QoS Provisioning in Wireless Networks

Dapeng Wu*

Abstract

The next-generation wireless networks such as the fourth generation (4G) cellular systems are targeted at supporting various applications such as voice, data, and multimedia over packet-switched networks. Providing quality of service (QoS) guarantees for these applications is an important objective in the design of the next-generation wireless networks. In this paper, we overview the issues and techniques in QoS provisioning for wireless networks, and present some of our recent results in this area. Specifically, we survey the results in five sub-areas, namely, network services models, traffic specification, packet scheduling for wireless transmission, call admission control in wireless networks, and wireless channel characterization. For each sub-area, we address the particular issues, review major approaches and mechanisms, and discuss the trade-offs of the approaches.

Key Words: QoS, network services model, traffic specification, call admission control, wireless channel modeling, packet scheduling.

*Please direct all correspondence to Prof. Dapeng Wu, University of Florida, Dept. of Electrical & Computer Engineering, P.O.Box 116130, Gainesville, FL 32611, USA. Tel. (352) 392-4954, Fax (352) 392-0044, Email: wu@ece.ufl.edu. URL: http://www.wu.ece.ufl.edu.

1 Introduction

The next-generation wireless networks such as the fourth generation (4G) cellular systems are targeted at supporting various applications such as voice, data, and multimedia over packet-switched networks. In these networks, person-to-person communication can be enhanced with high quality images and video, and access to information and services on public and private networks will be enhanced by higher data rates, quality of service (QoS), security measures, location-awareness, energy efficiency, and new flexible communication capabilities. These features will create new business opportunities not only for manufacturers and operators, but also for providers of content and services using these networks.

Providing QoS guarantees to various applications is an important objective in designing the next-generation wireless networks. Different applications can have very diverse QoS requirements in terms of data rates, delay bounds, and delay bound violation probabilities, among others. For example, applications such as power plant control, demand reliable and timely delivery of control commands; hence, it is critical to guarantee that no packet is lost or delayed during the packet transmission. This type of QoS guarantees is usually called *deterministic* or *hard* guarantees. On the other hand, most multimedia applications including video telephony, multimedia streaming, and Internet gaming, do not require such stringent QoS. This is because these applications can tolerate a certain small probability of QoS violation. This type of QoS guarantees is commonly referred to as *statistical* or *soft* guarantees.

For wireless networks, since the capacity of a wireless channel varies randomly with time, an attempt to provide deterministic QoS (*i.e.*, requiring zero QoS violation probability) will most likely result in extremely conservative guarantees. For example, in a Rayleigh or Ricean fading channel, the deterministically guaranteed capacity[1] (without power control) is zero!

[1]The capacity here is meant to be delay-limited capacity, which is the maximum rate achievable with a prescribed delay bound (see [21] for details).

Table 1: Components in a QoS architecture.

Traffic specification: specifies source traffic characteristics and desired QoS.
QoS routing: provides route(s) between source and destination(s) that have sufficient resources to support the requested QoS.
Call admission control: decides whether a connection request should be accepted or rejected, based on the requested QoS and the network status.
Resource reservation: allots resources such as wireless channels, bandwidth, and buffers at the network elements, which are required to satisfy the QoS guarantees.
Packet scheduling: is to schedule packets to be transmitted according to the QoS requirements of the connections.
Wireless channel characterization: specifies the statistical QoS measure of a wireless channel, *e.g.*, a data rate, delay bound, and delay-bound violation probability triplet.

This conservative guarantee is clearly useless. For this reason, we only consider statistical QoS in this paper.

To support QoS guarantees, two general approaches have been proposed. The first approach is *network-centric.* That is, the routers, switches, and base stations in the network are required to provide QoS support to satisfy data rate, bounded delay, and packet loss requirements requested by applications (*e.g.*, integrated services [11, 14, 49, 62] or differentiated services [10, 25, 40]). The second approach is solely *end-system-based* and does not impose any requirements on the network. In particular, the end systems employ control techniques to maximize the application-layer quality without any QoS support from the transport network. In this paper, we address the problem of QoS provisioning primarily from the network perspective (we refer the interested readers to [55] for the end-system-based approach).

To provide QoS guarantees in wireless networks, a network architecture should contain the following six components: traffic specification, QoS routing, call admission control, wireless channel characterization, resource reservation, and packet scheduling (see Table 1 for description).

The network architecture is illustrated in Figure 1. First, an end system uses a traffic

specification procedure to specify the source traffic characteristics and desired QoS. Then, the network employs QoS routing to find path(s) between source and destination(s) that have sufficient resources to support the requested QoS. At each network node, call admission control decides whether a connection request should be accepted or rejected, based on the requested QoS, the wired link status, and/or the statistics of wireless channels. For base stations, wireless channel characterization is needed to specify the statistical QoS measure of a wireless channel, *e.g.*, a data rate, delay bound, and delay-bound violation probability triplet; this information is used by call admission control. If a connection request is accepted, resource reservation at each network node allots resources such as wireless channels, bandwidth, and buffers that are required to satisfy the QoS guarantees. During the connection life time, packet scheduling at each network node schedules packets to be transmitted according to the QoS requirements of the connections. As shown in Figure 1, in a network node, QoS routing, call admission control, resource allocation, and wireless channel characterization, are functions on the control plane, *i.e.*, performed to set up connections; packet scheduling is a function on the data plane, *i.e.*, performed to transmit packets.

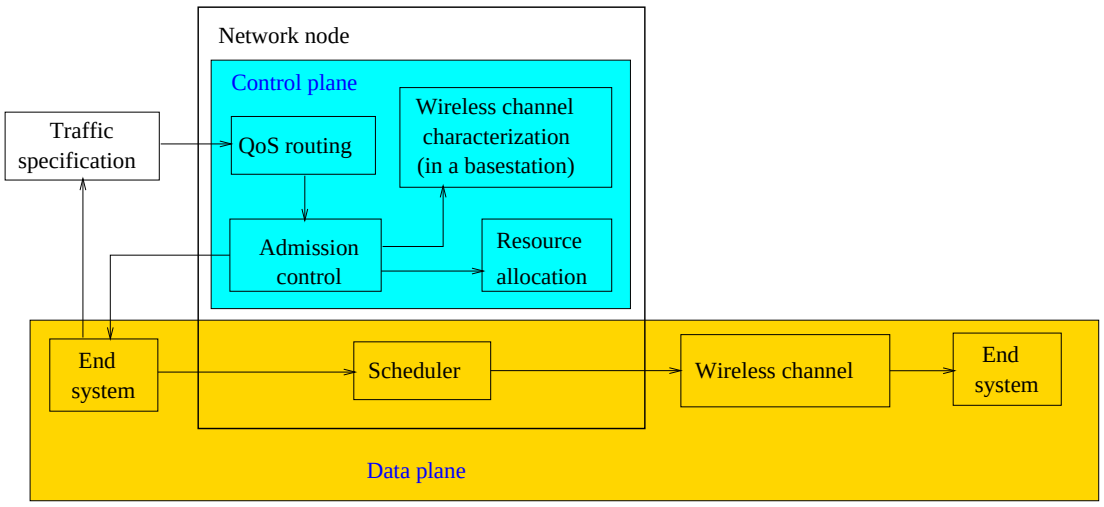

Figure 1: Network architecture for QoS provisioning.

