power spectral density of AWGN and T_c is the chip duration. The decoded bit error rate after correcting is denoted as P_b, which depends on E_b/N_e and the coding rate $r = x/n$, and is expressed as [31]

$$P_b = Q\left(\sqrt{W \cdot d(n, r) \cdot \frac{E_b}{N_e}}\right), \tag{4.11}$$

where W is a factor depending on the error-control coding scheme adopted and d is the minimum distance depending on n and r. When $r = 1$ and no checking bits are transmitted, P_b is simplified as

$$P_b = Q\left(\sqrt{2 \cdot \frac{E_b}{N_e}}\right). \tag{4.12}$$

Since voice traffic has constant rate, we assume all voice users adopt the same source coding rate but vary the channel coding rate, which in turn affects the error-correcting capability. The source coding rate can also be adaptive as proposed in AMR [32]. We here focus on the OVSF codes allocation and power adaptation. The system throughput is given by

$$T = \sum_{i=1}^{k} x_i + x_{k+1} = k + x_{k+1}, \tag{4.13}$$

assuming $x_i = 1$ for voice users. We are to maximize T, or x_{k+1} for k is given on allocation, with the constraints of BER performance, maximum transmitted power, and the system capacity.

Let BER_v and BER_d denote the BER requirements of voice users and data users. Let $S_{\max}$ be the maximum transmission power of each user. The constraints are expressed as

$$\begin{cases} P_{bi} \le BER_v, & i = 1, \ldots, k; \\ P_{bi} \le BER_d, & i = k+1; \\ 0 \le S_i \le S_{\max}, & i = 1, \ldots, k+1; \\ \sum_{i=1}^{k+1} n_i \le C. \end{cases} \tag{4.14}$$

where P_{bi} is given by formula (4.11). For OVSF codes, the capacity of a code is a power of 2, so n_i could only take some discrete values.

4.2.2 Scheduling Algorithm Description

We are to find the parameters S_i and n_i for $(k+1)$ number of users. The channel coding rate n and the coding efficiency r are discrete, while the transmission power

S is continuous. Since formula (4.11) is not linear when S_i and n_i are both variables, S_i and n_i cannot be optimized by linear programming simultaneously. We therefore optimize them in two loops. In external loop, we exhaustedly search all possible levels of n and r. With each possible combination of n_i's and r_i's, we find S_i's that satisfy the constraints in internal loop. The combination that achieves the maximum throughput is selected. We set three levels of channel coding rate for voice users. The data user attempts all possible channel coding rates up to the leftover capacity. We also set three levels of coding efficiency for data users. The external and internal loops are outlined as follows.

External loop:

To maximize $x_{k+1} = r_{k+1} \cdot n_{k+1}$, exhaust search all possible levels.

$$\begin{cases} n_i = \{1, 2, 4\}, \quad i = 1, \mathrm{K}, k+1; \\ r_i = \{3/4, 2/3, 1/2\}, \quad i = k+1; \\ \sum_{i=1}^{k+1} n_i \le C. \end{cases} \tag{4.15}$$

Internal loop:

To find S_i's that satisfy the constraints.

$$\begin{cases} P_{bi}(S_i) \le BER_i, \quad i = 1, \mathrm{K}, k+1; \\ 0 \le S_i \le S_{\max}, \quad i = 1, \mathrm{K}, k+1. \end{cases} \tag{4.16}$$

In the internal loop, given n_i's, S_i's can be found by linear programming [33]. In the external loop, voice users attempt all possible channel coding rates and data user attempts all possible coding efficiency. Let $O(S)$ be the complexity of finding S_i's in the internal loop given n_i's. The complexity of the whole procedure must be larger than $3^{k+1} \cdot O(S)$. We propose another method to find n_i's with less complexity. The solution, of course, is not optimal.

