CHAPTER 3

AN EFFECTIVE CAPACITY BASED CROSS-LAYER SCHEDULING ALGORITHM

3.1 INTRODUCTION

Quality of Service (QoS) guarantee plays a critically important role in mobile wireless networks. Depending on their distinct QoS requirements, differentiated mobile users are expected to tolerate different levels of delay for their service satisfactions. For instance, non-real-time services such as data dissemination aim at maximizing the throughput with a loose delay constraint. In contrast, for real-time services like multimedia video conference, the key QoS metric is to ensure a stringent delay-bound, rather than to achieve high spectral efficiency. There also exist some services falling in between, e.g., paging and interactive web surfing, which are delay-sensitive but whose delay requirements are not as stringent as those of real-time applications. The diverse mobile users impose totally different and sometimes even conflicting delay-QoS constraints, which impose great challenges to the design of mobile wireless networks.

Unlike its wired counterparts, supporting diverse delay QoS in wireless environment is much more challenging since the wireless channel has a significant impact on network performance. In particular, a deterministic delay-bound QoS guarantee over wireless networks is practically infeasible due to the time-varying nature of fading channels. Alternatively, a more practical solution is to provide the statistical QoS guarantees, where we guarantee the given delay-bound with a small violation probability.

For wireless QoS guarantees, Link Adaptation (LA) techniques have been widely considered as the key solution to overcome the impact of the wireless channel. At the physical layer, the scarcest resources are power and spectral bandwidth. The LA techniques such as adaptive modulation and power control, adopted by Choi and Shin (1999), are developed to enhance the spectral efficiency while maintaining a certain target error metric like bounded-delay, instead of high spectral efficiency.

QoS provisioning in wireless networks has been widely studied from different perspectives, such as packet scheduling, admission control, traffic specifications, resource reservations, etc. The authors Choi and Shin (1999) investigated the real time and non-real-time QoS provisioning for Code Division Multiple Access (CDMA)-based wireless networks. Several architectures/algorithms were discussed by Shakkottai et al (2003), Fattah and Leung (2002) and Razavilar et al (2002), for either implicit or explicit QoS provisioning. The integrated Finite-State Markov Chain (FSMC) model with Adaptive Modulation and Coding (AMC) and then jointly considered the physical layer channel and data link layer queuing characteristics were studied in Liu et al (2005, 2006). Their idea of resource allocation is to calculate the reserved bandwidth for each user by appropriate admission control and scheduling. This scheme is developed across the physical layer and data link layer and is thus capable of characterizing the impact of physical layer variation on the data link layer QoS performance. However, the main QoS requirement addressed is the average delay of the wireless transmission that does not effectively support the real-time multimedia services, where the key QoS metric is the bounded delay. Therefore, to support the real-time wireless multimedia QoS, we need to consider the LA techniques not only at the physical layer, but also at the upper-protocol-layers such as data-link layer when designing the wireless networks. To achieve this goal, the cross-layer model based adaptive resource allocation scheme is developed in this chapter to support the real-time multimedia QoS in the downlink heterogeneous mobile wireless networks.

A powerful concept termed “effective capacity” is proposed by Wu and Negi (2003). This concept turns out to be the dual problem of the so-called “effective bandwidth”, which has been extensively studied in the early 90’s in the contexts of wired Asynchronous Transfer Mode (ATM) networks. The effective capacity and effective bandwidth enable us to analyse the statistical delay-bound violation and buffer-overflow probabilities, which are very important for multimedia wireless networks. A set of resource-allocation schemes for statistical QoS guarantees in wireless networks is proposed by Wu and Negi (2004, 2005). The key techniques used are the integration of effective capacity with multiuser diversity, such that the scheme not only provides the statistical QoS for different mobile users, but also increases the throughput of total wireless network. However, the effective capacity approach has not been explored in cross-layer modelling and design for adaptive resource allocation and QoS guarantees in mobile wireless networks.

To overcome the aforementioned problems, in this chapter we propose an effective capacity based cross-layer scheduling to allocate resources adaptively for downlink heterogeneous mobile wireless networks. Based on the application of the effective capacity method employed in Tang and Zhang (2007), the system resources are allocated according to the heterogeneous fading channel statistics, the diverse QoS requirements, and different traffic characteristics.

Specifically, our proposed scheme adaptively assigns time slots for real-time mobile users in a dynamic time division multiple access (TDMA) mode to guarantee the bounded delays. The admission-control and power / time slot allocation conditions are analytically derived to guarantee the statistical delay-bound for real-time mobile users. Multiuser diversity scheme is not employed because of the following reasons: In a centralized heterogeneous multiuser network, the multiuser diversity will cause the serious fairness problem, i.e., the users with good channels may occupy most of the resources, while the users with poor channels may hardly have opportunity for information transmission. This would result in large queuing delay and hence the user's delay-bound QoS cannot be guaranteed. Whereas, the advantages of multiuser diversity only contribute to a small portion of mobile users whose channel quality is good, which may not lead to a significant QoS performance improvement from the entire network perspectives. The multiuser diversity under the proportional fairness constraint has been dealt in Viswanath et al (2002). However, this scheme can only support loose delay-bound QoS requirements, and is also not suitable for real-time multimedia services.

When designing the adaptive resource-allocation algorithm, impact of physical layer on the statistical QoS provisioning performance is discussed. Specifically, the influence of adaptive CSI feedback delay on our proposed scheme is studied. Based on the previous work proposed by Tang and Zhang (2007), effective capacity based cross-layer scheduling is applied in this work for heterogeneous mobile users and its performance is compared with constant power scheduling scheme. The numerical and simulation results show that our proposed effective capacity based cross-layer scheduling has significant advantages over the constant power scheme in terms of QoS guarantees. Moreover, our effective capacity based adaptive resource-allocation algorithm can efficiently support the QoS requirements for diverse real-time mobile users.

This chapter is organized as follows: Section 3.1 describes our system model of cross-layer scheduling. Section 3.2 introduces the fundamental concepts of effective capacity. The effective capacity based cross-layer scheduling algorithm is developed in section 3.3. Section 3.4 discusses the numerical and simulation results. The chapter summary is provided in section 3.5.