In this paper, we overview the issues and techniques in network-centric QoS provisioning

for wireless networks. The survey is not intended to be exhaustive. In addition, we do not survey results on QoS routing since we do not work on this topic. Resource allocation is often a part of call admission control and hence we do not explicitly survey results on resource allocation.

We organize the rest of the paper as follows. Section 2 presents various network services models. Understanding network service models and associated QoS guarantees is the first step in designing QoS provisioning mechanisms. In Section 3, we overview widely-used traffic models. Section 4 surveys packet scheduling schemes for wireless transmission. In Section 5, we discuss the issue of call admission control in wireless networks. Section 6 addresses wireless channel modeling, which plays an important role in QoS provisioning. Section 7 concludes the paper.

2 Network Services Models

2.1 The Integrated Services Model of the IETF

To support applications with diverse QoS guarantees in IP networks, the IETF Integrated Services (IntServ) Working Group has specified three types of services, namely, the *guaranteed service* [49], the *controlled-load service* [54], and the *best-effort service*.

The guaranteed service (GS) guarantees that packets will arrive within the guaranteed delivery time, and will not be discarded due to buffer overflows, provided that the flow's traffic conforms to its specified traffic parameters [49]. This service is intended for applications which need a hard guarantee that a packet will arrive no later than a certain time after it was transmitted by its sender. That is, the GS does not control the minimal or average delay of a packet; it merely controls the maximal queueing delay. Examples that have hard real-time requirements and require guaranteed service include certain audio and video applications which have fixed playback rates. Delay typically consists of two components,

namely, fixed delay and queueing delay. The fixed delay is a property of the chosen path, which is not determined by the guaranteed service, but rather, by the setup mechanism. Only queueing delay is determined by the GS.

The controlled-load (CL) service is intended to support a broad class of applications which have been developed for use in today's Internet, but are sensitive to heavy load conditions [54]. Important members of this class are the adaptive real-time applications (e.g., *vat* and *vic*) which are offered by a number of vendors and researchers [26]. These applications have been shown to work well over lightly-loaded Internet environment, but to degrade quickly under heavy load conditions. The controlled-load service does not specify any target QoS parameters. Instead, acceptance of a request for controlled-load service is defined to imply a commitment by the network to provide the requester with a service closely approximating the QoS the same flow would receive under lightly-loaded conditions.

Both the guaranteed service and the controlled-load service are designed to support real-time applications which need different levels of QoS guarantee from the network.

The best-effort (BE) service class offers the same type of service under the current Internet architecture. That is, the network makes effort to deliver data packets but makes no guarantees. This works well for non-real-time applications which can use an end-to-end retransmission strategy (i.e., TCP) to make sure that all packets are delivered correctly. These include most popular applications like Telnet, FTP, email, Web browsing, and so on. All of these applications can work without guarantees of timely delivery of data. Another term for such non-real-time applications is *elastic*, since they are able to stretch gracefully in the face of increased delay. Note that these applications can benefit from shorter-length delays but that they do not become unusable as delays increase.

2.2 The Differentiated Services Model of the IETF

The implementation of the IntServ models suffers severe scalability problem. To mitigate it, the IETF specifies the Differentiated Services (DiffServ) framework for the next generation Internet [10, 41]. The DiffServ architecture offers a framework within which service providers can offer each customer a range of network services differentiated on the basis of performance. Once properly designed, a DiffServ architecture can offer great flexibility and scalability, as well as meeting the service requirements for multimedia streaming applications. The IETF DiffServ working group has specified the Assured Forwarding (AF) per hop behavior (PHB) [22]. The AF PHB is intended to provide different levels of forwarding assurances for IP packets at a node and therefore, can be used to implement multiple priority service classes.

2.3 The Services Model of the ATM Forum

For ATM networks, the ATM Forum [4] defines the following services: constant bit rate (CBR), real-time variable bit rate (rt-VBR), non-real-time VBR (nrt-VBR), available bit rate (ABR), and unspecified bit rate (UBR).

Under the CBR service, traffic is specified by its peak cell rate (PCR) and its associated cell delay variation (CDV) tolerance; the connection is serviced at its peak rate at each network node. Under the VBR service, a connection is characterized by PCR, sustainable cell rate (SCR), and the maximum burst size (MBS). The rt-VBR service supports slightly bursty, isochronous streams such as packet voice and video, while the nrt-VBR service is suitable for interactive streams, which are asynchronous, but still delay sensitive. Under the ABR service, the end system transmits its packets at an (instantaneous) rate dynamically set by the network so as to avoid network congestion. Under the UBR service, a connection does not declare traffic parameters and receives no QoS guarantees.

2.4 The Services Model for Wireless Networks

Providing QoS guarantees such as data rate, delay, and loss rate is one of the main features of the next-generation wireless networks. As we mentioned in Section 1, these QoS guarantees can be either deterministic or statistical. However, due to the severely conservative nature of deterministic guarantees, we only consider statistical QoS guarantees in this paper.

In order to support the QoS requested by applications, network designers need to decide what kind of network services should be provided. According to the nature of wireless networks and the QoS guarantees offered, we classify network services into three categories: statistical QoS-assured service, adaptive service, and best-effort. Under best-effort services, no QoS guarantees are supported. Under statistical QoS-assured services, statistical QoS guarantees are explicitly provisioned. We define statistical QoS guarantees of a user as below. Assume that the user is allotted a single time-varying fading channel and the user source has a fixed rate r_s and a specified delay bound D_{max}, and requires that the delay-bound violation probability is not greater than a certain value ε, that is,

$$Pr\{D(\infty) > D_{max}\} \leq \varepsilon, \tag{1}$$

where $D(\infty)$ is the steady-state delay experienced by a flow, and $Pr\{D(\infty) > D_{max}\}$ is the probability of $D(\infty)$ exceeding a delay bound D_{max}. Then, we say that the user is specified by the (statistical) QoS triplet $\{r_s, D_{max}, \varepsilon\}$. This QoS triplet is essential in the design of statistical QoS provisioning mechanisms [60].

Adaptive services provide mechanisms to adapt traffic streams during periods of QoS fluctuations and hand-offs [56]. Adaptive services have been demonstrated to be able to effectively mitigate fluctuations of resource availability in wireless networks [5]. There have been many proposals on adaptive approaches and services in the literature, which include an "adaptive reserved service" framework [32], a wireless adaptive mobile information system (WAMIS) [2], an adaptive service based on QoS bounds and revenue [36], an adaptive

framework targeted at end-to-end QoS provisioning [38], a utility-fair adaptive service [9], a framework for soft QoS control [47], a teleservice model based on an adaptive QoS paradigm [24], an adaptive QoS framework called AQuaFWiN [52], and an adaptive QoS management architecture [31], among others. Although adaptive services provide a service better than best effort, no explicit QoS guarantees is enforced.

