# Power Control and Scheduling for Guaranteeing Quality of Service in Cellular Networks

Dapeng Wu* Rohit Negi†

**Abstract**

Providing Quality of Service(QoS) guarantees is important in the third generation (3G) and the fourth generation (4G) cellular networks. However, large scale fading and non-stationary small scale fading can cause severe QoS violations. To address this issue, we design QoS provisioning schemes, which are robust against time-varying large scale path loss, shadowing, non-stationary small scale fading, and very low mobility. In our design, we utilize our recently developed effective capacity technique and the time-diversity dependent power control proposed in this paper. The key elements of our QoS provisioning schemes are channel estimation, power control, dynamic channel allocation, and adaptive transmission. The advantages of our QoS provisioning schemes are 1) power efficiency, 2) simplicity in QoS provisioning, 3) robustness against large scale fading and non-stationary small scale fading. Simulation results demonstrate that the proposed algorithms are effective in providing QoS guarantees under various channel conditions.

**Key Words:** QoS, fading channel, effective capacity, power control, 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`.

†Carnegie Mellon University, Dept. of Electrical & Computer Engineering, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA. Tel. (412) 268-6264. Fax (412) 268-2860. Email: `negi@ece.cmu.edu`. URL: `http://www.ece.cmu.edu/~negi`.

# 1 Introduction

With the rapid growth of multimedia and data services in wireless networks, there is an increasing demand for Quality of Service(QoS) provisioning to support various applications. However, the task of explicit provisioning of QoS guarantees is not trivial since traditional methods typically incurs very high complexity [7, pp. 123–125]. To reduce the complexity in QoS provisioning, we [6] proposed a link-layer channel model based on the concept of *effective capacity* and developed a simple algorithm to estimate the parameters of the proposed channel model. Since effective capacity captures the effect of channel fading on the queueing behavior of the link, using a computationally simple yet accurate model, it is the critical device we need to design efficient QoS provisioning mechanisms. This has been shown in [9, 8], where we utilized the effective capacity channel model and developed simple and efficient schemes for admission control, resource allocation, and scheduling, which can yield substantial capacity gain.

However, QoS provisioning schemes (including our schemes in [9, 8]) typically suffer from time-varying large scale path loss, shadowing, non-stationary small scale fading, and very low mobility (i.e., very low degree of time diversity). In practice, the average power may change over time due to changes in the distance between the transmitter and the receiver and due to shadowing, while the Doppler rate may change over time due to changes in the velocity of the source/receiver (*i.e.*, non-stationary small scale fading). Such non-stationary behavior, which is possible in practical channels, can cause severe QoS violations. So, it is important to design QoS provisioning mechanisms that can mitigate large scale fading as well as non-stationary small scale fading.

In this paper, we design QoS provisioning schemes, which are robust against time-varying large scale path loss, shadowing, non-stationary small scale fading, and very low mobility. The key elements of our QoS provisioning schemes are channel estimation, power control, dynamic channel allocation, and adaptive transmission: to achieve the target QoS, we use the effective capacity channel model and propose a simple channel estimation algorithm; in power control, we utilize an efficient scheme called *time-diversity dependent power control* proposed in Section 3; we design a dynamic channel allocation mechanism that can adapt to changes in channel statistics, so as to achieve both efficiency and QoS guarantees; in adaptive transmission, we determine the transmission rate with the consideration of the effect of the physical layer (*i.e.*, practical modulation, channel coding, and signal-to-interference-plus-noise ratio (SINR) estimation error) on the link-layer performance. The nice features of our QoS provisioning schemes are 1) power efficiency, 2) simplicity in QoS provisioning, 3) robustness against large scale fading and non-stationary small scale fading. Our simulation results demonstrate that our proposed algorithms are effective in providing QoS guarantees under various channel conditions.

The remainder of this paper is organized as follows. In Section 2, we describe statistical QoS guarantees and our effective capacity channel model, which will be used in our QoS provisioning schemes. Section 3 discusses the trade-off between power control and time-diversity and proposes our time-diversity dependent power control. In Section 4, we present QoS provisioning schemes for downlink transmission. Section 5 describes our QoS provisioning schemes for uplink transmission. In Section 6, we present the simulation results that demonstrate the performance of our schemes. Section 7 concludes the paper.

## 2 Statistical QoS and Effective Capacity Channel Model

In wireless networking, statistical QoS guarantees are typically provisioned [10]. We formally 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 $\varepsilon$, that is,

$$Pr\{D(\infty) > D_{max}\} \le \varepsilon, \qquad (1)$$

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 $\{r_s, D_{max}, \varepsilon\}$. Even for this simple case, it is not immediately obvious as to which QoS triplets are feasible, for the given channel, since a rather complex queueing system (with an arbitrary channel capacity process) will need to be analyzed. The key contribution of [6] was to introduce a concept of statistical delay-constrained capacity termed *effective capacity*, which allows us to obtain a simple and efficient test, to check the feasibility of QoS triplets for a single time-varying channel. Next, we briefly explain the concept of effective capacity, and refer the reader to [6] for details.

Let $r(t)$ be the instantaneous channel capacity at time $t$. The *effective capacity function* of $r(t)$ is defined as [6]

$$\alpha(u) = -\lim_{t\to\infty} \frac{1}{ut} \log E[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}], \ \forall\ u > 0. \tag{2}$$

In this paper, since $t$ is a discrete frame index, the integral above should be thought of as a summation.

Consider a queue of infinite buffer size supplied by a data source of *constant* data rate $\mu$. It can be shown [6] 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

$$Pr\{D(\infty) > D_{max}\} \approx e^{-\theta(\mu) D_{max}}, \tag{3}$$

where the function $\theta(\mu)$ of source rate $\mu$ depends only on the channel capacity process $r(t)$. $\theta(\mu)$ can be considered as a "channel model" that models the channel at the link layer (in contrast to "physical layer" models specified by Markov processes, or Doppler spectra). The approximation (3) is accurate for large $D_{max}$.

In terms of the effective capacity function (2) defined earlier, the *QoS exponent function* $\theta(\mu)$ can be written as [6]

$$\theta(\mu) = \mu\alpha^{-1}(\mu) \tag{4}$$

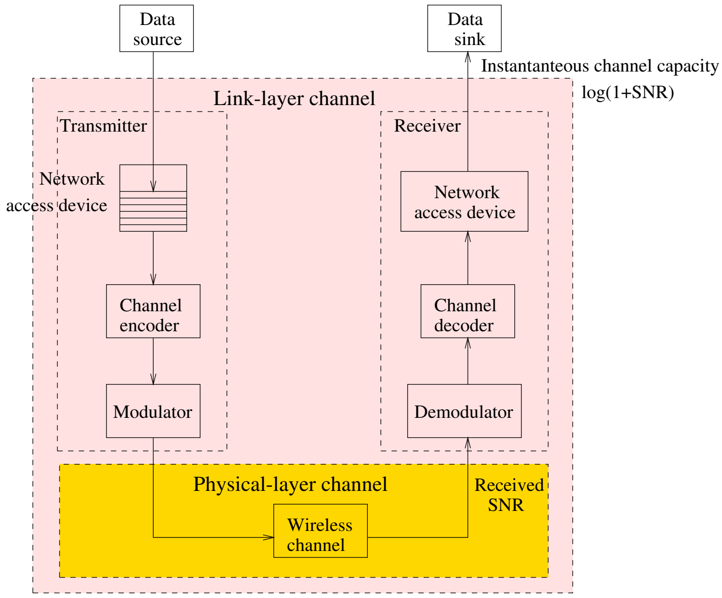



Figure 1: A packet-based wireless communication system.