3.2 SYSTEM MODEL

The system model of cross-layer scheduling is shown in Figure 3.1. Single-Input-Single-Output (SISO) antenna system with the downlink transmission from the base station to the mobile users is concentrated. A fixed average transmission power P of the base station and K total number of heterogeneous mobile users are assumed, i.e., they may experience different fading conditions and demand different QoS requirements.

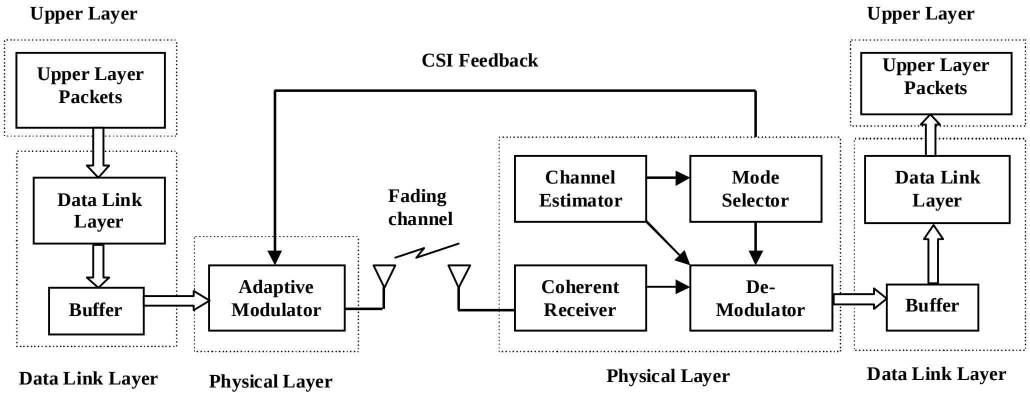

Figure 3.1 System Model of Cross-Layer Scheduling

As shown in Figure 3.1, the upper protocol layer packets are first divided into a number of frames at data link layer. The frames are stored at the transmitter infinite buffer and then split into bit-streams at physical layer. The adaptive modulation and coding is employed to enhance the system performance. The reverse operations are executed at the receiver side. Also, the CSI is estimated at the receiver and fed back to the transmitter for adaptive modulation and coding.

3.2.1 Data Link Layer Frame Structure

The data link layer frame structure of our proposed system is shown in Figure 3.2. In our system, each frame at data-link layer consists of L number of time slots. The time-duration of each frame is denoted by T_f.

Due to the employment of adaptive modulation, the number of bits per frame varies depending on each user's selected modulation modes. Within the frame duration T_f, the system runs in a dynamic TDMA mode. The k^{th} mobile user is assigned with $L^{(k)}$ number of time slots. The number $L^{(k)}$ is determined by the *k*th mobile user's QoS requirement.

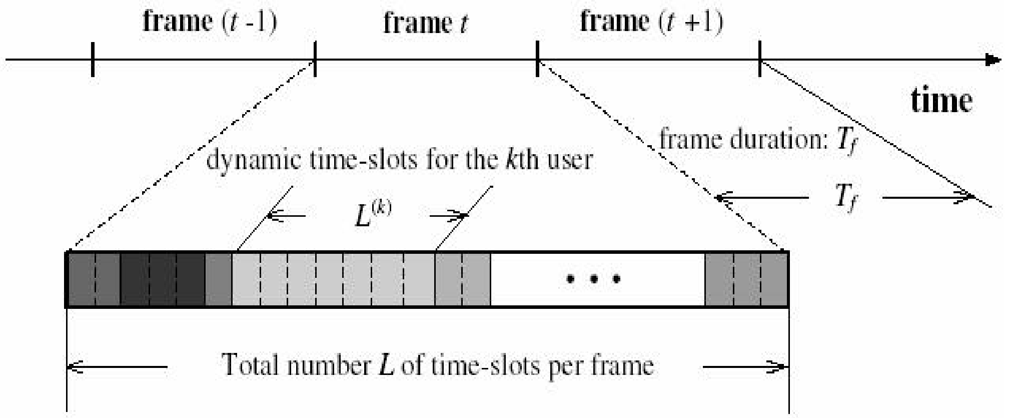

Figure 3.2 Data Link Layer Frame Structure

3.2.2 Channel Model

The Nakagami-*m* fading channel model is chosen here because it applies to a large class of fading channels. Simon and Alouini (2005) stated that this model is very general and often best fits the land-mobile and indoor mobile multipath propagations. As the fading parameter *m* varies, where

$m \in [1/2, +\infty)$, the model spans a wide range of fading environments, including one-sided Gaussian fading channel (*m* = 1/2, the worst fading case), the Rayleigh fading channel (*m* = 1), the precise approximations of Rician and lognormal fading channels (*m* > 1), and the conventional Gaussian channel (*m* = ∞, no fading). The channel is assumed to be invariant within a frame's time-duration T_f , but varies from one frame to another. Further, it is also assumed that the CSI is perfectly estimated at the receiver and reliably fed back to the transmitter with a time-delay denoted by τ . When the parameter τ = 0, it implies perfect CSI feedback. In practice, the CSI feedback delay is unavoidable in most situations. Due to imperfect CSI feedback, the normalized feedback delay $f_d\tau$ increases. Hence the number of time-slots requirement also increases to maintain the same statistical QoS requirements.

The probability density function (pdf) of the Signal-to-Noise-Ratio (SNR), denoted by $p_\gamma(\gamma)$, can be expressed as

$$p_\gamma(\gamma) = \frac{m^m \gamma^{m-1}}{\bar{\gamma}\Gamma(m)} \exp\left(-\frac{m\gamma}{\bar{\gamma}}\right) \tag{3.1}$$

where $\bar{\gamma} = \bar{P}\, E\{\alpha^2\} / (N_0 B)$ denotes the average received SNR of the combined signal, $\Gamma(m) = \int_0^\infty t^{m-1} e^{-t} dt$ is the complete Gamma function and *m* is the fading parameter of Nakagami-*m* distribution (where *m* value is chosen >1/2), $E\{\alpha^2\}$is the average path-gain of the Nakagami fading channel and N_0 is the single sided power spectral density (PSD) of the complex Additive White Gaussian Noise (AWGN).