3 Traffic Modeling

Traffic modeling plays an important role in QoS provisioning. It facilitates traffic specifications and accurate call admission control. Without a traffic model or characterization, measurement-based admission control needs to be employed with reduced accuracy and efficiency, compared to traffic-specification based admission control.

Traffic models fall into two categories: CBR and VBR as shown in Figure 2. For VBR, the traffic models can be either deterministic or stochastic.

The most commonly used deterministic model is linear bounded arrival process (LBAP) [16]. It has two parameters: token generating rate ρ and token bucket size σ. The amount of source traffic over any time interval of length τ is upper bounded by $\rho\tau + \sigma$, *i.e.*,

$$A(t, t+\tau) \leq \rho\tau + \sigma, \tag{2}$$

where $A(t, t+\tau)$ is the amount of source traffic generated during $[t, t+\tau)$. The simplicity of the LBAP model makes it very useful for traffic shaping and policing, which are required to ensure that the source traffic conforms to the declared characterization. Actually, the (σ, ρ)-leaky bucket regulator [51], which is the most widely used traffic shaping and policing scheme, produces traffic that can be characterized by the LBAP model. Hence, the LBAP is also called *leaky-bucket constrained* traffic model.

In stochastic modeling, there are three common approaches: the first approach is to use stochastic processes such as Markov processes to model the traffic arrival process it-

self; the second one is a stochastic bounding approach, *e.g.*, the exponentially bounded burstiness (EBB) model [61] provides an upper bound on the probability of violating the LBAP constraint; the third approach uses large deviations theory, specifically, the asymptotic log-moment generating function of the traffic process, to characterize the traffic. Effective bandwidth [12, 13] and self-similar traffic models [33] are two important stochastic models.

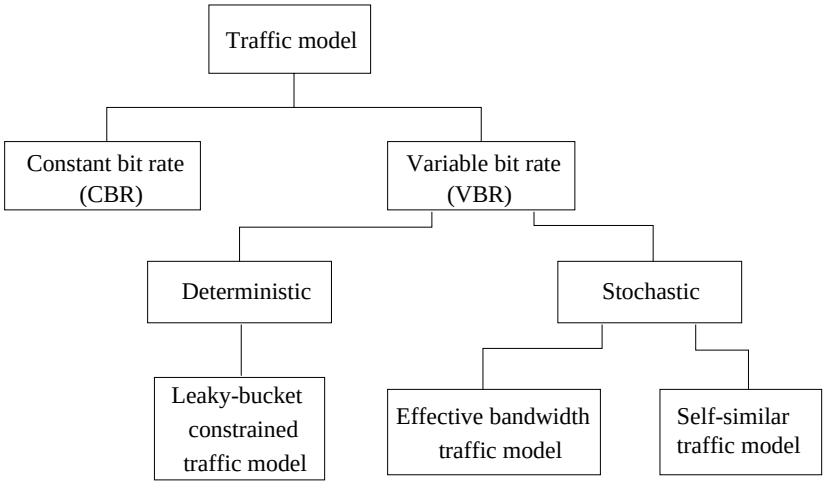

Figure 2: Classification of traffic models.

4 Scheduling for Wireless Transmission

Packet scheduling is an important QoS provisioning mechanism at a network node. Compared with the scheduler design for the wired networks, the design of scheduling for wireless networks with QoS guarantees, is particularly challenging. This is because wireless channels have low reliability, and time varying signal strength, which may cause severe QoS violations. Further, the capacity of a wireless channel is severely limited, making efficient bandwidth utilization a priority.

In wired networks, the task of packet scheduling is to associate a packet with a time slot. In wireless networks, packet scheduling can be more general than that; its function is to schedule such resources as time slots, powers, data rates, channels, or combination of

them, when packets are transmitted. (Note that a wired scheduler does not assign powers, data rates, and channels since packets are transmitted at a constant power, a constant data rate or link speed, and through one shared channel.)[2] Specifically, based on the source characteristics, QoS requirements, channel states, and perhaps the queue lengths, a wireless scheduler assigns time slots, powers, data rates, channels, or combinations of them, to the packets for transmission. For example, in TDMA systems, time slots, powers, and data rates can be scheduled [15, 44]; in FDMA systems, channels (*i.e.*, frequencies) can be assigned [59] (see Figure 3(b)); in CDMA systems, powers, channels (*i.e.*, signature sequences), and data rates (*i.e.*, variable spreading factor) can be allotted [6, 17, 48, 53].

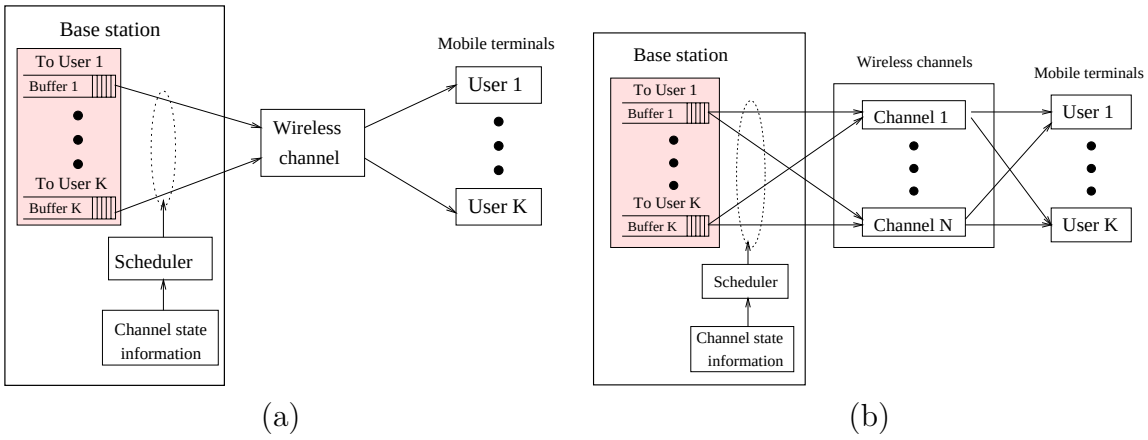

Figure 3: Wireless schedulers: (a) single channel, and (2) multiple channels.

A unique feature of wireless scheduling with QoS guarantees is its *channel state dependency*, *i.e.*, how to schedule the resources depends on the channel state (see Figure 3). This is necessary since without the knowledge about the channel state, it is impossible to guarantee QoS! A key difference between a wired scheduler and a wireless scheduler is that a wireless scheduler can utilize *asynchronous channel variations* or *multiuser diversity* [58] while a wired scheduler cannot.

[2]Here, we assume all flows share one wired channel/link. For the multiple shared channel case, a switch needs to be used.

Except the aforementioned differences, wireless and wired schedulers with QoS guarantees perform the same functions as below.