where $\alpha^{-1}(\cdot)$ is the inverse function of $\alpha(u)$. Hence, we call $\theta(\mu)$ effective capacity channel model. 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 (4)) to relate the channel capacity process $r(t)$ to statistical QoS. Since our effective capacity method predicts an exponential dependence (3) between $\varepsilon$ and $D_{max}$, we can henceforth consider the QoS pair $\{r_s, \rho\}$ to be equivalent to the QoS triplet $\{r_s, D_{max}, \varepsilon\}$, with the understanding that $\rho = -\log \varepsilon / D_{max}$.

Next, we discuss the trade-off between power control and time diversity.

# 3 Trade-off between Power Control and Time Diversity

It is well known that ideal power control can completely eliminate fading and convert the fading channel to an AWGN channel, so that deterministic QoS (zero queueing delay and zero delay-bound violation probability) can be guaranteed. However, fast fading (or time diversity) is actually useful. From the link-layer[1] perspective, the higher the degree of time diversity, the larger the effective capacity $\alpha(u)$ for a fixed QoS parameter $u$. But for a slow fading channel, we know that the effective capacity $\alpha(u)$ can be very small due to the stringent delay requirement, and therefore power control may be needed to provide the required QoS. Hence, it is conceivable that there is a trade-off between power control and the utilization of time diversity, depending on the degree of time diversity and the QoS requirements.

To identify this trade-off, we compare the following three schemes through simulations:

- **Ideal power control:** In order to keep the received signal-to-interference-plus-noise ratio (SINR) constant at a target value $SINR_{target}$, the transmit power at frame $t$ is determined as below

$$P_0(t) = \frac{SINR_{target}}{\tilde{g}(t)}, \tag{5}$$

  where the channel power gain $\tilde{g}(t)$ (absorbing the noise variance plus interference) is given by

$$\tilde{g}(t) = \frac{g(t)}{\sigma_n^2 + P_I(t)} \tag{6}$$

  where $g(t)$ is the channel power gain at frame $t$, $\sigma_n^2$ is the noise variance and $P_I(t)$ is the instantaneous interference power. Denote $P_{avg}$ the time average of $P_0(t)$ specified by (5); the time average is over the entire simulation duration. Note that the fast power control used in 3G networks [4, pp. 188-195] is an approximation of ideal power control, in that the fast power control in 3G has a peak power constraint and in that the power change (in dB) in each interval can only be a fixed integer, say 1 dB, rather than an arbitrary real number as in (5).

[1]As shown in Fig. 1, a link layer consists of a buffer at the transmitter, channel encoder, modulator, wireless channel, demodulator, channel decoder, and network access device at the receiver.

- **Fixed power:** The transmit power $P_0(t)$ is kept constant and is equal to $P_{avg}$. The objective of this scheme is to use time diversity only.

- **Time-diversity dependent power control:** This is our proposed scheme. To utilize time diversity, the transmit power at frame $t$ is determined as below

$$P_0(t) = \frac{\gamma_{coeff}}{g_{avg}(t)}, \tag{7}$$

where $\gamma_{coeff}$ is so determined that the time average of $P_0(t)$ in (7) is equal to $P_{avg}$; and $g_{avg}(t)$ is given by an exponential smoothing of $\tilde{g}(t)$ as below

$$g_{avg}(t) = (1 - \eta_g) \times g_{avg}(t-1) + \eta_g \times \tilde{g}(t) \tag{8}$$

where $\eta_g \in [0, 1]$ is a fixed parameter, chosen depending on the time diversity desired. It is clear that if $\eta_g = 0$, the time-diversity dependent power control reduces to the fixed power scheme; if $\eta_g = 1$, the time-diversity dependent power control reduces to ideal power control. Hence, by optimally selecting $\eta_g \in [0, 1]$, we expect to trade off time diversity against power control.

The three schemes have been so specified that they use the same amount of average power $P_{avg}$, for fairness of comparison. In all of the three schemes, the transmission rate at frame $t$ is given as

$$r(t) \;\; = \;\; B_c \times \log_2 \left( 1 + \frac{P_0(t) \times \tilde{g}(t)}{\Gamma_{link}} \right), \tag{9}$$

with the assumption that $\tilde{g}(t)$ is perfectly known at the transmitter. In (9), $B_c$ is the bandwidth of the channel; we use $\Gamma_{link}$ to accommodate the difference between the actual data rate achievable in practical systems and the Shannon channel capacity, since the Shannon channel capacity is typically not achievable by practical modulation and channel coding schemes (refer to [7, pp. 177–179] for how to obtain $\Gamma_{link}$ in practical systems). We assume that transmission at the rate $r(t)$ results in negligible decoding error probability (as compared to $Pr\{D(\infty) \geq D_{max}\}$ or buffer overflow probability).

Denote $\hat{\mu}(D_{max}, \varepsilon)$ the maximum data rate $\mu$, with $Pr\{D(\infty) > D_{max}\} \leq \varepsilon$ satisfied. That is, $\hat{\mu}(D_{max}, \varepsilon)$ is the maximum data rate achievable with delay bound $D_{max}$ and the delay-bound violation probability not greater than $\varepsilon$. Denote $T_c$ the coherence time of a fading channel. Figure 3 shows data rate $\hat{\mu}(D_{max}, \varepsilon)$ vs. time diversity index $D_{max}/T_c$ for the three schemes. It is clear that the larger the index $D_{max}/T_c$ is, the higher degree of time diversity the channel possesses. From the figure, we have the following observations:

1. *Power control vs. using time diversity.* If the degree of time diversity is low, ideal power control provides a substantial capacity gain as opposed to the fixed power scheme, which only uses time diversity; otherwise, the schemes utilizing time diversity can provide a higher rate $\hat{\mu}(D_{max}, \varepsilon)$ than ideal power control. The reason is as follows. When the degree of time diversity is low, which implies that the probability of having long deep fades is high, then ideal power control can keep the error-free data rate $r(t)$ constant at a high value even during deep fades, while the fixed power scheme suffers from low data rate $r(t)$ during deep fades. On the other hand, when the degree of time diversity is high and hence the probability of having long deep fades is small, one can leverage time diversity by buffering data during deep fades (limited by the delay bound $D_{max}$) and transmitting at a high data rate when the channel conditions are good.

2. The rate $\hat{\mu}(D_{max}, \varepsilon)$ under both the fixed power control and the time-diversity dependent power control, increases with the degree of time diversity. The reason is as given above.