Intuitively, voice users with similar channel conditions should adopt the coding schemes with the same error-correcting capability and/or the same transmission power. The difficulty is how to find these users and the proper coding schemes they adopt. A heuristic algorithm is that voice users are divided into groups according to their channel conditions and users in the same group adopt the same coding scheme. Transmission power is still adaptive for each user. Three kinds of coding schemes are provided in the algorithm corresponding to $n = 1, 2, 4$. Users are divided into groups according to their channel condition by Clustering Algorithm, e.g. K-Means method [34]. We detail the algorithm as follows.

1. Voice users are divided into three groups according to their channel conditions.
2. Voice users in each group attempt all possible values of n.
3. The data user attempts all possible n, which is no greater than

$$power\left(2, \left\lfloor \log_2\left(C - \sum_{i=1}^{k} n_i\right)\right\rfloor\right). \tag{4.17}$$

4. The data user attempts all possible coding efficiency r
5. Find S_i's that satisfy (4.16) by linear programming.
6. IF S_i's are found, calculate the system throughput.
7. Return the maximum system throughput.

The number of groups can be adjusted to find more optimal solution at the cost of complexity.

To select the data user for transmission in each slot, we combine the channel conditions in PHY layer and the queue information in MAC layer. Let D denote the due time and J denote the time jitter of data users. Users are given higher priority in the range of $[D - J, D]$; otherwise users with better channel conditions are preferred. It is the tradeoff between fair allocation and system throughput. The algorithm to select the data user is detailed as follows. The current slot is number as 0.

1. FOR $i = 0$ TO $i = J$, repeat

IF there are users with due time in slot-i, THEN

IF there is more than one user, THEN

The user with the best channel condition is preferred.

ELSE

The user with due time in slot-i is admitted.

Break.

2. The user with best channel condition is selected from the users with due time after slot-J.

4.2.3 Simulation Results

We will compare the performance achived by different algorithms. We have made the assumption that voice users are given higher priority to be served and the bandwidth is enough to admit all voice requests. Therefore, the difference is the system throughput achieved by different algorithms, which is affected mostly by the amount of data traffic admitted for transmission.

The channel condition is modeled as Rayleigh distribution [36]

$$p(g) = \begin{cases} \dfrac{g}{\sigma^2} \exp\left(-\dfrac{g^2}{2\sigma^2}\right), & 0 \le g \le \infty \\ 0, & g < 0 \end{cases}, \tag{4.18}$$

where σ is the rms value of the received voltage signal before envelope detection, and σ^2 is the time-average power of the received signal before envelope detection.

We compare the system throughput achieved by three algorithms for different channel conditions and for different real time traffic load. The first algorithm is our heuristic algorithm. The other algorithms are the traditional versions. Voice users adopt the constant coding scheme but vary the transmission power, or voice users adapt the coding scheme but transmit in constant power. The BER performance is lower than 10^{-3} for voice users and lower than 10^{-5} for data users.

The voice traffic has Poisson arrivals and exponential duration. Voice users have higher priority to be served. Since data traffic can tolerate some delay and is bursty in nature, we assume there is always data traffic in the queue. We are to find the maximum system throughput can be achieved. The system capacity is set to be $64R$ and the maximum transmission power ranges from 1 to 10. In the joint design algorithm, users adapt their coding schemes and transmission power according to their channel condition. In coding scheme constant algorithm, voice users all adopt the same coding scheme $[2,1]$, but adapt their transmission power to satisfy the BER requirements with different channel conditions. Data user in this scheme attempts all possible coding schemes to transmit information as much as possible. In power constant algorithm, all users adopt the same transmission power $S_{\max}/2$, but adapt their coding schemes.

In Figure 4.5, we increase the parameter σ of Rayleigh distribution, or in other words, improve the channel condition. The voice traffic load is set to be $12R$. We find that the joint design algorithm achieves the highest system throughput. The coding scheme constant algorithm achieves higher throughput than the power constant algorithm. This shows the disadvantage of power constant algorithm. The same power is used for each user no matter its channel condition. Large interference is translated to other users, leading to a capacity reduction. We also note with the improvement of channel condition, the difference between joint design algorithm and coding scheme constant algorithm is increased. It is because that the checking bits in coding scheme become redundant when the channel condition improves.