- *Isolation*: the scheduler supports the implementation of network service classes and provides isolation among these service classes in order to prevent one class from interfering with another class in achieving its QoS guarantees;

- *Sharing*: the scheduler controls bandwidth sharing among various service classes, and among flows in the same class, so that 1) statistical multiplexing gains can be exploited to efficiently utilize network resources, and 2) fair sharing of bandwidth among classes and sessions can be enforced.

There have been many proposals on scheduling with QoS constraints in wireless networks (see [19] for a survey). These schedulers fall into two classes: work-conserving and non-work-conserving. A work-conserving scheduler is never idle if there is a packet awaiting transmission. Examples include wireless fair queueing [37, 39, 45], modified largest weighted delay first (M-LWDF) [3], opportunistic transmission scheduling [34], lazy packet scheduling [43], Knopp and Humblet's (K&H) scheduling [28], Bettesh and Shamai's scheduling [7] and dynamic programming-based scheduling [8]. In contrast, a non-work-conserving scheduler may be idle even if there is a backlogged packet in the queue because the packet may not be eligible for transmission. Examples are weighted round robin, the joint Knopp&Humblet/round robin (K&H/RR) scheduler [58] and the reference channel scheduler [59]. A non-work-conserving scheduler is an important component of a hierarchical scheduler supporting both QoS-assured flows and best-effort traffic. Such a hierarchical scheduler for wireless transmission was described in [56].

Next we briefly overview some representative scheduling algorithms mentioned above.

Wireless fair queueing schemes [37, 39, 45] are aimed at applying wired fair queueing [42] to wireless networks. The objective of these schemes is to provide fairness, while providing

loose QoS guarantees. Although these schedulers make decisions based on the channel state information (*i.e.*, good or bad channel), they do not exploit asynchronous channel variations to improve efficiency since packets destined to different users are transmitted at the same bit-rate.

The M-LWDF algorithm [3] and the opportunistic transmission scheduling [34] optimize a certain QoS parameter or utility index. They both exploit asynchronous channel variations and allow different user to transmit at different bit-rate or signal-to-interference-noise ratio (SINR), so that higher efficiency can be achieved. However, they do not provide the explicit QoS guarantees such as data rate, delay bound, and delay-bound violation probability.

The lazy packet scheduling [43] is targeted at minimizing energy, subject to a delay constraint. The scheme only considers AWGN channels and thus allows for a deterministic delay bound, unlike fading channels and the general statistical QoS considered in this dissertation.

Time-division scheduling (or weighted round robin) has been proposed for 3-G WCDMA [23, page 226]. However, their proposal did not provide methods on how to use time-division scheduling to support statistical QoS guarantees explicitly.

In the K&H scheduling [28], Knopp and Humblet utilized a kind of diversity, which is inherent in a wireless network with multiple users sharing a time-varying channel. This diversity, termed *multiuser diversity* [20], comes from the fact that different users usually have *independent* channel gains for the same shared medium. With multiuser diversity, the strategy of maximizing the total Shannon (ergodic) capacity is to allow at any time slot only the user with the best channel to transmit. This strategy is called Knopp and Humblet's (K&H) scheduling. Results [28] have shown that the K&H scheduling can increase the total (ergodic) capacity dramatically, in the absence of delay constraints, as compared to the traditionally used (weighted) round robin (RR) scheduling where each user is *a priori* allocated fixed time slots.

The K&H scheduling intends to maximize ergodic capacity, which pertains to situations

of infinite tolerable delay. However, under this scheme, a user in a fade of an arbitrarily long period will not be allowed to transmit during this period, resulting in an arbitrarily long delay; therefore, this scheme provides no delay guarantees and thus is not suitable for delay-sensitive applications, such as voice or video. To mitigate this problem, Bettesh and Shamai [7] proposed an algorithm (called Bettesh and Shamai's scheduler), which strikes a balance between throughput and delay constraints. This algorithm combines the K&H scheduling with an RR scheduling, and it can achieve lower delay than the K&H scheduling while obtaining a capacity gain over a pure RR scheduling. However, it is very complex to theoretically relate the QoS obtained by this algorithm to the control parameters of the algorithm, and thus cannot be used to guarantee a specified QoS.

Bettesh and Shamai [8] also proposed a dynamic programming-based scheduler that can increase capacity, while also maintaining QoS guarantees. But the dynamic programming approach suffers from the curse of dimensionality, since the size of the dynamic program state space grows exponentially with the number of users and with the delay requirement.

To address the limitation of the above scheduling algorithms, *i.e.*, inability of provisioning explicit QoS, we proposed the joint K&H/RR scheduler in [58] and the reference channel scheduler in [59], respectively.

Our joint K&H/RR scheduling [58] simplifies the task of explicit provisioning of QoS guarantees while achieving efficiency in utilizing wireless channel resources (due to multiuser diversity). Specifically, we design our scheduler based on the K&H scheduling, but shift the burden of QoS provisioning to the resource allocation mechanism, thus simplifying the design of the scheduler. Such a partitioning would be meaningless if the resource allocation problem now becomes complicated. However, we are able to solve the resource allocation problem efficiently using the method of *effective capacity* developed in [57]. Effective capacity captures the effect of channel fading on the queueing behavior of the link, using a computationally simple yet accurate model, and thus, is the critical device we need to design an efficient

resource allocation mechanism. Our results show that compared to the RR scheduling, our joint K&H/RR scheduling can substantially increase the statistical delay-constrained capacity (a.k.a., effective capacity [57]) of a fading channel, when delay requirements are not very tight. For example, in the case of low signal-to-noise-ratio (SNR) and ergodic Rayleigh fading, our joint K&H/RR scheduling can achieve approximately $\sum_{k=1}^{K} \frac{1}{k}$ gain for K users with loose-delay requirements. But more importantly, when the delay bound is not loose, so that simple-minded K&H scheduling does not directly apply, our joint K&H/RR scheduling can achieve a capacity gain, and yet meet the QoS requirements.

In [59], we extended our work in [58] to the setting of multiple users sharing *multiple* parallel channels, by utilizing both multiuser diversity and frequency diversity. We first applied the joint K&H/RR scheduler in [58] to the multiple channel case. Due to the frequency diversity inherent in multiple wireless channels, the joint K&H/RR scheduler in the new setting can achieve higher capacity gain than that in [58], when delay requirements are loose or moderate. However, we then noted that when users' delay requirements are stringent, the joint K&H/RR reduces to the RR scheduling, and so the high capacity gain due to multiuser diversity associated with the K&H scheduling, vanishes. To extract more capacity in this case with tight delay requirements, it is desirable to have a scheduler, which at each instant, dynamically selects the best channel among multiple channels for each user to transmit, so as to obtain frequency diversity. In other words, this scheduler must find a channel-assignment schedule, at each time-slot, which minimizes the channel usage while yet satisfying users' QoS requirements. We therefore formulated this scheduling problem as a linear program, in order to avoid the 'curse of dimensionality' associated with optimal dynamic programming solutions. The key idea that allows us to do this, is what we call the reference channel approach, wherein the QoS requirements of the users, are captured by resource allocation (channel assignments). The reference channel approach allows us to obtain capacity gain under tight QoS constraints, by utilizing frequency diversity. Hence, we call the resulting scheduling as *reference channel* scheduling.