3. The time-diversity dependent power control, which jointly utilizes power control and time diversity, is the best among the three schemes. This is because the fixed power scheme and ideal power control are special cases of the time-diversity dependent power

control, when $\eta_g = 0$ and 1, respectively. Hence, by optimally selecting $\eta_g \in [0, 1]$, the time-diversity dependent power control can achieve the largest $\hat{\mu}(D_{max}, \varepsilon)$.

4. As the degree of time diversity increases, the capacity gain provided by the time-diversity dependent power control increases as compared to ideal power control; the capacity gain provided by the time-diversity dependent power control decreases as compared to the fixed power scheme. This is because, as the degree of time diversity increases, the effect of time diversity on $\hat{\mu}(D_{max}, \varepsilon)$ increases, while the effect of power control on $\hat{\mu}(D_{max}, \varepsilon)$ does not change.

With the effective capacity channel model and our time-diversity power control, we design QoS provisioning schemes for downlink transmission and uplink transmission, which are presented in the next two sections.

# 4 Downlink Transmission

We first describe the schemes for the case of downlink transmissions, *i.e.*, a base station (BS) transmits data to a mobile station (MS).

Assume that a connection requesting a QoS triplet $\{r_s, D_{max}, \varepsilon\}$ or equivalently $\{r_s, \rho = -\log \varepsilon / D_{max}\}$, is accepted by the admission control (described later in Algorithm 2). In the transmission phase, the following tasks are performed.

1. **SINR estimation at the MS:** The MS estimates instantaneous received SINR at frame $t$, denoted by $SINR(t)$, which is given by

$$SINR(t) = \frac{P_0(t) \times g(t)}{\sigma_n^2 + P_I(t)}. \tag{10}$$

   where $P_0(t)$ is the transmitted power at the BS in frame $t$, $g(t)$ is the channel power gain in frame $t$, $\sigma_n^2$ is the noise variance and $P_I(t)$ is the instantaneous interference power.

Then the MS conveys the value of $SINR(t)$ to the BS. Since the value of $SINR(t)$ is typically within a range of 30 dB, *e.g.*, from -19 to 10 dB, five bits should be enough to represent the value of $SINR(t)$ to within 1 dB. If the estimation frequency is 200 Hz, the signaling overhead is only 1 kb/s, which is low. Note that 3G allows for a 1500-Hz power control loop.

2. **Time-diversity dependent power control at the BS:** Since the BS knows the transmit power $P_0(t)$, upon receiving $SINR(t)$, it can derive the channel power gain $\tilde{g}(t)$ as below

$$\tilde{g}(t) = \frac{SINR(t)}{P_0(t)} = \frac{g(t)}{\sigma_n^2 + P_I(t)} \tag{11}$$

Denote $P_{peak}$ the peak transmit power at the BS. The BS determines the transmit power for frame $t+1$ by

$$P_0(t+1) = \min\left\{\frac{SINR_{target}}{g_{avg}(t)}, P_{peak}\right\}, \tag{12}$$

where $g_{avg}(t)$ is given by (8). Note that $\eta_g$ in (8) is time-diversity dependent; based on the current mobile speed $v_s(t)$, the value of $\eta_g$ is specified by a table, similar to Table 2.

Note that the downlink power control described here is different from the downlink power control in 3G systems, in that in our scheme, the BS initiates the power control while in 3G systems, the MS initiates the power control. Specifically, in our system, the BS determines the transmit power 'value' based on the value of $SINR(t)$ sent by the MS, while in 3G systems, the power control signal (*i.e.*, power-up or power-down signal) is sent from the MS to the BS. In 3G systems, the power-up signal requests an increase of transmit power by a preset value, *e.g.*, 1 dB, and the power-down signal requests a decrease of transmit power by a preset value, *e.g.*, 1 dB.

3. **Estimation of QoS exponent $\theta$ at the BS:** The BS measures the queueing delay $D(t)$ at the transmit buffer, and estimates the average queueing delay $D_{avg}(t)$ at frame $t$ by

$$D_{avg}(t) = (1 - \eta_d) \times D_{avg}(t-1) + \eta_d \times D(t) \tag{13}$$

where $\eta_d \in (0, 1)$ is a preset constant. Then, the BS estimates the QoS exponent $\theta$ at frame $t$, denoted by $\hat{\theta}(t)$, as below

$$\hat{\theta}(t) = \frac{1}{0.5 + D_{avg}(t)} \tag{14}$$

Eq. (14) is obtained from Eq. (22) in [6].

4. **Scheduling (dynamic channel allocation) at the BS:** Denote $\hat{\mu}(D_{max}, \varepsilon)$ the maximum data rate $\mu$, with $Pr\{D(\infty) > D_{max}\} \leq \varepsilon$ satisfied. That is, $\hat{\mu}(D_{max}, \varepsilon)$ is the maximum data rate achievable with delay bound $D_{max}$ and the delay-bound violation probability not greater than $\varepsilon$. It is known [7] that as the degree of time diversity, or equivalently the mobile speed, increases (resp., decreases), the data rate $\hat{\mu}(D_{max}, \varepsilon)$ increases (resp., decreases) and the QoS exponent $\theta(\mu = r_s)$ increases (resp., decreases), hence requiring less (resp., more) channel resource to support the requested QoS. This motivates us to design a dynamic channel allocation mechanism that can adapt to changes in channel statistics, so as to achieve both efficiency and QoS guarantees.

   The basic idea of dynamic channel allocation is to use the QoS measures $\hat{\theta}(t)$ and $D(t)$ in deciding channel allocation. Specifically, the BS allocates a fraction $\lambda(t+1)$ of frame $t+1$, to the connection, as below

$$\lambda(t+1) = \begin{cases} \min\{\lambda(t) + \Delta_\lambda, 1\} & \text{if } \hat{\theta}(t) < \gamma_{inc} \times \rho \text{ and } D(t) > D_h; \\ \max\{\lambda(t) - \Delta_\lambda, 0\} & \text{if } \hat{\theta}(t) > \gamma_{dec} \times \rho \text{ and } D(t) < D_l; \\ \lambda(t) & \text{otherwise.} \end{cases} \tag{15}$$

   where $\Delta_\lambda \in (0, 1)$, $\gamma_{inc} \geq 1$, $\gamma_{dec} \geq \gamma_{inc}$, low threshold $D_l \in (0, D_{max})$, and high threshold $D_h \in (D_l, D_{max})$ are preset constants.

It is clear that the control in (15) has hysteresis (due to $D_h > D_l$ and $\gamma_{dec} \geq \gamma_{inc}$), which helps reduce the variation in $\lambda(t)$ and hence reduce the signaling overhead for dynamic channel allocation. The condition $\hat{\theta} < \gamma_{inc} \times \rho$ means that the measured QoS exponent $\hat{\theta}$ does not meet the required $\rho$, scaled by $\gamma_{inc} \geq 1$ to allow a safety margin; the condition $D(t) > D_h$ means that the delay $D(t)$ is larger than the high threshold $D_h$; the two conditions jointly trigger an increase in $\lambda(t)$. Similarly, the condition $\{\hat{\theta} > \gamma_{dec} \times \rho$ and $D(t) < D_l\}$ causes a decrease in $\lambda(t)$.

