1. Introduction

We first introduce in Section 1.1 essential information about wireless air interface technology, radio channel characteristics, multiple access techniques, forward-link scheduling issues, and admission control. These provide the background adequate for catching the objective of the thesis work, which is then described in Section 1.2. Some previous works related to the study are addressed in Section 1.3. In Section 1.4, we give a brief summary and the remaining of the thesis organization.

1.1 Background

With great advantages of access anytime and anywhere, wireless communication services have enjoyed phenomenal growth of market penetration during the past decade. Meanwhile, wireless users are also demanding one more feature, access to any format of media types satisfactory. To fulfill the requirement in wireless networks, an access and transmission system effective in providing a wide variety of data services with different QoS requirements is necessitated. Advanced wideband packet-switched techniques facilitate the design of one radio interface for traffics of different characteristics, such as loss-sensitive bursty traffics and traditional time-sensitive voice ones. In particular, the wideband code-division-multiple access (CDMA) spread spectrum technique has been regarded as the choice of radio interface technology for

future wireless networks due to its flexibility for different traffic types, more resistance to interference and multipath delay spread, and better radio coverage in frequency selective fading environments [1].

One essential feature of wireless communications is data waveforms transmitted over time-varying channels. The time-varying characteristic is mainly caused by radio propagation environment, user mobility, and reuse of limited frequency spectrum, but also slightly affected by background thermal noise. Its effects at a radio receiver are the fluctuations of received signals in *fast fading*, *slow fading*, or *shadowing*, in addition to distance-related attenuation, which in general impairs the capacity of information carried by a data waveform. To resolve this issue present in the physical layer, *diversity* in space, time or frequency, *equalization*, and *channel coding* are three techniques which can be used to improve received signal quality or link performance over fast fading channels and small-scale distances [2]; While *power control* is a traditional technique in CDMA system to improve link quality over slow fading and shadowing channels and large-scale distances. Recent interests in high-speed data applications over wireless links have given rise to consideration of different approaches in the physical, link, and network layers to exploiting time-varying channel characteristics for wireless system performance and service QoS.

In order to support bursty (elastic) data services with different QoS constraint, high data rate transmission is vital. This necessitates transmitted data waveforms being received at high level of signal-to-interference and noise power ratio (SINR). The requirement has caused some significant changes in the design of wireless *multiple access* techniques, besides developing different diversity techniques to improve received SINR. One of the changes is to transmit data waveforms in time-division-multiplexing (TDM) format, instead of code-division-multiplexing (CDM) one, so that they can be received at high SINR. This change not only improves cell coverage of a service base station transmitter (BST) but also avoids the effect of extra intra-interference from non-orthogonal factors which usually exists in CDM. Another change is to use *rate control*, instead of power control, in order to efficiently exploit highly dynamic SINR. By rate control, a transmitter is supposed to adapt modulation schemes, data encoding, and information rate of waveforms to a radio link where the target mobile terminal (MT) is able to receive the data successfully.

Wireless data traffics are highly asymmetric, with the bottleneck of data transport at the forward link rather than the reverse link [1]. Besides, bursty (elastic) data traffics are remarkably tolerant of occasional transmission delay over a high speed channel. This thus gives room for scheduling TDM waveforms over forward link so as to provide some QoS guarantee as well as to make efficient use of cell capacity by

means of multiuser diversity. In the spirit, a number of *opportunistic* schedulers have been proposed to take advantage of fast fading [3]. The policy is to schedule data waveforms to various users when their radio SINRs are relatively favorable, while ensuring fair access to the transmission resources.

In summary, the main controllable variables for forward-link data services are data rate, transmission power, and the scheduling of transmission. These have given rise to lots of previous work in the literature investigating different algorithms for striking a balance between cell capacity and user fairness through rate control, power control, opportunistic scheduling or their different combinations. As far as QoS is concerned, it is inevitable to limit the amount of accepted traffic load in system. When overloaded, the system would rather block new service requests than further degrade the performance of ongoing services.