5 Call Admission Control in Wireless Networks

The objective of call admission control (CAC) is to provide QoS guarantees for individual connections while efficiently utilizing network resources. Specifically, a CAC algorithm makes the following decision:

> Given a call arriving to a network, can it be admitted by the network, with its requested QoS satisfied and without violating the QoS guarantees made to the existing connections?

The decision is made based on the availability of network resources as well as traffic specifications and QoS requirements of the users. If the decision is affirmative, necessary network resources need to be reserved to support the QoS. Hence, CAC is closely related to channel allocation, base station assignment, scheduling, power control, and bandwidth reservation. For example, whether the channel assignment is dynamic or fixed will result in different CAC algorithms.

The CAC problem can be formulated as an optimization problem, *i.e.*, maximize the network efficiency/utility/revenue subject to the QoS constraints of connections. The QoS constraints could be signal-to-interference ratio (SIR), the ratio of bit energy to interference density E_b/I_0, bit error rate (BER), call dropping probability, or connection-level QoS (such as a data rate, delay bound, and delay-bound violation probability triplet). For example, a CAC problem can be maximizing the number of users admitted or minimizing the blocking probability, subject to the BER violation probability not more than a required value ε_1, *i.e.*,

$$\text{maximize the number of users admitted} \tag{3}$$

$$\text{or minimize the blocking probability} \tag{4}$$

$$\text{subject to} \quad \Pr\{\ \text{BER} > \text{BER}_{th}\} \leq \varepsilon_1 \tag{5}$$

where BER_{th} denotes a threshold for the BER. The constraint (5) can be replaced by

$$\text{subject to} \qquad \text{the dropping probability} \leq \varepsilon_2 \tag{6}$$

where the value of ε_2 may be different from that of ε_1.

CAC can also be used to provide priority to some service classes, or to enforce some policies like fair resource sharing, which includes complete sharing, complete partitioning, and threshold-based sharing.

There have been many algorithms on CAC in wireless networks (refer to [1] for a survey). These CAC algorithms may differ in admission criteria; they may be centralized or distributed; they may use global (all-cell) or local (single-cell) information about resource availability and interference levels to make admission decisions. The design of distributed CAC for cellular networks is not an easy task since intra-cell and inter-cell interference needs to be considered. The associated intra-cell and inter-cell resource allocation are complicated due to the interference.

A typical admission criterion is SIR. For example, Ref. [35] employs SIR to define a measure called *residual capacity*, and uses it as the admission criterion: if the residual capacity is positive, accept the new call; otherwise, reject it. Ref. [18] uses the concept of effective bandwidth[3] to measure whether the signal to interference density ratio (SIDR) can be satisfied for each class with certain probability. There, SIDR is defined as $r_s \times E_b/I_0$, where r_s denotes the source data rate in bits/sec. If the total effective bandwidth including that for the new call, is less than the available bandwidth, the new call will be accepted; otherwise, it will be rejected.

Another important admission criterion is transmitted or received power. In [29], a new call is admitted if the total transmitted power does not exceed a preset value. In [30], the CAC uses the 95-percentile of the total received power as the admission criterion.

[3]Note that the effective bandwidth defined in [18] is different from that defined in [12].

However, none of the existing CAC algorithms provides explicit connection-level QoS guarantees such as a data rate, delay bound, and delay-bound violation probability triplet. In [58, 59], we proposed simple CAC algorithms that are capable of providing connection-level QoS guarantees explicitly. The key idea of our CAC algorithms is the following. We first measure the QoS exponent function $\theta(\mu)$ (defined in Section 6) for a given channel. Then we use $\theta(\mu)$ to check the feasibility of user's statistical QoS requirements specified by $\{r_s, D_{max}, \varepsilon\}$. Specifically, the channel can support a QoS triplet $\{r_s, D_{max}, \varepsilon\}$ if $\theta(r_s) \geq \rho$, where $\rho \doteq -\log \varepsilon / D_{max}$.

6 Wireless Channel Modeling

Figure 4 shows a wireless communication system. The data source generates packets and the packets are first put into a buffer to accommodate the mismatch between the source rate and the time-varying wireless channel capacity. Then the packets traverse a channel encoder, a modulator, a wireless channel, a demodulator, a channel decoder, a network access device, and finally reach the data sink.

As shown in Figure 4, one can model the communication channel at different layers as below

- Radio-layer channel: is the part between the output of the modulator and the input of the demodulator.
- Modem-layer channel: is the part between the output of the channel encoder and the input of the channel decoder.
- Codec-layer channel: is the part between the output of the network access device at the transmitter, and the input of the network access device at the receiver.
- Link-layer channel: is the part between the output of the data source and the input of

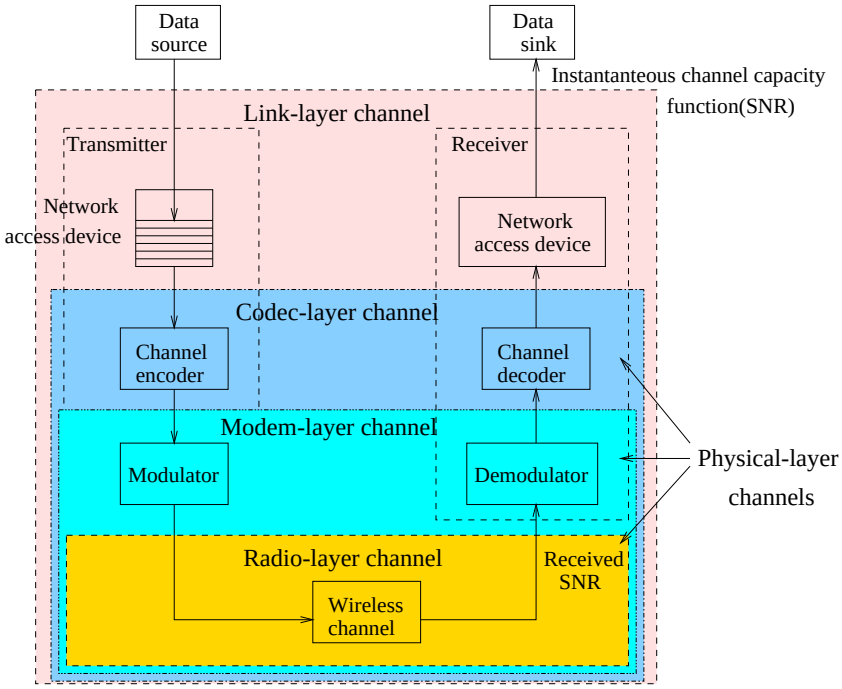

Figure 4: A wireless communication system and associated channel models.

the data sink.

The above radio-layer, the modem-layer, and the codec-layer channels can all be regarded as physical-layer channels.