In practice, $\lambda(t)$ can be interpreted in different ways, depending on the type of the system. For CDMA, TDMA, and FDMA systems, $\lambda(t)$ can be implemented by using variable spreading codes, variable number of mini-slots, and variable number of frequency carriers, respectively.

For ease of implementation, one can set $\Delta_\lambda = 0.1$ so that $\lambda(t)$ only takes discrete values from the set $\{0, 0.1, 0.2, \cdots, 0.9, 1\}$. Then, in a TDMA system, if a frame consists of ten mini-slots, $\lambda(t) = 0.3$ would mean using three mini-slots to transmit the data at frame $t$; the remaining seven mini-slots in the frame can be used by other users, *e.g.*, best-effort users.

5. **Adaptive transmission at the BS:** Once the channel allocation $\lambda(t+1)$ is given, the BS determines the transmission rate at frame $t+1$ as below

$$r(t+1) = \lambda(t+1) \times \varpi^* \times B_c \times \log_2\left(1 + \frac{P_0(t+1) \times g(t)}{(\sigma_n^2 + P_I(t)) \times \Gamma_{link} \times \gamma_{safe}}\right) \quad (16)$$

$$= \lambda(t+1) \times \varpi^* \times B_c \times \log_2\left(1 + \frac{P_0(t+1) \times SINR(t)}{P_0(t) \times \Gamma_{link} \times \gamma_{safe}}\right) \quad (17)$$

where $B_c$ is the channel bandwidth, $\varpi^*$ denotes the amount of channel resource allocated by the admission control (described later in Algorithm 2), $\Gamma_{link}$ characterizes the effect of practical modulation and coding and $\gamma_{safe}$ introduces a safety margin to mitigate the effect of the SINR estimation error at the MS. The BS uses (17) to compute $r(t+1)$ since all variables in (17) are known.

The values of $\Gamma_{link}$ and $\gamma_{safe}$ are so chosen that transmitting at the rate $r(t+1)$ specified by (17) will result in negligible bit error rate (w.r.t. $P_{loss}$, which is the packet loss probability due to buffer overflow at the transmitter). So, $r(t+1)$ specified by (17) can be regarded as an error-free data rate. Since (17) takes into account the effect of the physical layer (*i.e.*, practical modulation, channel coding, and SINR estimation error), we can focus on the queueing behavior and link-layer performance.

Once $r(t+1)$ is determined, an M-ary QAM can be used for the transmission, where $M = 2^b$ and $b$ is given by

$$b = floor\left(\log_2\left(\lambda(t+1) \times \varpi^* \times \log_2\left(1 + \frac{P_0(t+1) \times SINR(t)}{P_0(t) \times \Gamma_{link} \times \gamma_{safe}}\right)\right)\right) \tag{18}$$

where $floor(x)$ is the largest integer that is not larger than $x$.

The above tasks are summarized in Algorithm 1.

**Algorithm 1 Downlink power control, channel allocation, and adaptive transmission**

*In the transmission phase, the following tasks are performed.*

1. **SINR estimation at the MS:** *The MS estimates the received $SINR(t)$ and conveys the value of $SINR(t)$ to the BS.*

2. **Power control at the BS:** *The BS derives the channel power gain $\tilde{g}(t)$ using (11), estimates $g_{avg}(t)$ using (8), and then determines the transmit power $P_0(t+1)$ using (12).*

3. **Estimation of QoS exponent $\theta$ at the BS:** *The BS measures the queueing delay $D(t)$, estimates $D_{avg}(t)$ using (13), and estimates the QoS exponent $\hat{\theta}(t)$ using (14).*

4. **Scheduling at the BS:** *The BS allocates a fraction of frame* $\lambda(t+1)$ *to the connection, using (15).*

5. **Adaptive transmission at the BS:** *The BS determines the transmission rate* $r(t+1)$ *using (17).*

The key elements in Algorithm 1 are power control and scheduling. The power control is intended to mitigate large scale path loss, shadowing, and low mobility. The scheduler specified by (15) is targeted at achieving both efficiency and QoS guarantees.

In Algorithm 1, the power control allocates the power resource, while the scheduler allocates the channel resource; their effects on the 'error-free' transmission rate $r(t)$ in (17) are different: $r(t)$ is linear in channel allocation $\lambda(t)$, but is a log-function of power $P_0(t)$.

**Remark 1 Power control vs. channel allocation in QoS provisioning**

*From (17), we see that the error-free data rate* $r(t)$ *is determined by the channel resource allocated* $\lambda(t)$ *and power* $P_0(t)$. *A natural question is how to optimally allocate the channel and power resource to satisfy the required QoS.*

*There are two extreme cases. First, if the transmit power* $P_0(t)$ *is fixed and we suppose* $\lambda(t) \in [0, \infty)$, *then given arbitrary channel gain* $g(t)$ *(which includes the effect of the noise and interference), we can obtain arbitrary* $r(t) \in [0, \infty)$ *by choosing appropriate* $\lambda(t) \in [0, \infty)$. *Second, if the channel resource allocated* $\lambda(t)$ *is fixed and we suppose* $P_0(t) \in [0, \infty)$, *then given arbitrary channel gain* $g(t)$, *we can obtain arbitrary* $r(t) \in [0, \infty)$ *by choosing appropriate* $P_0(t) \in [0, \infty)$.

*However, in practical situations, we have both a peak power constraint* $P_0(t) \leq P_{peak}$ *and a peak channel usage constraint* $\lambda(t) \leq 1$, *assuming that* $\lambda(t)$ *is the fraction of allotted channel resource. Hence, we cannot obtain arbitrary* $r(t) \in [0, \infty)$, *given arbitrary channel gain* $g(t)$. *Since applications can tolerate a certain delay and there is a buffer at the link layer,* $r(t)$ *is allowed to be less than the arrival rate, with a small probability. Therefore, there could* *be feasible solutions* $\{P_0(t), \lambda(t)\}$ *that satisfy the QoS constraint, peak power constraint, and peak channel usage constraint. If such feasible solutions do exist, the next question is which one is the optimal solution, given a certain criterion. If we want to minimize average power usage (resp., average channel usage) under the QoS constraint, peak power constraint, and peak channel usage constraint, an optimal solution must have* $\lambda(t) = 1$ *(resp.,* $P_0(t) = P_{peak}$*). Hence, we cannot simultaneously minimize both average power or average channel usage; and we are facing a multi-objective optimization problem. A classical multi-objective optimization method is to convert a multi-objective optimization problem to a single-objective optimization problem by a weighted sum of multiple objectives, the solution of which is Pareto optimal [2, page 49]. Using this method, we formulate an optimization problem as follows*

$$\underset{\{P_0(t),\lambda(t):t=0,1,\cdots,\tau-1\}}{\textit{maximize}} \quad \frac{1}{\tau}\sum_{t=0}^{\tau-1}\mathbf{E}[\beta_{weight} \times P_0(t) + (1-\beta_{weight}) \times \lambda(t)] \tag{19}$$

$$\textit{subject to} \quad Pr\{D(\infty) \geq D_{max}\} \leq \varepsilon, \qquad \textit{for a fixed rate } r_s \tag{20}$$

$$0 \leq P_0(t) \leq P_{peak} \tag{21}$$

$$0 \leq \lambda(t) \leq 1 \tag{22}$$

*where* $\tau$ *is the connection life time, and* $\beta_{weight} \in [0, 1]$*. Dynamic programming often turns out to be a natural way to solve (19). However, the complexity of solving the dynamic program is high. If the statistics of the channel gain process are unpredictable (due to large scale fading and time-varying mobile speed), we cannot use dynamic programming to solve (19). This motivates us to seek a simple (sub-optimal) approach, which can enforce the specified QoS constraints explicitly, and yet achieve an efficient channel and power usage. Our scheme is based on the tradeoff between power and time diversity: we use the time-diversity dependent power control to maximize the data rate* $\hat{\mu}(D_{max}, \varepsilon)$*, and use scheduling to determine the minimum amount of resource that satisfies the required QoS, given the choice of power control. This leads to two separate optimization problems, which simplifies the complexity, while achieving good performance. Algorithm 1 is designed according to this* *idea.*