To guarantee some QoS constraint, it is imperative that admission control be employed to limit the amount of traffic load in system. For voice traffic (resp. virtual circuit) service, a general approach is to block new call (resp. connection) requests when system becomes overloaded (resp. is unable to guarantee QoS parameters). In wireless communications, admission control has to further take into account some preallocated guard resources required for ensuring service continuity, due to user

mobility and dynamic available resources. Consequently, it tends to be more conservative, which thus raises the issue of system resource utilization. On the other hand, congestion control is considered in this context for data networks. It tends to be more aggressive, which results in better system resource utilization but poor QoS guarantee, if any.

In the absence of admission control for wireless forward link data services, user-level performance deteriorates when the number of active users is large. For example, the channel-aware scheduling algorithm for a symmetric rate distribution [4] provides each active user a fraction of $G(N)/N$ of its time average rate C_i if the weight-based strategy with weight $1/C_i$ for user i is employed, where N is the number of active users and $G(N)$ represents multiuser diversity gain. It was shown that $G(N)$ is an increasing function of N while $G(N)/N$ decreases with N. Considering uniform distribution of MTs in two dimensional space of radio coverage, most of active users in N will be located in far field of a service BST while their time average rate C_i will be very low due to poor received SINR. The effect of low mean service rate $C_i\ G(N)/N$ is a large number of active users suffering from long service delay.

In the presence of admission control for resolving the above issue, a natural criterion of admission for a service request is the number of active users $N \leqq C_R G(N)/$

C_{min} or the minimum data rate guarantee $C_i G(N)/N \geqq C_{min}$ for all admitted user i, where C_R is the average data rate at cell boundary. For short, the approach based on the former is thereafter referred to as number-based admission control and the latter as rate-based admission control, respectively. This entailed the study in [5], which relates user performance, in terms of blocking probability and data throughput, to cell size and traffic density.

In particular, the authors in [6] considered BST as a proxy server and service requests as a batch arrival process. They defined the ratio of service time to job time as a *normalized delay* (or *stretch*) and developed several forward link scheduling algorithms based on whether rate information or request size is available. It was shown that if both rate information and request size are available, prioritizing jobs with smaller “air time” requirement leads to distinct performance advantages. Our traffic model in the thesis turns out to be the same but our focus is not entirely on scheduling algorithm.

1.2 Objective

In the literature, there have been many research studies on exploiting forward link time-varying characteristics for optimizing system throughput [7] or admission capacity region [8]. Particularly, the work [9] involved rate-based and number-based admission control for elastic data services, but their focus was the performance of cell dimensioning under the constraint of a given “hard guarantee” on the minimum service rate for admitted users. The number-based scheme is able to give fair blocking probability when the system becomes overloaded, however the users with requests for short air time of service may not perceive it as real fairness. The rate-based scheme gives preference to users with high received SINR when the system becomes overloaded, but it may not totally acceptable for the users with low received SINR for short service air time. Both essentially provide hard guarantees of minimum service rates at the expense of service blocking. Besides, the inherent “elasticity” of data transfers is not entirely explored for alleviating the impact of temporal and spatial random traffic fluctuations which are severe in wireless environments. On the other hand, part of the work [6] took the elasticity in the extreme by assuming that data traffic arrives according to a batch arrival process and that radio resources can be allocated “offline”. They investigated the effect of different scheduling algorithms taking into account whether rate or job size information is available, but overlooked

admission control for QoS guarantee.