3.2.3 Adaptive Modulation

Adaptive modulation is an efficient LA technique to improve the spectral-efficiency at physical layer. Following the work of Liu et al (2006), the specific modulation and coding modes for the AMC scheme as per IEEE 802.16 standard setting is given in Figure 3.3. The specific modulation modes for the adaptive-modulation scheme are constructed as follows: The entire SNR range is partitioned into *N* non-overlapping consecutive intervals, resulting in *N* + 1 boundary points denoted by $\{\gamma\}_{n=0}^{N+1}$, $\gamma_0 < \gamma_1 < \cdots < \gamma_N$ with $\gamma_0 = 0$ and $\gamma_N = \infty$. Correspondingly, the adaptive modulation is selected to be in mode *n* if the SNR denoted by γ falls into the range of $\gamma_0 < \gamma_1 < \cdots < \gamma_N$. The zero-th mode corresponds to the outage mode of the system, i.e., the transmitter does not transmit data in mode 0. The constellation used for the *n*th mode is M_n-QAM, where $M_n = 2^n$ with $n \in \{0, 1, \cdots, N-1\}$. $M_0 = 0$ and $M_N = \infty$. Thus, the spectral-efficiency of the adaptive modulation ranges from 0 to *N* - 1 bits/sec/Hz. As the SNR increases, the system selects the AMC module with higher spectral efficiency to transmit data. On the other hand, as the SNR gets worse, the system decreases the transmission rate to adapt to the degraded channel conditions. In the worst case, the transmitter stops transmitting data as in the outage mode of the system.

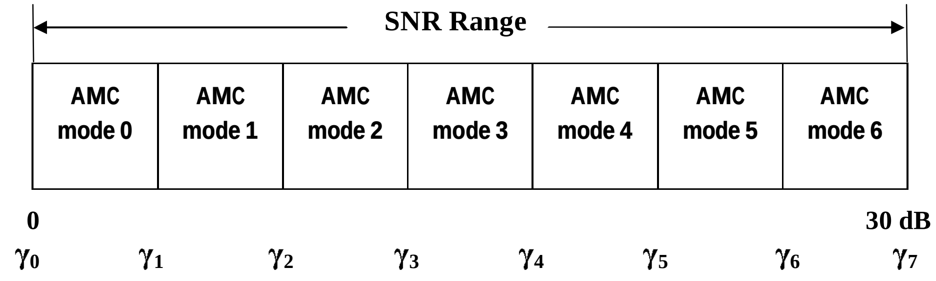

Figure 3.3 Construction of AMC Modes

In Liu et al (2004a), the PER when using the *n*th AMC mode for $n \in \{1, 2, \cdots, N-1\}$ denoted by PER$_n$(γ) can be approximated as

$$PER_n(\gamma) = \begin{cases} 1, & \text{if } \; 0 < \gamma < \gamma_{pn} \\ a_n \exp(-g_n \gamma), & \text{if } \; \gamma \geq \gamma_{pn} \end{cases} \qquad (3.2)$$

where a_n, g_n and γ_{pn} are transmission mode dependent parameters.

Based on the pdf given in Equation (3.1), the probability Pr(*n*), that the SNR falls into mode *n* is determined by

$$\Pr(n) = \int_{\gamma_n}^{\gamma_{n+1}} p_\gamma(\gamma)\, d\gamma = \frac{\Gamma\left(m, \frac{m\gamma_n}{\bar{\gamma}}\right)}{\Gamma(m)} - \frac{\Gamma\left(m, \frac{m\gamma_{n+1}}{\bar{\gamma}}\right)}{\Gamma(m)} \tag{3.3}$$

where Γ(*m*) is the complete gamma function and Γ(*m*,*x*) represents the complementary incomplete Gamma function.

In general, the Forward Error Control (FEC) and automatic retransmission request (ARQ) are also employed at the physical / data link layer. However, uncoded system is only focussed due to the following reasons: First, some simple analytical power-control policies are proposed by Goldsmith and Chua (1997) and Tang and Zhang (2006) for uncoded transmissions, while for coded transmission, it is difficult to find such a policy. Thus, we assume uncoded transmission for analytical convenience. Second, based on the work by Tang and Zhang (2007), it is observed that the performance trend of FEC/ARQ-based transmission is similar to that of uncoded systems, as long as the link BER is not too high. Therefore, the investigation of the uncoded system also provides a guideline on designing the coded system.

3.2.4 Service Process Modelling by Using FSMC

The finite state markov chain (FSMC) model is employed to characterize the variation of the wireless service process. Each state of FSMC corresponds to a mode of the adaptive-modulation scheme. Let $p_{i,j}$ denote the transition probability from state *i* to state *j*. A slow-fading channel model is assumed such that the transition happens only between adjacent states.

Under such an assumption, $p_{i,j} = 0$ for all $|i-j| > 1$. Wang and Moayeri (1995) have dealt the approximation of adjacent transition probability as

$$P_{n,n+1} \approx \frac{N_\gamma(\gamma_{n+1})T_f}{\Pr(n)}, \quad \text{if} \quad n = 0, \cdots, N-2$$

$$P_{n,n-1} \approx \frac{N_\gamma(\gamma_n)T_f}{\Pr(n)}, \quad \text{if} \quad n = 1, \cdots, N-1 \tag{3.4}$$

where $N_{1}(\gamma)$ is the Level Crossing Rate (LCR) determined by SNR of γ. Simon and Alouini (2005) proposed the LCR as

$$N_n = \sqrt{2\pi \frac{m\gamma_n}{\overline{\overline{\gamma}}}} \frac{f_d}{\Gamma(m)} \left(\frac{m\gamma_n}{\overline{\overline{\gamma}}} \right)^{m-1} \exp\left(-\frac{m\gamma_n}{\overline{\overline{\gamma}}} \right) \tag{3.5}$$

where f_d is the Doppler frequency of the mobile user. Then, the remaining transition probabilities can be derived as

$$\begin{cases} p_{1,1} = 1 - p_{1,2} \\ p_{N,N} = 1 - p_{N,N-1} \\ p_{n,n} = 1 - p_{n,n-1} - p_{n,n+1}, \quad n = 1, \cdots, K-1 \end{cases} \tag{3.6}$$

Applying Equations (3.4) and (3.6), the probability transition matrix of the FSMC, denoted by $P = \lfloor p_{ij} \rfloor_{KxK}$ can be obtained. Correspondingly, the stationary distribution of the FSMC is obtained as

$$\pi = [\Pr(0), \Pr(1), \cdots \Pr(N-1)] \tag{3.7}$$

where Pr(*n*) is given by Equation (3.3) for $n \in \{1, 2, \cdots, N-1\}$.