Now, we get to the issue of admission control. Assume that a user initiates a connection request, requiring a QoS triplet $\{r_s, D_{max}, \varepsilon\}$. In the connection setup phase, we use Algorithm 2 (see below) to test whether the required QoS can be satisfied. Specifically, the algorithm measures the QoS that the link-layer channel can provide; if the measured QoS satisfies the required QoS, the connection request is accepted; otherwise, it is rejected.

Algorithm 2 uses the methods in Algorithm 1. The key difference between the two algorithms is that in Algorithm 2, the BS creates a fictitious queue, that is, the BS uses $r_s$ as the arrival rate and $r(t)$ as the service rate to 'simulate' a fictitious queue, but no actual packet is transmitted over the wireless channel. In the admission test, there is no need for the BS to transmit actual data in order to obtain QoS measures of the link-layer channel. This is because 1) the MS can use the common pilot channel [4, page 103] to measure the received $SINR(t)$, and 2) the simulated fictitious queue provides the same queueing behavior as if actual data was transmitted over the wireless channel.

To facilitate resource allocation, we simulate $N_{fic}$ fictitious queues, each of which is allocated with different amount of resource $\varpi_i$ $(i = 1, \cdots, N_{fic})$. Assume that $\varpi_i$ represents the proportion of the resource allocated to queue $i$, to the total resource, and $\varpi_i$ $(i = 1, \cdots, N_{fic})$ takes a discrete value in $(0, 1]$, e.g., $\varpi_i \in \{0, 0.1, 0.2, \cdots, 0.9, 1\}$. If the connection is accepted, the BS allocates the minimum amount of resource (denoted by $\varpi^*$) that satisfies the QoS requirements, to the connection. That is, $\varpi^*$ is the minimum of all feasible $\varpi_i$ that satisfy the QoS requirements. The algorithm for admission control and resource allocation is as below.

**Algorithm 2 Downlink admission control and resource allocation:**
*Upon the receipt of a connection request requiring a QoS triplet* $\{r_s, D_{max}, \varepsilon\}$*, the following tasks are performed.*

1. **SINR estimation at the MS:** *The MS estimates the received $SINR(t)$ using the common pilot channel, and conveys the value of $SINR(t)$ to the BS.*

2. **Power control at the BS:** *The BS derives the channel power gain $\tilde{g}(t)$ using (11), where $P_0(t)$ is meant to be the actual transmit power for the common pilot channel at frame $t$. Then, the BS estimates $g_{avg}(t)$ by computing (8). Finally, the BS determines the fictitious transmit power $\tilde{P}_0(t+1)$ using (12).*

3. **Estimation of QoS exponent $\theta$ at the BS:** *For each fictitious queue $i$ ($i = 1, \cdots, N_{fic}$), the BS generates fictitious arrivals with data rate $r_s$, measures the queueing delay $D_i(t)$, estimates $D_{avg}^{(i)}(t)$ using (13), and estimates the QoS exponent $\hat{\theta}_i(t)$ using (14).*

4. **Scheduling at the BS:** *For each fictitious queue $i$ ($i = 1, \cdots, N_{fic}$), the BS allocates a fraction of frame $\lambda_i(t+1)$, using (15).*

5. **Adaptive transmission at the BS:** *For each fictitious queue $i$ ($i = 1, \cdots, N_{fic}$), the BS determines the transmission rate $r_i(t+1)$ as below*

$$r_i(t+1) = \lambda_i(t+1) \times \varpi_i \times B_c \times \log_2 \left( 1 + \frac{\tilde{P}_0(t+1) \times SINR(t)}{\tilde{P}_0(t) \times \Gamma_{link} \times \gamma_{safe}} \right) \quad (23)$$

6. **Admission control and resource allocation:** *If there exists a queue $\tilde{i}$ such that its QoS exponent average $\hat{\theta}_{avg}^{(\tilde{i})}(t) = \frac{1}{t+1} \sum_{\tau=0}^{t} \hat{\theta}_{\tilde{i}}(\tau)$ is not less than a preset threshold $\theta_{th}$, accept the connection request; otherwise, reject it. If the connection is accepted, the BS allocates the minimum amount of resource $\varpi^* = \min_{\tilde{i}} \varpi_{\tilde{i}}$, to the connection.*

Note that in Algorithm 2, the MS needs to convey $SINR(t)$ to the BS in the connection setup phase, which is different from the current 3G standard.

It is required that Algorithm 2 be fast and accurate in order to implement it in practice. Our simulation results in Section 6.2.7 show that $\hat{\theta}_{avg}(t)$ is a reliable QoS measure for the purpose of admission control; moreover, within a short period of time, say two seconds, the system can obtain a reasonably accurate $\hat{\theta}_{avg}(t)$ and hence can make a quick and accurate admission decision.

As long as the large scale path loss and shadowing can be mitigated by the power control in (12), the required QoS can be guaranteed. It is known that the large scale path loss within a coverage area can be mitigated by the power control. To mitigate shadowing more effectively as compared to power control, our scheme can be improved by macro-diversity, which employs the collaboration of multiple base stations. We leave this for future study.

# 5 Uplink Transmission

For uplink transmissions, *i.e.*, an MS transmits data to a BS, the design methodology for QoS provisioning is the same as that for downlink transmissions. Specifically, we use Algorithms 3 and 4, which are modifications of Algorithms 2 and 1. Algorithms 3 uses the common random access channel [4, page 106] instead of the common pilot channel as in Algorithm 2.