In this thesis, we investigate alternative admission control algorithms for wireless forward link data services. For the benefit of prioritizing service opportunities and exploiting the elasticity of data traffic, we consider admission criteria which are based on service air-time requirement and admission delay tolerance. Moreover, we assume that forward-link data service requests arrive according to a batch arrival process, and that the information of feasible data rates for each user is available. Essentially, the admission control strategy is to accept requests for short service air time but to buffer requests for long service air time when the forward link data service system is temporally overloaded. Those buffered requests are later accepted for service when the system is not overloaded. However, if the buffered but not accepted request waits for more than a time threshold, it is discarded (dropped, a forced renege). The objective is to minimize service dropping as well as to maximize the instantaneous rate of serviced requests (transactions) when the system is temporally overloaded. We envisage that these goals to alleviate the impact of random traffic fluctuations meet the general expectation by most users and system operators in wireless temporal traffic hotspots. Since the metric of admission criterion is air time, we will refer to the proposed strategy as time-based admission control.

The spirit of time-based admission control is to take advantage of “the shortest processing time (SPT) rule [10]” and queueing requests for alleviating the impact of temporal fluctuations of traffic loads on service blocking. This approach generally results in the shortest service delay and the least number of requests waiting for service. As to QoS issues, the time-based scheme can simply use the admission criterion of the number-based or the rate-based as the condition for turning on the time-based admission control. This method allows the time-based scheme to provide more flexible QoS parameters than the minimum service rates. However, we will develop novel time-based schemes taking into account requirements for soft or hard QoS guarantee on given parameters, without involving any number-based or rate-based admission criteria.

Key to the design of time-based scheme are when to turn on and to turn off admission control, what criterion is based for admission of a request into service schedulers, and what criterion is based for discarding a buffered request. For these purposes, we simply use three time constants, called *busy threshold*, *admission threshold*, and *dropping threshold*, respectively. In this work, the service time required for the scheduler to finish all admitted traffic remaining in queue, thus called “residual service time”, is constantly monitored. Specifically, we will investigate the time-based admission control, by which system residual service time is constantly compared with

busy threshold; If it is larger (resp. smaller) than the *busy threshold*, the time-based admission control is turned on (resp. off). When the time-based admission control is turned on, a service request is accepted into scheduler (resp. queued in buffer) if it requires service air-time shorter (resp. longer) than *admission threshold*. When the time-based admission control is turned off, any new service request can be accepted and any queued request can be selected into scheduler (In this work, we only consider FIFO discipline for queued requests). Additionally, if any of queued requests waits for more than *dropping threshold*, it is discarded and causes a service blocking. In fact, the *dropping threshold* is aimed to somewhat emulate the action of reneging that some users or the time-out mechanism of some reliable transport protocols may take. Also note that the use of one *admission threshold* actually gives rise to two service priorities. (In the future, we will develop another type of time-based scheme for guarantee of service stretch.)

We will investigate the impact of varying different thresholds in the proposed time-based scheme on the instantaneous rate of serviced requests, mean service delay, and the blocking probability of new service requests all through simulations. These involve *ergodic* analysis for mean service delay and blocking probability and *transient* analysis for instantaneous rates of serviced requests and service curve [11]. For ergodic analysis, arrival traffic for forward link service is assumed to be a stationary

Poisson batch arrival process. For transient analysis, it is assumed to be a Poisson batch arrival process with two states of arrival rates. Some typical results from transient analysis will be shown to highlight the effect of time-based schemes on relieving traffic hotspots. We will also look into issues on the feasibility and effect of using adaptive time thresholds, besides the static ones.

For the radio environment in simulations, we will consider only path-loss components without fast fading and shadowing effects. This results in static channel conditions and thus allows us to focus on admission control issues without having to involve opportunistic scheduling for capturing fast channel dynamics. Furthermore, we will consider discrete rate functions of received SINR with four available rates. Since opportunistic schedulers gain nothing in such radio environment and the *Round-Robin* service discipline generally supports favorable performance metrics in most time-invariant environments [12], we will simply use the Round-Robin policy for scheduling accepted traffics in all simulations. Results from simulations for time-based, number-based, and rate-based schemes will be obtained and compared. (In the future, we will look into the effect of time-varying channels from fast fading and shadowing on the performance of the time-based admission control.)