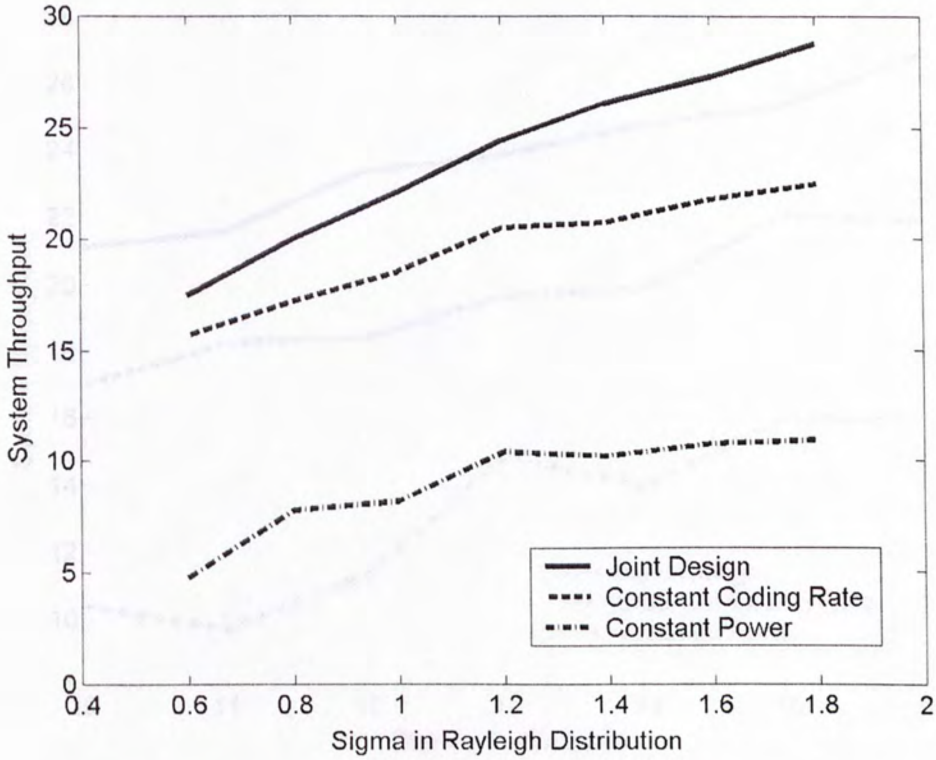

Figure 4.5 System throughput versus channel condition

We then compare the system throughput by increasing the traffic load. The channel condition is modeled as Rayleigh distribution with $\sigma = 1$. The voice traffic is increased from $10R$ to $16R$. The system throughput achieved by the algorithms is shown in Figure 4.6. The joint design algorithm achieves the highest system throughput and the power constant algorithm achieves the lowest system throughput. The voice traffic load is low to meet our assumption that the voice requests can all be served.