3.3 THE FUNDAMENTAL CONCEPT OF EFFECTIVE CAPACITY

As per Jabbari (1996), efficient and practical mechanisms for QoS support require accurate and simple channel models. Toward this end, it is essential to model a wireless channel in terms of QoS metrics such as data rate, delay, and delay-violation probability. However, the existing channel models (e.g., Rayleigh-fading model with a specified Doppler spectrum) do not explicitly characterize a wireless channel in terms of these QoS metrics. To use the existing channel models for QoS support, the parameters for the channel model are estimated, and then QoS metrics are extracted from the model. This two-step approach is obviously complex, and may lead to inaccuracies due to possible approximations in extracting QoS metrics from the models.

To address this issue, Wu (2003) proposed and developed a link layer channel model termed, the Effective Capacity (EC) model. In this approach, the author first modelled a wireless link by two EC functions, namely, the probability of nonempty buffer and the QoS exponent of the connection and then proposed a simple and efficient algorithm to estimate these EC functions. The physical layer analogy of these two link-layer EC functions is the marginal distribution (e.g., Rayleigh–Ricean distribution) and the Doppler spectrum, respectively. The key advantages of EC link-layer modelling and estimation are: 1) ease of translation into QoS guarantees, such as delay bounds; 2) simplicity of implementation; and 3) accuracy and hence efficiency in admission control and resource reservation. Simulation results of Wu (2003) showed that the actual QoS metric was closely approximated by the estimated QoS metric obtained from EC channel estimation algorithm, under a wide range of conditions. This demonstrated the effectiveness of the EC link-layer model in guaranteeing QoS.

Conventional channel models directly characterize the fluctuations in the amplitude of a radio signal. These models are called as physical layer channel models. Physical layer channel models provide a quick estimate of the physical layer performance of wireless communications systems [e.g., symbol error rate versus SNR]. However, physical layer channel models cannot be easily translated into complex link-layer QoS guarantees for a connection, such as bounds on delay. The reason is that these complex QoS requirements need an analysis of the queuing behaviour of the connection, which is difficult to extract from physical layer models. Thus, it is hard to use physical layer models in QoS support mechanisms, such as admission control and resource reservation. The limitation of physical layer channel models in QoS support is the difficulty in analysing queues using them.

In this thesis, this EC channel model is adopted by moving up the protocol stack from the physical layer to the link layer, so that it captures a generalized link-level capacity notion of the fading channel. The EC link model aims to characterize wireless channels in terms of functions that can be easily mapped to link-level QoS metrics, such as delay-bound violation probability.

3.3.1 Statistical QoS Guarantees

The real-time multimedia services such as video and audio require the bounded delay, or equivalently, the guaranteed bandwidth. Once a received real-time packet violates its delay bound, it is considered as useless and will be discarded. However, over the mobile wireless networks, a hard delay bound guarantee is practically infeasible to be achieved due to the impact of the time-varying fading channels. For example, over the Rayleigh fading channel, the only lower-bound of the system bandwidth that can be deterministically guaranteed is a bandwidth of zero (Wu 2003). Thus, an alternative solution is considered by providing the statistical QoS guarantees, where the delay-bound with a small violation probability is guaranteed.

During the early 90's, the statistical QoS guarantees theories have been extensively studied in the contexts of so-called effective bandwidth theory with the emphasis on wired ATM networks by Chang (1994) and Courcoubetis and Weber (1995). The asymptotic results given by Chang (1994) showed, for stationary arrival and service processes with the average arrival-rate less than the average service-rate, the probability that the queue size Q exceeds a certain threshold C decays exponentially fast as the threshold C increases (Wu and Negi 2003), i.e.,

$$\Pr\{Q > C\} \approx e^{-\theta C} \tag{3.8}$$

where θ is a certain positive constant called QoS exponent. Furthermore, when delay-bound is the main QoS metric of interest (i.e., when the focus is on delay-bound violation probability), an expression similar to Equation (3.8) can be obtained as

$$\Pr\{Delay > D_{\max}\} \approx \varepsilon\, e^{-\theta\delta D_{\max}} \tag{3.9}$$

where D_{max} denotes the delay-bound, and δ is jointly determined by both arrival and service processes.

From Equations (3.8) and (3.9), it is seen that the parameter θ plays an important role for the statistical QoS guarantees, which indicates the decaying-rate of the QoS violation probability. The smaller θ corresponds to the slower delaying-rate, which implies that the system can only provide a looser QoS requirement, while a larger θ leads to a faster delaying-rate, which means a more stringent QoS requirement can be guaranteed. Consequently θ is called as QoS exponent.

3.3.2 Effective Bandwidth and Cross-Layer Designs

Inspired by the effective bandwidth theory, Wu and Negi (2003) proposed a powerful concept termed as effective capacity, which turns out to be the dual problem of the original effective bandwidth. The effective capacity function $E_C(\theta)$, characterizes the attainable wireless-channel service-rate as a function of the QoS exponent θ. Specifically, they defined the effective capacity $E_C(\theta)$, as the constant arrival rate that the channel can support in order to guarantee a QoS requirement specified by θ.

Although the original concept of effective capacity is proposed based on constant-arrival assumption, it actually can be generalized to investigate the QoS performance of any stationary arrival process. Under such a condition, the arrival process should be represented by its effective bandwidth while the service process should be characterized by its effective capacity, respectively. Note that for a constant arrival-process, the corresponding effective-bandwidth is equal to its constant arrival-rate. Thus, the problem discussed by them can be considered as the special case of our more general scenario addressed in this section, where both arrival and service processes are time-varying.