1.3 Related Works

Because of the enormous increase in demand for Internet access, it is expected that next-generation wireless networks will support this type of data service satisfactorily besides traditional real-time voice traffic. In general, the Internet data traffic comprises time-sensitive streaming data from real-time applications and loss-sensitive bursty data from non-real-time and computer applications. The Internet data traffic thus has quite different traffic characteristics and different quality of services requirements than traditional voice traffic. Wireless channel resources are, however, very limited and highly fluctuate. To use the resources efficiently and adapt the time varying nature of channel, wireless systems have to allocate resources such as power, code space and timeslots dynamically [13]

There are two ways to share a transmission channel. They are Code Division Multiplexing (CDM) mode and Time Division Multiplexing (TDM) mode. Users with different codes can use a channel at the same time under CDM mode. Some methods [14] were proposed to solve the interference problem under CDM mode, arising from the fact that the base station might transmit data to many users at the same time. In contrast to CDM mode, only one user can use the transmission channel at a given time under TDM mode. Therefore, the intra-cell interference by other users in the same cell does

not exist[16]. Moreover, using TDM waveform can eliminate power sharing among active users, which improves cell coverage, by allocating full power and all code channels to a single user at any moment [15].

It was mentioned in [1] that the bottleneck of high-speed data system is the forward link rather than the reverse link. There are, nevertheless, some works on the subject of improving reverse link capacity [17] [18]. For data applications, most of the data traffic is from the base station to users. As a result, the amount of traffic over the forward link is much more than that over the reverse link [5] [19]. This property of highly "asymmetric" service requirement supports the fact that the forward link is the limit of system capacity. Most of recent research has focused on forward link [7] [20].

In cellular networks, the main controllable variables are data rate, transmission power, number of active users, and the scheduling of transmissions. They gave rise to lots of previous work investigating different algorithms, such as rate control, power control, admission control, and scheduling. Because of the time varying channel quality, rate control adapts transmission rate under constant power [21], while power control adjusts transmit power in order to keep the transmission rate fixed [22]. Scheduling algorithms decide which user should be served in each time slot [4] [23]

[24]. Although these algorithms above may improve or optimize system throughput, they do not consider how many users could exist in the system at same time. When overloaded, the system would rather block new requests than further degrade the performance of ongoing connections, so that the minimum data rate of all users could be guaranteed [9].

1.4 Summary and Thesis Organization

Previous admission control algorithms for high speed wireless data service over forward link all focused on optimizing system throughput, but the issue of user performance has been rarely addressed. Reference [9] provides two admission control algorithms, based on the number of active users and based on the minimum data rate, respectively. Both are designed to account for the random nature of traffic, the inherent“elasticity”of data transfers, and the spatial component of offered traffic.

In view of the fact that a significant amount of transaction-based services is being carried over wireline data networks, it is vital for future wireless data networks to support this type of service satisfactorily. An essential feature of transaction-based service is “short service time” required by each request which expects not to be denied for service often. The study in [9] gave two simple admission control algorithms for

minimum rate guarantees but did not address the above issue. In this thesis, we particularly focus on this issue of “short service time” requirement. Our admission control algorithms are based on the service time that each user requests. They are designed not only to maximize the number of served requests but also to minimize average latency by prioritizing the request of short service time for admission. In addition to static traffic conditions, we also consider the bursty traffic characteristics of data applications and apply the algorithms to a number of practical issues about temporal traffic hot spots. Results will be compared with those provided in [9].

1.4.1 Thesis Organization

The remaining of the thesis is organized as follows. In chapter 2, we present preliminary analysis of time-based admission control without buffering requests and chapter 3 presents an analytical system model and our time-based algorithms. In chapter 4, we present discrete rate system model. In chapter 5, we present simulation model and some numerical results respectively. Finally, chapter 6 concludes the thesis and provides some points to future research directions.