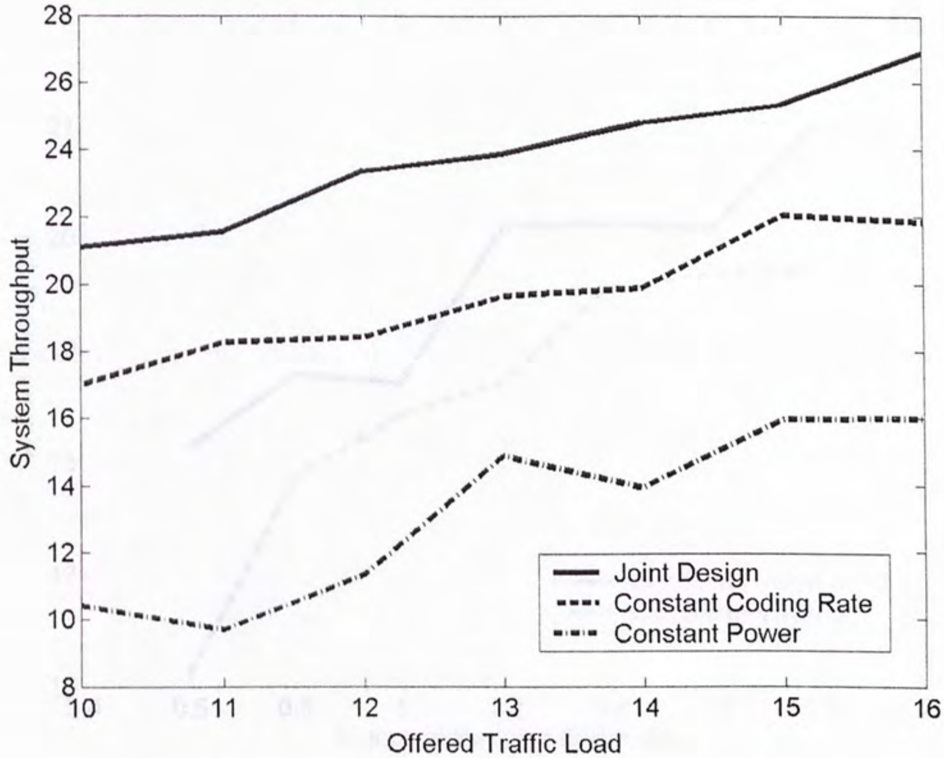

Figure 4.6 System throughput versus offered traffic load

We simulate the scenario when data packets are not infinite. Data users in the range of $[D-J, D]$ are given higher priority; otherwise, users with better channel conditions are preferred. In each slot, number as slot-0, one data user arrives with due time D uniformly distributed in the slot-6 to slot-9. The size of data packets has geometric distribution with mean value of 7 or 10 in unit of R. We set the time jitter $J=3$. The system throughput is show as Figure 4.7 and the blocking probability of data traffic is shown as Figure 4.8.