For any given arrival process and service process, let $E_C(\theta)$ and $E_B(\theta)$ be the effective capacity and effective bandwidth functions respectively. The two limiting values are defined as follows (Wu and Negi 2003):

$$\begin{aligned} \mu_A &\overset{\Delta}{=} \lim_{\theta \to 0} E_B(\theta) \\ \mu_C &\overset{\Delta}{=} \lim_{\theta \to 0} E_C(\theta) \end{aligned} \tag{3.10}$$

The effective bandwidth theory demonstrates that μ_A is equal to the average arrival-rate of the traffic process (Chang 1994). Also Tang and Zhang

(2007), had shown that μ_C is equal to the average service rate of the service process. Using the approximation pointed out by Chang (2000), the buffer non-empty probability ε in Equation (3.9) can be expressed as

$$\varepsilon \approx \frac{\mu_A}{\mu_C} \tag{3.11}$$

Increasing the service-process bandwidth results in higher effective capacity, which will lead to a larger QoS- exponent solution θ^*. This implies that the higher bandwidth service-process can support a more stringent QoS for a given arrival process. On the other hand, increasing the arrival-process bandwidth makes the effective bandwidth to increase, which generates a smaller QoS-exponent solution θ^* for a given service process. This implies that only a looser QoS can be guaranteed. When the bandwidth of the arrival process further increases such that $\mu_A > \mu_C$, there is no solution for $\theta^* > 0$ existing. Thus, the service process cannot support any QoS for the given arrival process, which is consistent with the queuing theory that if $\mu_A > \mu_C$, both queue length and the queuing delay will approach to infinity.

The QoS exponent θ can be used as a bridge in cross-layer design modelling between the physical layer system infrastructure and the upper layer network protocols' statistical QoS performance. The characterization of the QoS performance guarantees are equivalent to investigating the dynamics of the effective capacity function, which turns out to be a simple and efficient cross-layer modelling approach.

Based on the duality between effective bandwidth and effective capacity, the effective capacity function denoted $E_C(\theta)$, has the following properties (Wu et al 2003):

- $E_C(\theta)$ is a monotonically decreasing function of θ, i.e., $\frac{dE_C^{(\theta)}}{d\theta} \leq 0$, for all $\theta > 0$.

- $\lim_{\theta \to 0} E_C(\theta)$ converges to the average service rate , i.e., $\lim_{\theta \to 0} E_C(\theta) = \overline{\mu}$

- $\lim_{\theta \to \infty} E_C(\theta)$converges to the minimum service rate, i.e., $\lim_{\theta \to \infty} E_C(\theta) = \mu_{\min}$.

Intuitively, these properties can be explained as follows: As the QoS constraint becomes more stringent, the given channel can support lower traffic arrival rates in order to guarantee the more stringent delay QoS requirement. Hence the effective capacity is a decreasing function of θ. On the other hand, when the system can tolerate long delay, the maximum arrival rate a given channel can support is equal to its average service rate. However, if the arrival rate increases beyond the average service rate, from queuing theory we know that a large queue will build up and the queue size will eventually approach to infinity. This is the reason effective capacity converges to its average service rate when $\theta \to 0$. When the system cannot tolerate any delay, the arrival rate can be restricted to be equal to or less than the minimum service rate to ensure that the queue will never build up. Therefore, the effective capacity converges to the minimum service rate as $\theta \to \infty$.

To help demonstrate the principles and identify the relationship between effective bandwidth and effective capacity, let us consider two cases as illustrated in Figure 3.4.

Case I — Service process (ii) is given with arrival process (i) having higher bandwidth than that of arrival process (ii).

For the fixed service process (ii) in Figure 3.4, a higher bandwidth traffic arrival process (i) plotted in terms of effective bandwidth, intersects with $E_C(\theta)$ at the QoS exponent θ_1, while a lower bandwidth traffic arrival process (ii) plotted according to the effective bandwidth function intersects with $E_C(\theta)$ at the QoS exponent θ_2.

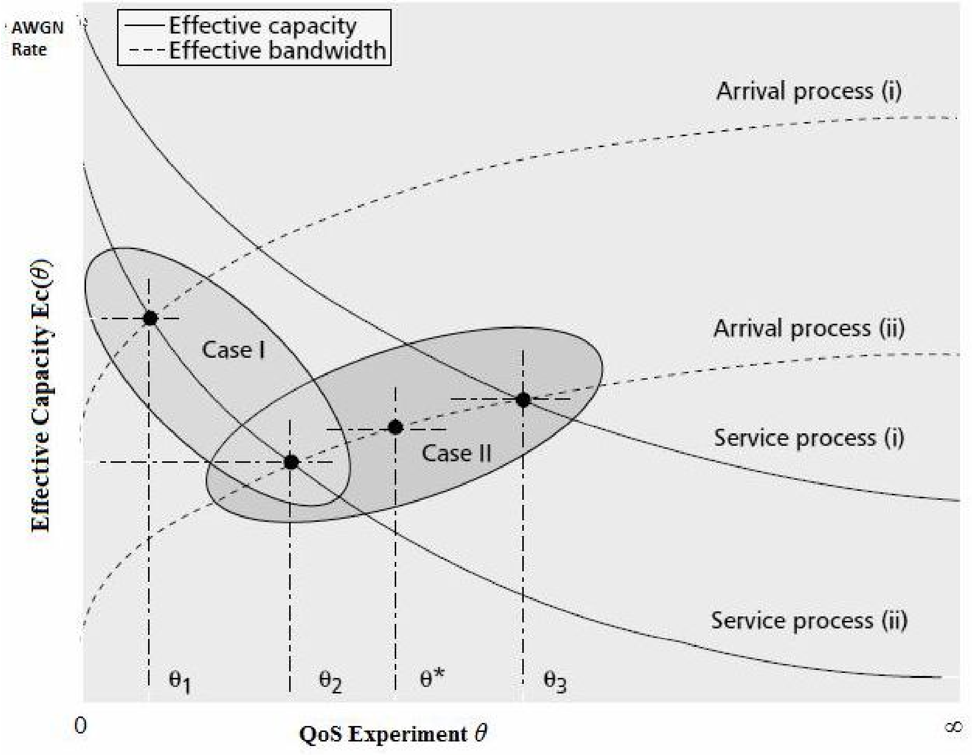

Figure 3.4 Relationship Between Effective Bandwidth and Effective Capacity as a Function of the QoS Exponent *θ*. (*After Wu 2003*)

Clearly, the figure shows $\theta_1 < \theta_2$. This obviously implies that the given service process (ii) can support a more stringent QoS for the slow arrival process (ii) than the faster arrival process (i), since intuitively the higher bandwidth arrival process (i) results in larger QoS violation probability.