As shown in Figure 5, radio-layer channel models can be classified into two categories: large-scale path loss and small-scale fading. Large-scale path loss models, also called propagation models, characterize the underlying physical mechanisms (*i.e.*, reflection, diffraction, scattering) for specific paths. These models specify signal attenuation as a function of distance, which is affected by prominent terrain contours (buildings, hills, forests, etc.) between the transmitter and the receiver. Path loss models describe the mean signal attenuation vs. distance in a deterministic fashion (*e.g.*, nth-power law [46]), and also the statistical variation about the mean (*e.g.*, log-normal distribution [46]).

Small-scale fading models describe the characteristics of generic radio paths in a statistical

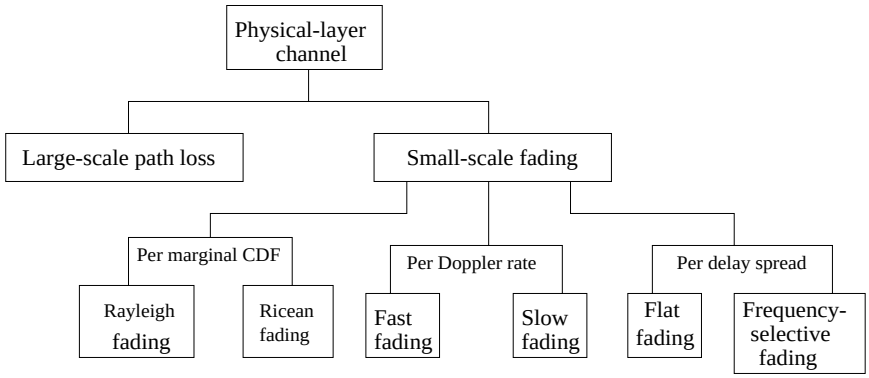

Figure 5: Classification of physical-layer channel models.

fashion. Small-scale fading refers to the dramatic changes in signal amplitude and phase that can be experienced as a result of small changes (as small as a half-wavelength) in the spatial separation between the receiver and the transmitter [50]. Small-scale fading can be slow or fast, depending on the Doppler rate. Small-scale fading can also be flat or frequency-selective, depending on the delay spread of the channel. The statistical time-varying nature of the envelope of a flat-fading signal is characterized by distributions such as Rayleigh, Ricean, Nakagami, etc. [46]. Uncorrelated scattering is often assumed, to extend these distributions to the frequency-selective case. The large-scale path loss and small-scale fading together characterize the received signal power over a wide range of distances.

A modem-layer channel can be modeled by a finite-state Markov chain [63], whose states are characterized by different bit error rates (BER). For example, in [63], a Rayleigh fading with certain Doppler spectrum is converted to a BER process, modeled by a finite-state Markov chain. The idea is the following: 1) quantize the continuous Rayleigh random variable into a discrete random variable, based on certain optimal criterion (*e.g.*, minimum mean squared error), 2) map the resulting discrete random variable or SNR to discrete BER, for a given modulation scheme (say, binary phase shift keying), and 3) estimate the state transition probabilities, which reflect the Doppler spectrum. This procedure gives the states (*i.e.*, BER's) and the transition probability matrix of the Markov chain.

A codec-layer channel can also be modeled by a finite-state Markov chain, whose states can be characterized by different data-rates [27], or symbol being error-free/in-error, or channel being good/bad [64]. The two state Markov chain model with good/bad states [64] is widely used in analyzing the performance of upper layer protocols such as TCP [65]. If the decoder uses hard decisions from the demodulator/detector, a codec-layer channel model can be easily obtained from a modem-layer channel model. For example, the good/bad channel model can be derived from a finite-state Markov chain with BER's as the states in the following way: first compute symbol error probability from BER; then decide the channel being good if the symbol error probability is less than a preset threshold, otherwise decide the channel being bad. The resulting good/bad channel process is a two state Markov chain.

Radio-layer channel models provide a quick estimate of the performance of wireless communications systems (*e.g.*, symbol error rate vs. signal-to-noise ratio (SNR)). However, radio-layer channel models cannot be easily translated into complex QoS guarantees for a connection, such as bounds on delay violation probability and packet loss ratio. The reason is that, these complex QoS requirements need an analysis of the queueing behavior of the connection, which is hard to extract from radio-layer models [57]. Thus it is hard to use radio-layer models in QoS support mechanisms, such as admission control and resource reservation.

Finite-state Markov chain models for a modem-layer or codec-layer channel also require a queueing analysis of very high complexity to obtain connection-level QoS such as a data rate, delay bound, and delay-bound violation probability triplet. We showed this high complexity through an example in [60, pp. 123–125].

Recognizing that the limitation of the physical-layer channel models in QoS support, is the difficulty in analyzing queues, we propose moving the channel model up the protocol stack, from the physical-layer to the link-layer. We call the resulting link-layer channel model *effective capacity* model, because it captures a generalized link-level capacity notion

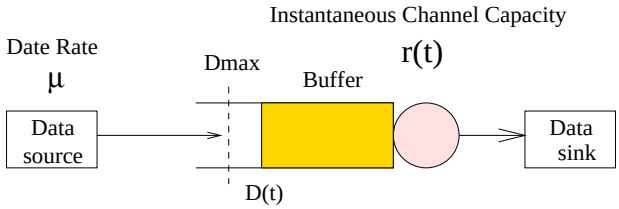

Figure 6: A queueing system model.

of the fading channel. Next, we briefly explain the concept of effective capacity, and refer the reader to [57] for details.

Let $r(t)$ be the instantaneous channel capacity at time t. The *effective capacity function* of $r(t)$ is defined as [57]

$$\alpha(u) = -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}], \ \forall\ u > 0. \tag{7}$$

Consider a queue of infinite buffer size supplied by a data source of *constant* data rate μ (see Figure 6). It can be shown [57] that if $\alpha(u)$ indeed exists (*e.g.*, for ergodic, stationary, Markovian $r(t)$), then the probability of $D(\infty)$ exceeding a delay bound D_{max} satisfies

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta(\mu)D_{max}}, \tag{8}$$

where the function $\theta(\mu)$ of source rate μ depends only on the channel capacity process $r(t)$. The approximation (8) is accurate for large D_{max}. $\theta(\mu)$ **is the effective capacity channel model** that models the channel at the link layer (in contrast to "physical layer" models).

In terms of the effective capacity function (7) defined earlier, the *QoS exponent function* $\theta(\mu)$ can be written as [57]

$$\theta(\mu) = \mu\alpha^{-1}(\mu) \tag{9}$$

where $\alpha^{-1}(\cdot)$ is the inverse function of $\alpha(u)$. Once $\theta(\mu)$ has been measured for a given channel, it can be used to check the feasibility of QoS triplets. Specifically, a QoS triplet $\{r_s, D_{max}, \varepsilon\}$ is feasible if $\theta(r_s) \geq \rho$, where $\rho \doteq -\log\varepsilon/D_{max}$. Thus, we can use the effective

capacity model $\alpha(u)$ (or equivalently, the function $\theta(\mu)$ via (9)) to relate the channel capacity process $r(t)$ to the statistical QoS guarantees specified by $\{r_s, D_{max}, \varepsilon\}$. In [57], we presented a simple and efficient algorithm to estimate $\theta(\mu)$ by direct measurement of the queueing behavior resulting from $r(t)$.