**Algorithm 3 Uplink admission control and resource allocation:**
*Upon the receipt of a connection request requiring a QoS triplet* $\{r_s, D_{max}, \varepsilon\}$*, the following tasks are performed.*

1. **SINR estimation at the BS:** *The MS transmits a signal of constant power* $P_{MS}$ *over the common random access channel to the BS. The value of* $P_{MS}$ *is known to the BS. The BS estimates received* $SINR(t)$ *from the common random access channel.*

2. **Power control at the BS:** *The BS derives the channel power gain* $\tilde{g}(t)$ *by (11), where* $P_0(t)$ *is equal to* $P_{MS}$*. Then the BS estimates* $g_{avg}(t)$ *by computing (8). Finally,*

*the BS determines the fictitious transmit power $\tilde{P}_0(t+1)$ by (12), where $P_{peak}$ is with respect to the MS and is specified in the 3G standard.*

3. **Estimation of QoS exponent $\theta$ at the BS:** *For each fictitious queue $i$ ($i = 1, \cdots, N_{fic}$), the BS generates fictitious arrivals with data rate $r_s$, measures the queueing delay $D_i(t)$, estimates $D^{(i)}_{avg}(t)$ using (13), and estimates the QoS exponent $\hat{\theta}_i(t)$ using (14).*

4. **Scheduling at the BS:** *For each fictitious queue $i$ ($i = 1, \cdots, N_{fic}$), the BS allocates a fraction of frame $\lambda_i(t+1)$, using (15).*

5. **Adaptive transmission at the BS:** *For each fictitious queue $i$ ($i = 1, \cdots, N_{fic}$), the BS determines the transmission rate $r_i(t+1)$ using (23).*

6. **Admission control and resource allocation:** *If there exists a queue $\tilde{i}$ such that its QoS exponent average $\hat{\theta}^{(\tilde{i})}_{avg}(t) = \frac{1}{t+1}\sum_{\tau=0}^{t}\hat{\theta}_{\tilde{i}}(\tau)$ is not less than a preset threshold $\theta_{th}$, accept the connection request; otherwise, reject it. If the connection is accepted, the BS allocates the minimum amount of resource $\varpi^* = \min_{\tilde{i}} \varpi_{\tilde{i}}$, to the connection.*

**Algorithm 4 Uplink power control, channel allocation, and adaptive transmission**

*In the transmission phase, the following tasks are performed.*

1. **SINR estimation at the BS:** *The BS estimates received $SINR(t)$ and conveys the value of $SINR(t)$ to the MS.*

2. **Power control at the MS:** *The MS derives the channel power gain $\tilde{g}(t)$ by (11) and estimates $g_{avg}(t)$ by computing (8). Then, the MS determines the transmit power $P_0(t+1)$ by (12).*

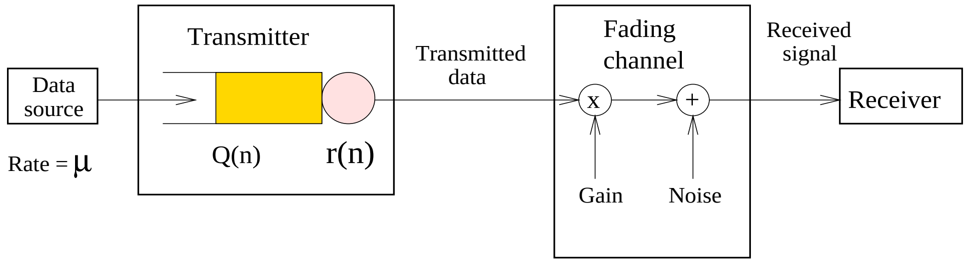



Figure 2: The queueing model used for simulations.

*3.* **Estimation of QoS exponent $\theta$ at the MS:** *The MS measures the queueing delay $D(t)$, estimates $D_{avg}(t)$ by (13), and estimates the QoS exponent $\hat{\theta}(t)$ using (14).*

*4.* **Renegotiation of channel allocation:** *The MS computes $\lambda(t+1)$, using (15). The MS sends a renegotiation request to the BS, asking for a fraction of frame $\lambda(t+1)$ for the connection. Based on the resource availability, the BS determines the value of $\lambda(t+1)$, and then notifies the MS of the final value of $\lambda(t+1)$, which will be used by the MS in frame $t+1$.*

*5.* **Adaptive transmission at the MS:** *The MS determines the transmission rate $r(t+1)$ by (17).*

# 6 Simulation Results

In this section, we simulate the discrete-time wireless communication system as depicted in Figure 2, and demonstrate the performance of our algorithms. We focus on Algorithm 1 for downlink transmission of a single connection, since the performance of Algorithm 4 for uplink transmission would be the same as that for Algorithm 1 if the simulation parameters are the same and fast feedback of channel gains is assumed. Section 6.1 describes the simulation setting, while Section 6.2 illustrates the performance of our algorithms.

## 6.1 Simulation Setting

### 6.1.1 Mobility Pattern Generation

We simulate the speed behavior of the MS using the model described in Ref. [1]. Under the model, an MS moves away from the BS, at a constant speed $v_s$ for a random duration; then a new target speed $v^*$ is randomly generated; the MS linearly accelerates or decelerates until this new speed $v^*$ is reached; following which, the MS moves at the constant speed $v^*$, and the procedure repeats again.

The speed behavior of an MS at frame $t$ can be described by three parameters:

- its current speed $v_s(t) \in [0, v_{max}]$ in units of m/s
- its current acceleration $a_s(t) \in [a_{min}, a_{max}]$ in m/s$^2$
- its current target speed $v^*(t) \in [0, v_{max}]$

where $v_{max}$ denotes the maximum speed, $a_{min}$ the minimum acceleration (which is negative), and $a_{max}$ the maximum acceleration.

At the beginning of the simulation, the MS is assigned an initial speed $v_s(0)$, which is generated by a probability density function $f_v(v_s)$, given by

$$f_v(v_s) = \begin{cases} p_0 \times \delta(v_s) & \text{if } v_s = 0; \\ p_{max} \times \delta(v_s - v_{max}) & \text{if } v_s = v_{max}; \\ \frac{1-p_0-p_{max}}{v_{max}} & \text{if } 0 < v_s < v_{max}; \\ 0 & \text{otherwise.} \end{cases} \tag{24}$$

where $p_0 + p_{max} < 1$. That is, the random speed has high probabilities at speed 0 (imitating stops due to red lights or traffic jams) and at the maximum speed $v_{max}$ (a preferred speed when driving); and it is uniformly distributed between 0 and $v_{max}$.

The speed change events are modeled as a Poisson process. That is, the time between two consecutive speed changes is exponentially distributed with mean $m_{v^*}$. Note that a speed change event happens at an epoch determined by the Poisson process but it does not include the speed changes during acceleration/deceleration periods.

Now, we know the epochs of speed change events follow a Poisson process and the new target speed $v^*$ follows the PDF $f_v(v_s)$. Denote $t^*$ the time at which a speed change event occurs and $v^* = v^*(t^*)$ the associated new target speed. Then, an acceleration $a_s(t^*) \neq 0$ is generated by the PDF

$$f_a(a_s) = \begin{cases} \frac{1}{a_{max}} & \text{if } 0 < a_s \leq a_{max}; \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

if $v^*(t^*) > v_s(t^*)$, or by the PDF

$$f_a(a_s) = \begin{cases} \frac{1}{|a_{min}|} & \text{if } a_{min} \leq a_s < 0; \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

if $v^*(t^*) < v_s(t^*)$. Obviously, $a_s$ is set to 0 if $v^*(t^*) = v_s(t^*)$. If $a_s(t) \neq 0$, the speed continuously increases or decreases; at frame $t$, a new speed $v_s(t)$ is computed according to

$$v_s(t) = v_s(t-1) + a_s(t) \times T_s \tag{27}$$

until $v_s(t)$ reaches $v^*(t)$; $T_s$ is the frame length in units of second. Then, we set $a_s = 0$ and the MS moves at constant speed $v_s(t) = v^*(t^*)$ until the next speed change event occurs. Figure 5 shows a trace of the speed behavior of an MS.