Case II — Arrival process (ii) is given with service process (i) having higher bandwidth than that of service process (ii). For the fixed arrival process (ii) in the above Figure 3.4, statistical QoS requirement is specified as the QoS exponent θ^*. Then the higher bandwidth service process (i) can guarantee the required statistical QoS since the intersection θ_3 between arrival process (ii) and service process (i) satisfies $\theta_3 > \theta^*$, while the lower bandwidth service process (ii) cannot support the required QoS provisioning, because the intersection θ_2 between arrival process (ii) and service process (ii) results in $\theta_2 < \theta^*$.

From the above observations and analyses, it is proposed to use the effective bandwidth and effective capacity as the controlling functions for cross-layer modelling. The characterizations of the QoS performance guarantees are equivalent to investigating the dynamics of the QoS exponent θ, which turns out to be a very simple and efficient cross-layer modelling approach are discussed in the next section.

3.3.3 Effective Capacity of the Proposed Scheduling Scheme

The effective capacity is known as the dual problem of the effective bandwidth (Wu et al 2003). Thus, utilizing the well-established effective bandwidth theory, it is feasible to formulate the effective capacity problem analytically. Let the sequence of non-negative discrete random variables $\{R(t),\ \ t=1,2,\cdots,k\}$ denote the attainable rate of the discrete service process. The values of $R(t)$ are chosen from a discrete set $R\overset{\Delta}{=}\{\mu_1,\mu_2,\cdots,\mu_k\}$. The marginal probability density function (pdf) of $R(t)$ is denoted by $p_R(\tau)$.

Moreover, $S(t)\overset{\Delta}{=}\sum_{i=1}^{t}R(i)$ is defined as the time cumulated service process over the time sequence of $i=1,2,\cdots,t$. The Gartner-Ellis limit of $S(t)$ defined as

$$\Lambda_C(\theta)\overset{\Delta}{=}\lim_{t\to\infty}\frac{1}{t}\log\left(E\left\{e^{\theta S(t)}\right\}\right) \tag{3.12}$$

is a convex function and differentiable for all real θ, where $E\{\cdot\}$denotes the expectation. Then, the effective capacity of the service process, denoted by $E_C(\theta)$, where $\theta > 0$, is defined as

$$E_C(\theta)\overset{\Delta}{=}-\frac{\Lambda_C(-\theta)}{\theta}=-\lim_{t\to\infty}\frac{1}{\theta t}\log\left(E\left\{e^{-\theta S(t)}\right\}\right) \tag{3.13}$$

Based on the physical layer FSMC model discussed in section 3.2.4, the effective capacity of the FSMC based service process is determined by (Wu et al 2003)

$$E_c(\theta)=-\frac{1}{\theta}\log(\rho\{P\Phi(\theta)\}),\ \theta>0 \tag{3.14}$$

where $\rho\{.\}$ represents the spectral radius of the matrix, P is the transition matrix of the developed FSMC determined by Equations (3.4 - 3.6) and

$$\Phi(\theta)\overset{\Delta}{=}diag\left\{e^{-\lambda_0\theta},e^{-\lambda_1\theta},\cdots,e^{-\lambda_{N-1}\theta}\right\} \tag{3.15}$$

where $\Phi(\theta)$ is K X K diagonal matrix with the k^{th} diagonal entry $\Phi_k(\theta)=\exp(-\lambda_n\theta)$, $\lambda_n=\tilde{R}_nT_fW$, with $n\in\{0,1,\ldots,N\text{-}1\}$ is the number of bits per frame transmitted by the n^{th} mode of the adaptive-modulation scheme (transmission rate), W is the system spectral bandwidth and $\tilde{R}_n$ is the achieved spectral efficiency of the n^{th} mode expressed as

$$\tilde{R}_n = R_n \left(1 - \overline{PER_n}\right) \tag{3.16}$$

3.4 EFFECTIVE CAPACITY BASED ADAPTIVE RESOURCE SCHEDULING

The cross-layer modelling introduced in section 3.2 establishes the analytical framework to investigate the impact of physical layer infrastructure variations on the statistical QoS provisioning performance at the data link layer through the effective capacity function. In this section, we develop the adaptive resource allocation algorithms based on the developed cross-layer model to guarantee the desired QoS requirements, adopting the simple round-robin (RR) scheduling algorithm for the real-time mobile users.

3.4.1 Effective Capacity of the Service Process

As described in section 3.2, the proposed system operates in a dynamic TDMA mode. The k^{th} user is assigned with $L^{(k)}$ number of time slots per frame for information transmission. In order to determine the number $L^{(k)}$ time slots allocated to the k^{th} user to support its statistical QoS, the effective capacity of the service process is to be derived.

Considering $L^{(k)}$ = 1 time slot as a basic unit to the k^{th} user, the effective capacity of the k^{th} user, denoted by

$$E_C^{(k,1)}(\theta) = -\frac{1}{\theta} \log\left(\rho\left\{P^{(k)} \Phi^{(1)}(\theta)\right\}\right), \quad \theta > 0 \tag{3.17}$$

where $P^{(k)}$ is the transition probability matrix of the k^{th} user, determined by the k^{th} user's channel statistics and is independent of $L^{(k)}$, and $\Phi^1(\theta)$ is given by

$$\Phi^{(1)}(\theta) \overset{\Delta}{=} diag\left\{e^{-\lambda_0^{(1)}\theta}, e^{-\lambda_1^{(1)}\theta}, \cdots, e^{-\lambda_{N-1}^{(1)}\theta}\right\} \tag{3.18}$$

Here, λ_ns are independent of the channel statistics.