Our effective capacity channel model characterizes wireless channels in terms of functions that can be easily mapped to link-level QoS metrics, such as delay-bound violation probability. Since effective capacity captures the effect of channel fading on the queueing behavior of the link, using a computationally simple yet accurate model, the effective capacity channel model is a critical tool for designing efficient QoS provisioning mechanisms.

7 Concluding Remarks

Due to the marriage of wireless networks and Internet, high-speed next-generation wireless networking has gained strong momentum. To support various applications such as voice, data, and multimedia in the next-generation wireless networks, providing QoS guarantees for these applications is particularly important. In this paper, we discussed the issues and techniques in QoS provisioning for wireless networks, and presented some of our recent results in this area. Specifically, we surveyed the results in five sub-areas, namely, network services models, traffic specification, packet scheduling for wireless transmission, call admission control in wireless networks, and wireless channel characterization. For each sub-area, we examined important issues, reviewed major approaches and mechanisms, and discussed the trade-offs of the approaches. The objective of this paper is not to provide an exhaustive review of existing approaches and mechanisms, but instead to give the reader a perspective on the range of options available and the associated trade-offs.

References

[1] M. Ahmed, H. Yanikomeroglu, and S. Mahmoud, "Call admission control in wireless communications: a comprehensive survey," to be submitted to *IEEE Wireless Communications Magazine.*

[2] A. Alwan, R. Bagrodia, N. Bambos, M. Gerla, L. Kleinrock, J. Short, and J. Villasenor, "Adaptive mobile multimedia networks," *IEEE Personal Communications Magazine,* vol. 3, no. 2, pp. 34–51, April 1996.

[3] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine,* vol. 39, no. 2, pp. 150–154, Feb. 2001.

[4] ATM Forum Technical Committee, "Traffic management specification (version 4.0)," ATM Forum, Feb. 1996.

[5] A. Balachandran, A. T. Campbell, M. E. Kounavis, "Active filters: delivering scalable media to mobile devices," in *Proc. Seventh International Workshop on Network and Operating System Support for Digital Audio and Video (NOSSDAV'97),* St Louis, MO, USA, May 1997.

[6] N. Bambos and S. Kandukuri, "Power-controlled multiple access schemes for next-generation wireless packet networks," *IEEE Wireless Communications Magazine,* vol. 9, no. 3, pp. 58–64, June 2002.

[7] I. Bettesh and S. Shamai, "A low delay algorithm for the multiple access channel with Rayleigh fading," in *Proc. IEEE Personal, Indoor and Mobile Radio Communications (PIMRC'98),* 1998.

[8] I. Bettesh and S. Shamai, "Optimal power and rate control for fading channels," in *Proc. IEEE Vehicular Technology Conference,* Spring 2001.

[9] G. Bianchi, A. T. Campbell, and R. Liao, "On utility-fair adaptive services in wireless networks," in *Proc. 6th International Workshop on Quality of Service (IWQOS'98),* Napa Valley, CA, May 1998.

[10] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated services," *RFC 2475,* Internet Engineering Task Force, Dec. 1998.

[11] R. Braden, D. Clark, and S. Shenker, "Integrated services in the Internet architecture: An overview," *RFC 1633,* Internet Engineering Task Force, July 1994.

[12] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications,* vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[13] C.-S. Chang, "Performance guarantees in communication networks," Springer, 2000.

[14] D. Clark, S. Shenker, and L. Zhang, "Supporting real-time applications in an integrated services packet network: architecture and mechanisms," in *Proc. ACM SIGCOMM'92,* Baltimore, MD, Aug. 1992.

[15] B. E. Collins and R. L. Cruz, "Transmission policies for time-varying channels with average delay constraints," in *Proc. Allerton Conference on Communication, Control, and Computing,* Monticello IL, Sept. 1999.

[16] R. L. Cruz, "A calculus for network delay, Part I: network elements in isolation," *IEEE Trans. on Information Theory,* vol. 37, no. 1, pp. 114–131, Jan. 1991.

[17] M. Elaoud and P. Ramanathan, "Adaptive allocation of CDMA resources for network level QoS assurances," in *Proc. ACM Mobicom'00,* Aug. 2000.

[18] J. S. Evans and D. Everitt, "Effective bandwidth-based admission control for multiservice CDMA cellular networks," *IEEE Trans. on Vehicular Technology,* vol. 48, no. 1, pp. 36–46, Jan. 1999.

[19] H. Fattah and C. Leung, "An overview of scheduling algorithms in wireless multimedia networks," *IEEE Wireless Communications Magazine,* vol. 9, no. 5, pp. 76–83, Oct. 2002.

[20] M. Grossglauser and D. Tse, "Mobility increases the capacity of wireless adhoc networks," in *Proc. IEEE INFOCOM'01,* April 2001.

[21] S. Hanly and D. Tse, "Multi-access fading channels: part II: delay-limited capacities," *IEEE Trans. on Information Theory,* vol. 44, no. 7, pp. 2816–2831, Nov. 1998.

[22] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski, "Assured forwarding PHB group," *RFC 2597,* Internet Engineering Task Force, June 1999.

[23] H. Holma and A. Toskala, *WCDMA for UMTS: Radio Access for Third Generation Mobile Communications,* Wiley, 2000.

[24] A. Iera, A. Molinaro, and S. Marano, "Wireless broadband applications: the teleservice model and adaptive QoS provisioning," *IEEE Communications Magazine,* vol. 37, no. 10, pp. 71–75, Oct. 1999.

[25] V. Jacobson, K. Nichols, and K. Poduri, "An expedited forwarding PHB," *RFC 2598,* Internet Engineering Task Force, June 1999.

[26] S. Jamin, P. B. Danzig, S. Shenker and L. Zhang, "A measurement-based admission control algorithm for integrated services packet networks," *IEEE/ACM Trans. on Networking,* vol. 5, no. 1, pp. 56–70, Feb. 1997.

[27] Y. Y. Kim and S.-Q. Li, "Capturing important statistics of a fading/shadowing channel for network performance analysis," *IEEE Journal on Selected Areas in Communications,* vol. 17, no. 5, pp. 888–901, May 1999.

[28] R. Knopp and P. A. Humblet, "Information capacity and power control in single-cell multiuser communications," in *Proc. IEEE International Conference on Communications (ICC'95),* Seattle, USA, June 1995.

[29] J. Knutsson, P. Butovitsch, M. Persson, R. D. Yates, "Downlink admission control strategies for CDMA systems in a Manhattan environment," *Proc. IEEE 48th Vehicular Technology Conference (VTC98),* May 1998.