### 6.1.2 Channel Gain Process Generation

The channel power gain process $g(t)$ is given by

$$g(t) = g_{small}(t) \times g_{large}(t) \times g_{shadow}(t) \tag{28}$$

where $g_{small}(t)$, $g_{large}(t)$, and $g_{shadow}(t)$ denote channel power gains due to small-scale fading, large scale path loss, and shadowing, respectively.

**Non-stationary small scale fading**

Given the mobile speed $v_s(t)$, the Doppler rate $f_m(t)$ can be calculated by [5, page 141]

$$f_m(t) = v_s(t) \times \cos\varphi \times f_c/c, \tag{29}$$

where $\varphi$ is the angle between the direction of motion of the MS and the direction of arrival of the electromagnetic waves, $f_c$ is the carrier frequency and $c$ is the speed of light, which is $3 \times 10^8$ m/sec. We choose $\varphi = 0$ in all the simulations.

We assume Rayleigh flat-fading for the small scale fading. Rayleigh flat-fading voltage-gains $h(t)$ are generated by an AR(1) model as below. We first generate $\bar{h}(t)$ by

$$\bar{h}(t) = \kappa(t) \times \bar{h}(t-1) + u_g(t), \tag{30}$$

where $u_g(t)$ are i.i.d. complex Gaussian variables with zero mean and unity variance per dimension. Then, we normalize $\bar{h}(t)$ and obtain $h(t)$ by

$$h(t) = \bar{h}(t) \times \sqrt{\frac{1-[\kappa(t)]^2}{2}}. \tag{31}$$

$\kappa(t)$ is determined by 1) computing the Doppler rate $f_m(t)$ for given mobile speed $v_s(t)$, using (29), 2) computing the coherence time $T_c$, through $T_c = \frac{9}{16\pi f_m}$, and 3) calculating $\kappa = 0.5^{T_s/T_c}$. Then we obtain $g_{small}(t) = |h(t)|^2$.

**Large scale path loss**

Next, we describe the generation of large scale path loss. Denote $\{x_t, y_t, z_t\}$ and $\{x_r, y_r, z_r\}$ the 3-dimensional locations of the transmit antenna and the receive antenna, respectively. Specifically, $z_t$ and $z_r$ are the heights of the transmit antenna and the receive antenna, respectively. The initial distance $d_0$ between the MS and BS is given by

$$d_0 = \sqrt{(x_t - x_r)^2 + (y_t - y_r)^2}. \tag{32}$$

Denote $d_{tr}(t)$ the distance between the BS (transmitter) and the MS (receiver) at $t$. Hence, we have $d_{tr}(0) = d_0$. Assume that the MS moves directly away from the BS. Then, for $t > 0$, we have

$$d_{tr}(t) = d_{tr}(t-1) + v_s(t) \times T_s. \tag{33}$$

We use two path loss models: Friis free space model and the ground reflection model. Friis free space model is given by [5, page 70]

$$g_{large}(t) = \left( \frac{c}{f_c \times 4 \times \pi \times d_{tr}(t)} \right)^2, \tag{34}$$

where $c$ is light speed, and $f_c$ is carrier frequency. The ground reflection (two-ray) model [5, page 89] is given as below

$$g_{large}(t) = \frac{z_t^2 \times z_r^2}{[d_{tr}(t)]^4}. \tag{35}$$

We need to compute the cross-over distance $d_{cross}$ to determine which model to use. $d_{cross}$ is given by [5, page 89]

$$d_{cross} = \frac{20 \times \pi \times z_t \times z_r \times f_c}{3 \times c}. \tag{36}$$

If $d_{tr}(t) \leq d_{cross}$, we choose Friis free space model (34) to generate $g_{large}(t)$; otherwise, we use the ground reflection model (35) to generate $g_{large}(t)$.

**Shadowing**

We generate the shadow fading process $\tilde{g}_{shadow}(t)$ in units of dB by an AR(1) model as below [3]

$$\tilde{g}_{shadow}(t) = \kappa_{shadow}^{v_s(t) \times T_s / D_{shadow}} \times \tilde{g}_{shadow}(t-1) + \sigma_{shadow} \times \tilde{u}_g(t) \tag{37}$$

where $\kappa_{shadow}$ is the correlation between two locations separated by a fixed distance $D_{shadow}$, $\tilde{u}_g(t)$ are i.i.d. Gaussian variables with zero mean and unity variance, $\sigma_{shadow}$ is a constant in units of dB, $v_s(t)$ is obtained from the above mobility pattern generation, and hence $v_s(t) \times T_s$ is the distance that the MS traverses in frame $t$. It is obvious that the shadowing gain $g_{shadow}(t) = 10^{\tilde{g}_{shadow}(t)/10}$ follows a log-normal distribution with standard deviation $\sigma_{shadow}$.

### 6.1.3 Simulation Parameters

Table 1 lists the parameters used in our simulations. Since we target at interactive real-time applications, we set the QoS triplet as below: $r_s = 50$ kb/s, $D_{max} = 50$ msec, and $\varepsilon = 10^{-3}$. In addition, we set the values of $\gamma_{inc}$, $\gamma_{dec}$, $D_l$, and $D_h$ in (15) in such a way that can reduce the signaling overhead for dynamic channel allocation, while meeting the QoS requirements $\{r_s, \rho = -\log_e \varepsilon / D_{max}\}$. Further, we set the values of $\sigma_{shadow}$, $\kappa_{shadow}$, and $D_{shadow}$ according to Ref. [3]. The maximum speed $v_{max} = 15.6$ m/s corresponds to 35 miles per hour. We set $P_{peak} = 24$ dBm according to the specification of 3G systems on mobile stations [4, page 159], so that our results are also applicable to uplink transmissions. We assume total intra-cell and inter-cell interference $P_I(t)$ is constant over time.

Assume that the random errors in estimating $SINR(t)$ are i.i.d. Gaussian variables with zero mean and variance $\sigma_{est}^2$. Denote the random estimation error in dB by $\tilde{g}_{est}(t)$. Then, the estimated $SINR(t)$ is given by

$$SINR(t) = \frac{P_0(t) \times g(t) \times 10^{\tilde{g}_{est}(t)/10}}{\sigma_n^2 + P_I(t)} \tag{38}$$

To be realistic, the power $P_0(t+1)$ specified in (12) only takes integer values in units of dB and can only change 1 dB in each frame. We also assume that M-ary QAM is used for modulation. In addition, each simulation run is 100-second long.