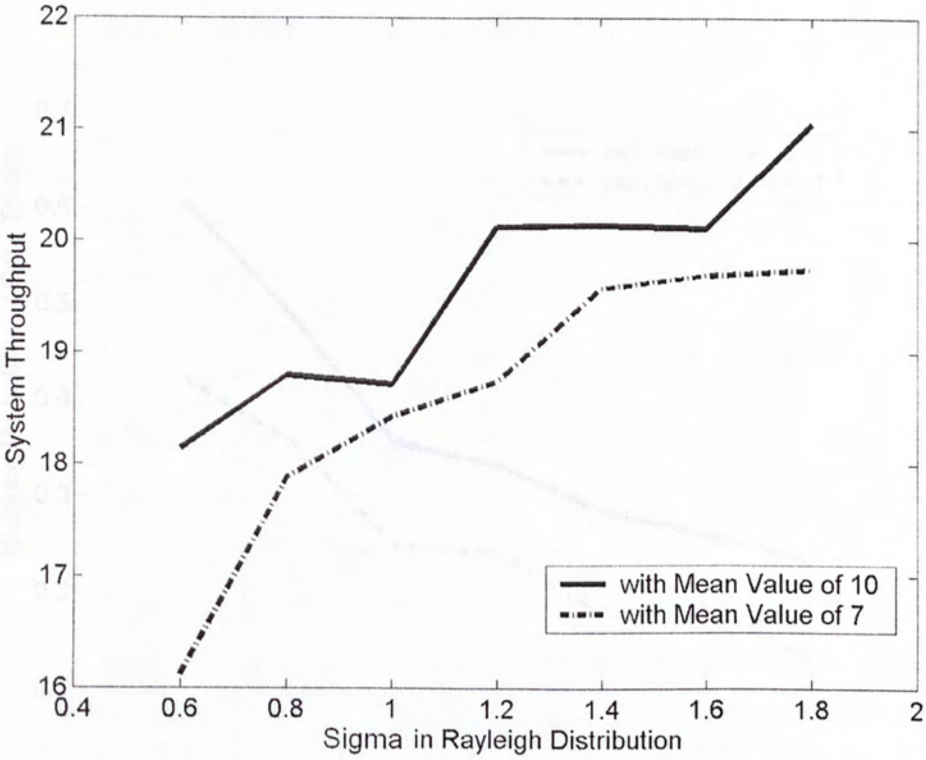

Figure 4.7 System throughput with different data traffic load

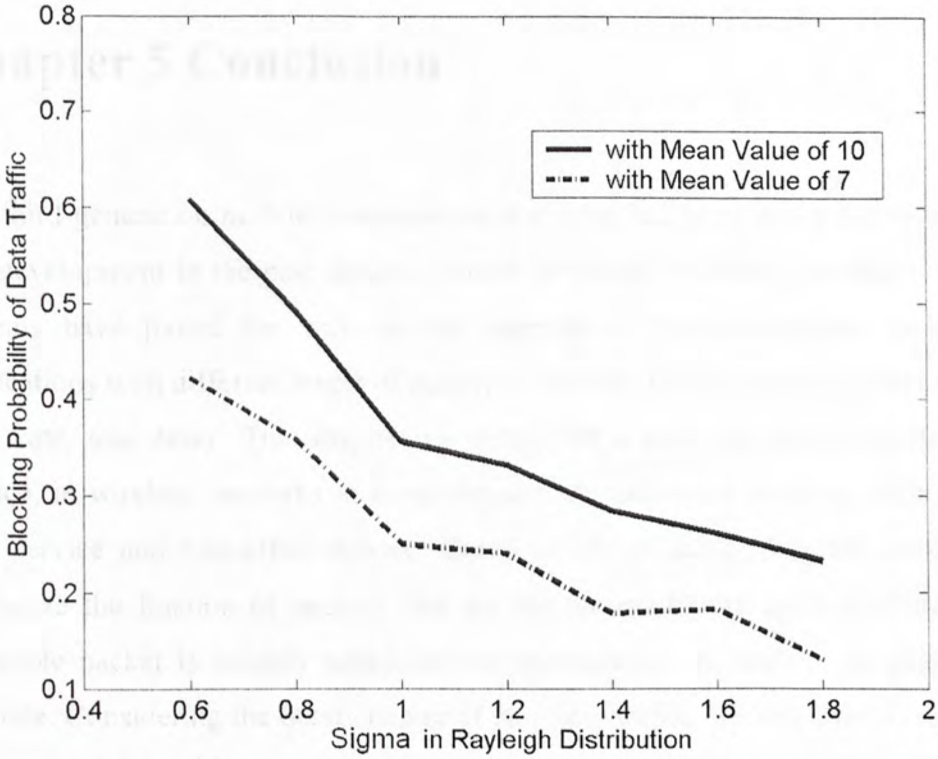

Figure 4.8 Blocking probability of data traffic with different data traffic load

When larger data traffic load is offered, the system achieves higher system throughput but larger blocking probability of data traffic. When data traffic load has mean value of 7, system throughput does not increase much after γ reaches a certain value. It is because there is not sufficient data traffic in the queue although bandwidth is available.

Chapter 5 Conclusion

The third-generation mobile communication system has been under active research and development in the past decade. Recent proposals for third-generation wireless systems have paved the way for the support of various wireless multimedia applications with different levels of quality of service (QoS) in terms of data rate, bit error rate, and delay. This inspires us to support a new service, called delayable service, in wireless networks in concurrence with traditional services, such as real-time service and best-effort service. Based on the intuitive idea that, in order to maximize the fraction of packets that are transmitted before their deadlines, each delayable packet is initially scheduled for transmission as close to its deadline as possible. Considering the bursty nature of real-time traffic, we proposed to rearrange the accepted delayable packets to release similar number of channels in the following slots. More real-time connections are admitted consequently. Simulation results show that the bandwidth utilization is increased without affecting service requirements and the blocking probability of real-time service is decreased with the rearrangement of delayable traffic.