When allocating $L^{(k)} = l$ time slots for the user, applying the results developed in Wu and Negi (2005), the effective capacity $E_C^{(k,l)}(\theta)$ can be expressed as

$$E_C^{(k,l)}(\theta) = l\, E_C^{(k,l)}(\theta)(l\theta) \tag{3.19}$$

3.4.2 Admission Control and Time Slot Allocation

Let the kth user's statistical QoS requirement be denoted by $\{D_{\max}^{(k)}, \varepsilon^{(k)}\}$, where $D_{\max}^{(k)}$ is the delay bound and $\varepsilon^{(k)}$is the violation probability. Similar to the procedure described in section 3.3, the time slot allocation algorithms can be designed in the following steps:

S1: Denote the effective bandwidth of the kth user's arrival-process by $E_B^{(k)}(\theta)$. Find the solution of the rate and QoS-exponent (δ_l, θ_l) such that $E_B^{(k)}(\theta_l) = E_C^{(k)}(\theta_l)\,\delta_l$

S2: Using $L^{(k)} = l$ number of time slots, the delay bound violation probability can be derived as

$$\Pr\left\{Delay > D_{\max}^{(k)}\right\} \approx \exp\left(-\theta_l \delta_l D_{\max}^{(k)}\right) \tag{3.20}$$

S3: The number $L^{(k)}$ is determined by

$$L^{(k)} = \min_{1 \le l \le L} \{l\}, \text{ such that } \exp\left(-\theta_l\, \delta_l\, D_{\max}^{(k)}\right) \le \phi\varepsilon^{(k)} \tag{3.21}$$

For each real-time user, $L^{(k)}$ can be calculated using Equation (3.21). This clearly shows that the total number of time slots allocated to the real-time users' need to satisfy the following equation:

$$\sum_{k=1}^{K} L^{(k)} \leq L \tag{3.22}$$

When a new mobile real-time user applies to join the system, the admission-control algorithm examines if the number of available time slot resources is sufficient to support the new user's statistical QoS. If yes, the new real time mobile user is admitted to join the system; otherwise, this new real-time mobile user is rejected to join the system.

3.5 NUMERICAL AND SIMULATION RESULTS

The proposed time slot allocation algorithm is evaluated by computer simulations. The number of adaptive modulation modes is set as N=6, the total system has spectral-bandwidth B of 1MHz with 50 users, the data-link layer frame duration T_f as 2ms, the number of time slots per frame L as 100 and the Doppler frequency spread f_d as15Hz. Moreover, two types of real-time services are generated. The first type simulates the low speed audio service, where the arrival traffic is modelled by the well known ON-OFF fluid model. The holding times in "ON" and "OFF" states are exponentially distributed with the mean equal to 8.9 ms and 8.4 ms, respectively. The "ON" state traffic is modelled as a constant-rate of 32 Kbps. The second type simulates a high-speed video traffic flow. Jabbari (1996) employed a first order auto-regressive (AR) process to simulate video traffic characteristics, the bit-rate of which can be expressed as

$$v(t) = av(t-1) + bw\,(25) \tag{3.23}$$

where a = 0.8781, b = 0.1108 and w is a Gaussian random variable with the mean 80 Kbps and the standard deviation of 30 Kbps. The effective bandwidths of the audio and video traffic are derived according to Chang

(1994) and Courcoubetis and Weber (1995), respectively. The QoS requirements of these two types of services are shown in Table 3.1.

Table 3.1 QoS Requirements for Audio and Video Services

Services	**Bit Error rate (BER)**	**Delay-bound ($\mathbf{D_{max}}$)**	**Delay Violation Probability**
Audio	10^{-3}	50 ms	10^{-2}
Video	10^{-4}	150 ms	10^{-3}

Using the time slot allocation algorithm proposed in section 3.3.2, the Figure 3.5 and 3.6 show the simulated results of allocated time slots for audio and video services as a function of the average SNR. As shown by the figures, for both audio and video services, the required time slots for supporting the QoS decrease as the average SNR increases. It indicates the reduction in the amount of resource allocation when the channel is in good condition (high SNR state) and vice versa in bad channel condition (low SNR state).

When the average SNR is low, the time slot allocation algorithms may not find the feasible solution of the $L^{(k)}$ to support the required QoS, since $L^{(k)}$ must satisfy $1 \leq L^{(k)} \leq L$. It can be observed that the developed effective capacity based scheduling approach reveals improved significance in terms of minimum number of time slots allocated to audio and video services over conventional constant power approach (Tang and Zhang 2007).

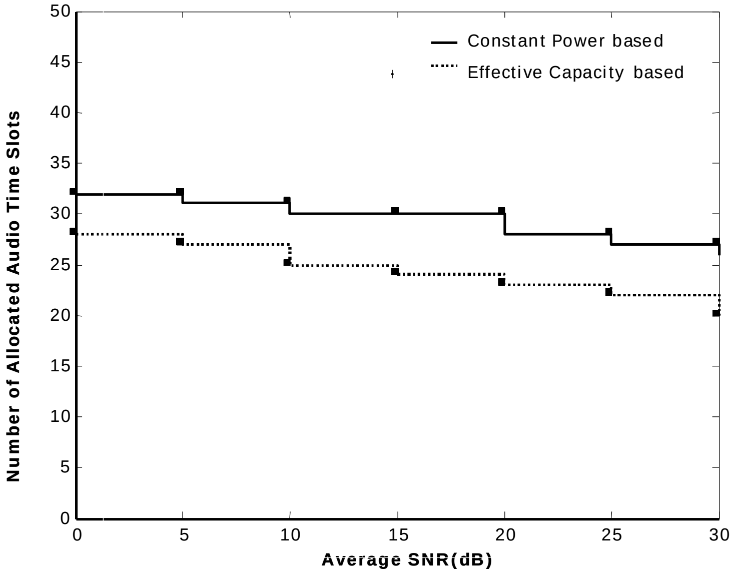

Figure 3.5 Allocated Audio Time Slots as a Function of Average SNR (dB)