[30] J. Kuri and P. Mermelstein, "Call admission on the uplink of a CDMA system based on total received power," *Proc. IEEE International Conference on Communications (ICC'99),* June 1999.

[31] O. Lataoui, T. Rachidi, L. G. Samuel, S. Gruhl, and R.-H., Yan, "A QoS management architecture for packet switched 3rd generation mobile systems," in *Proc. Networld+Interop 2000 Engineers Conference,* Las Vegas, Nevada, USA, May 10–11, 2000.

[32] K. Lee, "Adaptive network support for mobile multimedia," in *Proc. ACM Mobicom'95,* Berkeley, CA, USA, Nov. 13–15, 1995.

[33] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson, "On the self-similar nature of Ethernet traffic," *IEEE/ACM Transactions on Networking,* vol. 2, no. 1, pp. 1–15, Feb. 1994.

[34] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE Journal on Selected Areas in Communications,* vol. 19, no. 10, pp. 2053–2064, Oct. 2001.

[35] Z. Liu and M. El Zarki, "SIR-based call admission control for DS-CDMA cellular systems," *IEEE Journal on Selected Areas in Communications,* vol. 12, no. 4, pp. 638–644, May 1994.

[36] S. Lu, K.-W. Lee, and V. Bharghavan, "Adaptive service in mobile computing environments," in *Proc. 5th International Workshop on Quality of Service (IWQOS'97),* Columbia University, New York, May 21–23, 1997.

[37] S. Lu, V. Bharghavan, and R. Srikant, "Fair scheduling in wireless packet networks," *IEEE/ACM Trans. on Networking,* vol. 7, no. 4, pp. 473–489, Aug. 1999.

[38] M. Naghshineh and M. Willebeek-LeMair, "End-to-end QoS provisioning in multimedia wireless/mobile networks using an adaptive framework," *IEEE Communications Magazine,* vol. 35, no. 11, pp. 72–81, Nov. 1997.

[39] T. S. E. Ng, I. Stoica, and H. Zhang, "Packet fair queueing algorithms for wireless networks with location-dependent errors," in *Proc. IEEE INFOCOM'98,* pp. 1103–1111, San Francisco, CA, USA, March 1998.

[40] K. Nichols, V. Jacobson, and L. Zhang, "A two-bit differentiated services architecture for the Internet," *RFC 2638,* Internet Engineering Task Force, July 1999.

[41] K. Nichols, S. Blake, F. Baker and D. Black, "Definition of the differentiated services field (DS field) in the IPv4 and IPv6 headers," *RFC 2474,* Internet Engineering Task Force, Dec. 1998.

[42] A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks: the single node case," *IEEE/ACM Trans. on Networking,* vol. 1, no. 3, pp. 344–357, June 1993.

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

[44] D. Rajan, A. Sabharwal, and B. Aazhang, "Delay and rate constrained transmission policies over wireless channels," in *Proc. IEEE GLOBECOM'01,* Nov. 2001.

[45] P. Ramanathan and P. Agrawal, "Adapting packet fair queueing algorithms to wireless networks," in *Proc. ACM MOBICOM'98,* Oct. 1998.

[46] T. S. Rappaport, *Wireless Communications: Principles & Practice,* Prentice Hall, 1996.

[47] D. Reininger, R. Izmailov, B. Rajagopalan, M. Ott, and D. Raychaudhuri, "Soft QoS control in the WATMnet broadband wireless system," *IEEE Personal Communications Magazine,* vol. 6, no. 1, pp. 34–43, Feb. 1999.

[48] A. Sampath, P. S. Kumar, and J. M. Holtzman, "Power control and resource management for a multimedia CDMA wireless system," in *Proc. IEEE PIMRC'95,* Sept. 1995.

[49] S. Shenker, C. Partridge, and R. Guerin, "Specification of guaranteed quality of service," *RFC 2212,* Internet Engineering Task Force, Sept. 1997.

[50] B. Sklar, "Rayleigh fading channels in mobile digital communication systems Part I: characterization," *IEEE Communications Magazine,* vol. 35, no. 7, pp. 90–100, July 1997.

[51] J. S. Turner, "New directions in communications (or which way to the information age?)," *IEEE Communications Magazine,* vol. 24, no. 10, pp. 8–15, Oct. 1986.

[52] B. Vandalore, R. Jain, S. Fahmy, and S. Dixit, "AQuaFWiN: adaptive QoS framework for multimedia in wireless networks and its comparison with other QoS frameworks," in *Proc. IEEE Conference on Local Computer Networks (LCN'99),* Boston, MA, USA, Oct. 17–20, 1999.

[53] H. Wang, "Opportunistic transmission for wireless data over fading channels under energy and delay constraints," *Ph.D. Dissertation,* Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey New Brunswick, New Jersey, Jan. 2003.

[54] J. Wroclawski, "Specification of the controlled-load network element service," *RFC 2211,* Internet Engineering Task Force, Sept. 1997.

[55] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Transporting real-time video over the Internet: challenges and approaches," *Proceedings of the IEEE,* vol. 88, no. 12, Dec. 2000.

[56] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Scalable video soding and transport over broadband wireless networks," *Proceedings of the IEEE,* vol. 89, no. 1, Jan. 2001.

[57] D. Wu and R. Negi, "Effective capacity: a wireless link model for support of quality of service," *IEEE Trans. on Wireless Communications,* vol. 2, no. 4, pp. 630–643, July 2003.

[58] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE ICC'03,* Anchorage, Alaska, USA, May 2003.

[59] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality of service assurance," *IEEE Vehicular Technology Conference (VTC) Fall 2003,* Orlando, Florida, USA, October 2003.

[60] D. Wu, "Providing quality of service guarantees in wireless networks," *Ph.D. Dissertation,* Dept. of Electrical & Computer Engineering, Carnegie Mellon University, Aug. 2003. Available at `http://www.wu.ece.ufl.edu/mypapers/Thesis.pdf`.

[61] O. Yaron and M. Sidi, "Performance and stability of communication networks via robust exponential bounds," *IEEE/ACM Trans. on Networking,* vol. 1, no. 3, pp. 372–385, June 1993.

[62] L. Zhang, S. Deering, D. Estrin, S. Shenker, and D. Zappala, "RSVP: A new resource ReSerVation Protocol," *IEEE Network Magazine,* vol. 7, no, 5, pp. 8–18, Sept. 1993.

[63] Q. Zhang and S. A. Kassam, "Finite-state markov model for Rayleigh fading channels," *IEEE Trans. Commun.,* vol. 47, no. 11, pp. 1688–1692, Nov. 1999.

[64] M. Zorzi, R. R. Rao, and L. B. Milstein, "Error statistics in data transmission over fading channels," *IEEE Trans. Commun.,* vol. 46, no. 11, pp. 1468–1477, Nov. 1998.

[65] M. Zorzi, A. Chockalingam, and R. R. Rao, "Throughput analysis of TCP on channels with memory," *IEEE Journal on Selected Areas in Communications,* vol. 18, no. 7, pp. 1289–1300, July 2000.