When the transmitter is provided with channel fading information, the transmission schemes are adapted to use the channels more efficiently. However it is too complex to optimize all these parameters simultaneously. We propose some algorithms to find sub-optimal solutions for practice. We combine the number of users admitted and their error-control coding schemes, based on the assumption that power adaptation is used only to reduce the near-far effect in CDMA systems. We combine transmission power and error-control coding schemes, based on the assumption that all voice users can be accommodated and only one data user is admitted for transmission in each slot. We further reduce the complexity by dividing users into groups according to their channel conditions, based on the idea that users in the same group experience similar channel fading and adopt the same transmission schemes. We select the data user for transmission in each slot based on the channel conditions in PHY layer and queue information in MAC layer crossly. It is the tradeoff between system throughput and fair allocation among flows. Simulation results show that higher system throughput is achieved by joint design algorithms.

Bibliography

[1] http://www.umtsworld.com

[2] T.-S. P. Yum and M. Chen, "Dynamic Channel Assignment in Integrated Services Cable Networks," *IEEE Transactions on Communications*, vol. 42, pp. 2023-2027, Feb./Mar./Apr., 1994.

[3] S. Lin and D. Costello, Error Control Coding: Fundamentals and Applications. Englewood Cliffs, N. J.: Prentice Hall, 1983.

[4] 3GPP TS 25.213 (V4.2.0), "Spreading and modulation (FDD)," Technical Specification (Release 4), Technical Specification Group Radio Access Network, 3GPP, Dec. 2001.

[5] 3GPP TS 25.223 (V4.2.0), "Spreading and modulation (TDD)," Technical Specification (Release 4), Technical Specification Group Radio Access Network, 3GPP, Dec. 2001.

[6] E. Dahlman and K. Jamal, "Wide-band services in a DS-CDMA based FPLMTS system," in Proc. IEEE Vehicular Technology Conf., vol. 3, Apr. 1996, pp. 1656-1660.

[7] F. Adachi, M. Sawahashi, and K. Okawa, "Tree-structured generation of orthogonal spreading codes with different lengths for forward link of DS-CDMA mobile radio," Electronics Letters, 33:27-28, Jan. 1997.

[8] T. Minn, and K.-Y. Siu, "Dynamic assignment of orthogonal variable spreading factor codes in WCDMA," IEEE Journal on Selected Area in Communications, 18:1429-1440, Aug. 2000.

[9] N. Passas, S. Paskalis, D. Vali, and L. Merakos, "Quality-of-Service-Oriented Medium Access Control for Wireless ATM Networks," *IEEE Communications Magazine*, vol. 35, pp. 42-50, Nov. 1997.

[10] J. F. Hayes, "Adaptive feedback communications," IEEE Trans. Commun., vol. COM-16, pp. 29-34, Feb. 1968.

[11] J. K. Cavers, "Variable-rate transmission for Rayleigh fading channels," IEEE Trans. Commun., vol. COM-20, pp. 15-22, Feb. 1972.

[12] W. T. Webb and R. Steele, "Variable rate QAM for mobile radio," IEEE Trans. Commun., vol. 43, pp. 2223-2230, July 1995.

[13] S. M. Alamouti and S. Kallel, "Adaptive trellis-coded multiple phase shift keying for Rayleigh fading channels," IEEE Trans. Commun., vol. 42, pp. 2305-2314, June 1994.

[14] S.-J. Oh and K. M. Wasserman, "Dynamic spreading gain control in multi-service CDMA networks," IEEE J. Select. Areas Commun., vol. 17, pp. 918-927, May 1999.

[15] ---, "Optimality of greedy power control in DS-CDMA mobile networks," Proc. ACM/IEEE 5th Annual Int. Conf. Mobile Comput. and Network. (MobiCom'99), 1999.

[16] H. D. Schotten, H. E. Boll, and A. Busboom, "Adaptive multi-code CDMA systems for variable data rates," in Proc. ICPWC, 97, pp.334-337.