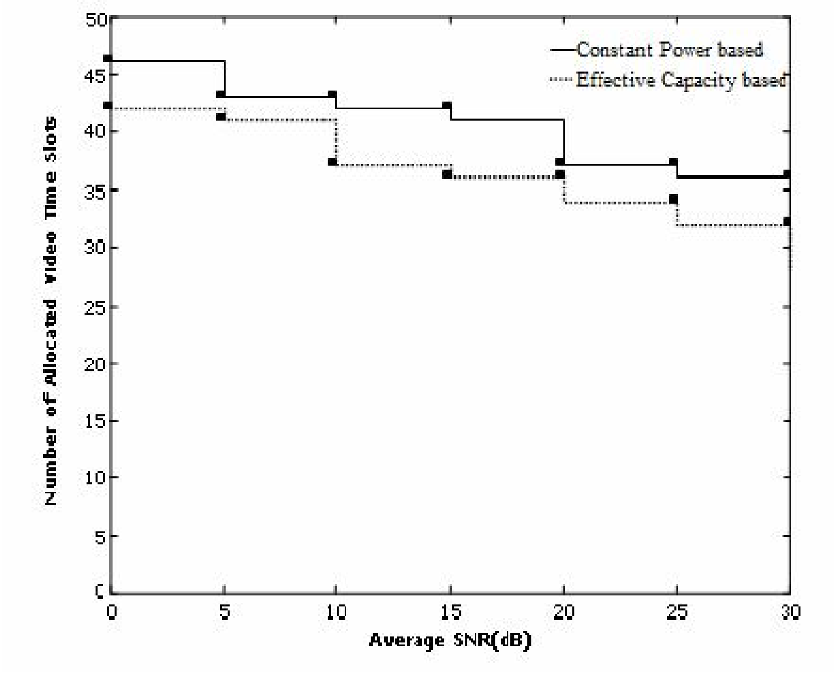

Figure 3.6 Allocated Video Time Slots as a Function of Average SNR (dB)

Table 3.2 shows the numerical results of resource allocation algorithms for audio and video services.

Table 3.2 Performance of the Effective Capacity Based Cross-Layer Scheduling

Average SNR (dB)	No. of audio time slots allocated			No. of video time slots allocated		
	Constant Power approach	Effective Capacity based approach	No. of audio time slots Saved	Constant Power approach	Effective Capacity based approach	No. of video time slots Saved
0	32	28	4	46	42	4
5	31	27	4	43	41	2
10	30	25	5	42	37	5
15	30	24	6	41	36	5
20	28	23	5	37	34	3
25	27	22	5	36	32	4
30	26	20	6	33	28	5

From the results, it is understood that the proposed approach is providing significant reduction in the number of time slots for the same type of services. As the SNR increases, generally the number of time slots taken or assigned for transmission decreases and vice versa. From the tabulation, by comparing the allocation and saving in the audio and video time slots, it is clear that the proposed effective capacity based scheduling approach is better than constant power based scheduling in terms of number of time slots saved after allocation.

To evaluate whether the allocated time slots can support the required statistical QoS, Figure 3.7 is plotted to show the variations of the simulated delay-bound violation probabilities for video and audio services using the proposed effective capacity based scheduling algorithm.

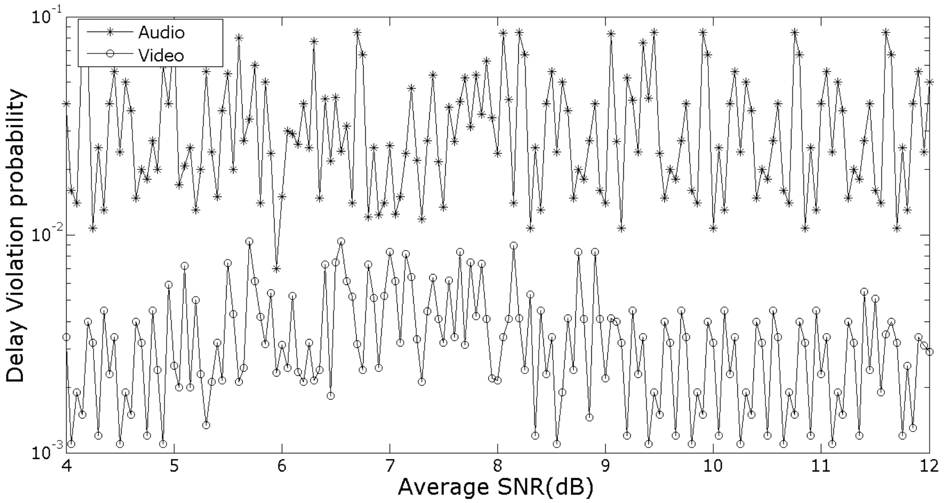

Figure 3.7 Variations of Simulated Delay Bound Violation Probabilities for Both Audio and Video Services

It is observed that for both audio and video services, the simulated delay-bound violation probabilities are below the required upper-bounds ε's. This is due to the fact that the approximations in Equations (3.8) and (3.9) are conventional. It is also observed that the conventional constant power control scheme achieves the similar QoS violation performance by using much more resources (i.e., time slots) than proposed QoS-driven power control approach. Figure 3.7 shows that the QoS violation probability fluctuates according to the time slot allocation outcomes varying within a discrete set due to the proposed time slot allocation.

3.6 SUMMARY

In this chapter, an effective capacity based cross-layer scheduling is proposed and analysed for statistical QoS guarantees over wireless networks. The critical relationship between effective bandwidth and effective capacity is identified and the effective capacity function is obtained analytically in the proposed system configuration. The proposed scheme allocates time slots adaptively for real-time users to guarantee statistical delay-bound QoS requirements. The admission control and time slot allocation algorithms are developed by extending the effective capacity method of scheduling. The numerical results showed that the AMC with FSMC based service modelling at the physical layer and effective capacity based scheduling at the link layer have significant impact on the statistical QoS performance at upper protocol layers. The admission control and time slot allocation algorithms are developed by extending the existing effective capacity method. Compared to the conventional constant power approach, the proposed time slot allocation scheme shows significant advantages in terms of allocating minimum number of time slots with diverse guarantee of the statistical delay-bound QoS requirements. Due to this, it can significantly reduce the transmit power or equivalently increase the admission region.