[17] B. Hashem and E. Sousa, "A combined power/rate control scheme for data transmission over a DS/CDMA system," in Proc. IEEE Vehicular Technology Conf. VTC'98, pp. 1096-1100.

[18] S. W. Kim and Y. H. Lee, "Combined rate and power adaptation in DS/CDMA communications over Nakagami fading channels," IEEE Trans. Commun., vol. 48, pp. 162-168, Jan. 2000.

[19] R. Assarut, K. Kawanishi, U. Yamamoto, Y. Onozato, and M. Matsushita, "Region division assignment of orthogonal variable spreading factor codes in W-CDMA," in Proc. IEEE Veh. Technol. Conf., vol. 3, Fall 2001, pp. 1884-1888.

[20] Y. Yang and T.-S. P. Yum, "Maximally flexible assignment of orthogonal variable spreading factor codes for multirate traffic," IEEE Transactions on Wireless Communications, 3:781-792, May, 2004.

[21] M. Dell'Amico, M.L. Merani, and F. Maffioli, "Efficient algorithms for the assignment of OVSF codes in wideband CDMA," in Proc. IEEE Int. Conf. Commun. ICC, 5:3055-3060, 2002.

[22] R. Fantacci and S. Nannicini, "Multiple access protocol for integration of variable bit rate multimedia traffic in UMTS/IMT-2000 based on wideband CDMA," IEEE J. Select. Areas Commun., vol. 18, pp. 1441-1454, Aug. 2000.

[23] C.E. Fossa Jr. and N. J. Davis IV, "Dynamic code assignment improves channel utilization for bursty traffic in third-generation wireless networks," in Proc. IEEE Int. Conf. Commun. ICC, vol. 5, 2002, pp. 3061-3065.

[24] W. T. Chen, Y. P. Wu, and H. C. Hsiao, "A novel code assignment scheme for W-CDMA systems," in Proc. IEEE Vehicular Technology Conf., vol. 2, Fall 2001, pp. 1182-1186.

[25] J. S. Kaufman, "Blocking in a Shared Resource Environment," *IEEE Transactions on Communications*, vol. COM-29, pp. 1474-1481, Oct. 1981.

[26] L. Kleinrock, *Queueing Systems---Volume II: Computer Applications*. New York: Wiley, 1975.

[27] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo-codes," In Proc. IEEE ICC'93, pp. 1064-1070, Geneva, Switzerland, May 1993.

[28] T. Verhoeff, "An updated table of minimum-distance bounds for binary linear codes," IEEE Transactions on Information Theory, IT-33(5):665-680, Sept. 1987.

[29] T. Ottosson and A. Svensson, "Multi-rate schemes in DS/CDMA systems," In Proc. IEEE VTC'95, pp. 1006-1010, Chicago, USA, July 1995.

[30] S. A. Jafar and A. Goldsmith, "Adaptive multirate CDMA for uplink throughput maximization," IEEE Transactions on Wireless Communications, vol. 2, no. 2, pp. 218-228, Mar. 2003.

[31] J. B. Cain, C. C. Clark, JR., and J. M. Geist, "Punctured convolutional codes rate $(n-1)/n$ and simplified maximum likelihood decoding," IEEE Transactions on Information Theory, vol. IT-25, no. 1, Jan. 1979.

[32] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," IEEE Transactions on Speech and Audio Processing, vol. 10, Issue: 8, pp. 620-636, Nov. 2002.

[33] G. B. Dantzig, *Linear programming*, New York : Springer, 1997.

[34] H. Spath, *Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples*, Halsted Press, New York, 1985.

[35] J. G. Proakis, Digital Communications: McGraw-Hill, 1995.

[36] T. S. Rappaport, Wireless Communications: Principles and Practice. Englewood Cliffs, NJ: Prentice-Hall, 1996.

[37] E. Dahlman, B. Gudmundson, M. Nilsson, and J. Skold, "UMTS/IMT-2000 based on wideband CDMA," IEEE Communications Magazine, 36:70-80, Sept. 1998.