## 6.2 Performance Evaluation

We organize this section as follows. Sections 6.2.1 identifies the trade-off between power control and time diversity. In Section 6.2.2, we show the accuracy of the exponentially

Table 1: Simulation parameters.

| | | |
|---|---|---|
| QoS requirement | Constant bit rate $r_s$ | 50 kb/s |
| | Delay bound $D_{max}$ | 50 msec |
| | Delay-bound violation probability $\varepsilon$ | $10^{-3}$ |
| Channel | Bandwidth $B_c$ | 300 kHz |
| | Sampling-interval (frame length) $T_s$ | 1 msec |
| | Noise plus interference power $\sigma_n^2 + P_I(t)$ | -100 dBm |
| Mobility pattern | Maximum speed $v_{max}$ | 15.6 m/s |
| | Minimum acceleration $a_{min}$ | -4 m/s$^2$ |
| | Maximum acceleration $a_{max}$ | 2.5 m/s$^2$ |
| | Probability $p_0$ | 0.3 |
| | Probability $p_{max}$ | 0.3 |
| | Mean time between speed change $m_{v^*}$ | 25 sec |
| Shadowing | Standard deviation $\sigma_{shadow}$ | 7.5 dB |
| | Correlation $\kappa_{shadow}$ | 0.82 |
| | Distance $D_{shadow}$ | 100 m |
| Receive antenna | $x_r$ | 100 m |
| | $y_r$ | 100 m |
| | Height $z_r$ | 1.5 m |
| Transmit antenna | $x_t$ | 0 |
| | $y_t$ | 0 |
| | Height $z_t$ | 50 m |
| BS | Carrier frequency $f_c$ | 1.9 GHz |
| | Channel coding gain $\gamma_{code}$ | 3 dB |
| | Target bit error rate $\epsilon_{error}$ | $10^{-6}$ |
| | Smoothing weight for average delay $\eta_d$ | 0.0005 |
| | Standard deviation of estimation error $\sigma_{est}$ | 1 dB |
| | Safety margin $\gamma_{safe}$ | 1 dB |
| Power control | Peak transmission power $P_{peak}$ | 24 dBm |
| | $SINR_{target}$ | 5 dB |
| Scheduling | Step size $\Delta_\lambda$ | 0.1 |
| | $\gamma_{inc}$ | 1 |
| | $\gamma_{dec}$ | 1 |
| | Low threshold $D_l$ | $0.1 \times D_{max}$ |
| | High threshold $D_h$ | $0.5 \times D_{max}$ |

Table 2: Simulation parameters.

| Speed $v_s$ (m/s) | 0.011 | 0.11 | 0.23 | 0.57 | 1.1 | 5.7 | 11 | 56 |
|---|---|---|---|---|---|---|---|---|
| Speed $v_s$ (km/h) | 0.041 | 0.41 | 0.81 | 2 | 4.1 | 20 | 41 | 204 |
| Doppler rate $f_m$ (Hz) | 0.072 | 0.72 | 1.4 | 3.6 | 7.2 | 36 | 72 | 358 |
| Coherence time $T_c$ (s) | 2.5 | 0.25 | 0.125 | 0.05 | 0.025 | 0.005 | 0.0025 | 0.0005 |
| $D_{max}/T_c$ | 0.02 | 0.2 | 0.4 | 1 | 2 | 10 | 20 | 100 |
| $\eta_g$ | 0.2 | 0.08 | 0.08 | 0.04 | 0.04 | 0.02 | 0.02 | 0 |

smoothed estimate of $\theta$. Sections 6.2.3 to 6.2.6 evaluates the performance of Algorithm 1 under four cases, namely, a time-varying mobile speed, large scale path loss, shadowing, and very low mobility. In Section 6.2.7, we investigate whether our admission control test in Algorithms 2 and 3 is quick and accurate.

### 6.2.1 Power Control vs. Time Diversity

This experiment is to identify the trade-off between power control and time diversity.

We compare the three schemes, namely, ideal power control, the fixed power scheme, and the time-diversity dependent power control, defined in Section 3.

All the three schemes use the same amount of average power $P_{avg}$, for the purpose of fair comparison. Assuming that the channel power gain $\tilde{g}(t)$ is perfectly known by the transmitter, all the three schemes determine the transmission rate at frame $t$ using (9).

In each simulation, we generate Rayleigh fading with fixed mobile speed $v_s$ specified in Table 2. We do not simulate large scale path loss and shadowing. For the time-diversity dependent power control, the smoothing factor $\eta_g$ in (8) is given by Table 2. Note that for different mobile speeds $v_s$ in Table 2, we use different $\eta_g$; the value of $\eta_g$ is chosen so as to maximize the data rate $\hat{\mu}(D_{max}, \varepsilon)$.

Table 2 lists the parameters used in our simulations, where $D_{max} = 50$ ms. The range of

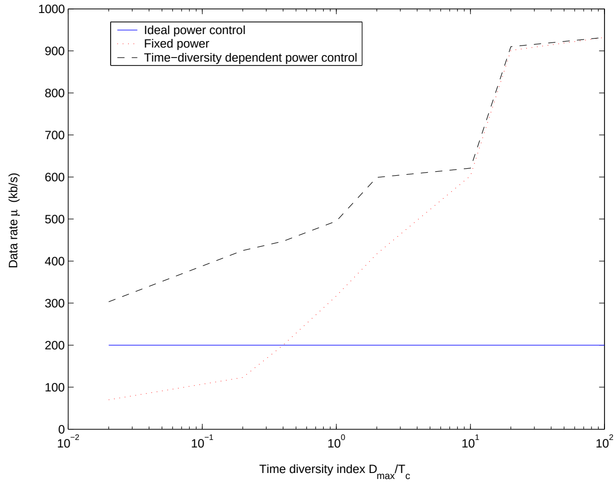



Figure 3: Data rate $\hat{\mu}(D_{max}, \varepsilon)$ vs. time diversity index $D_{max}/T_c$.

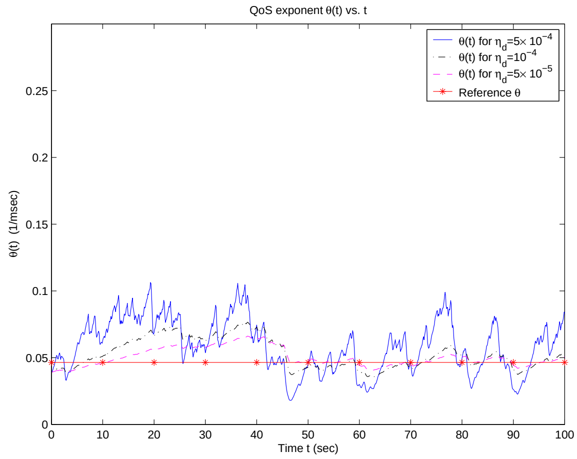
speed $v_s$ is from 0.011 to 56 m/s, which covers both downtown and highway speeds. Doppler rate $f_m$ is computed from $v_s$ using (29) and coherence time $T_c$ is computed from $f_m$ by (**??**). There is no need to simulate the case for $v_s = 0$ since for $v_s = 0$, the theory gives $\hat{\mu}(D_{max}, \varepsilon) = 200$, 0, and 200 kb/s under ideal power control, the fixed power scheme, and the time-diversity dependent power control, respectively. For $v_s = 0$, we have $\eta_g = 1$ and hence the time-diversity dependent power control reduces to ideal power control.

Figure 3 shows data rate $\hat{\mu}(D_{max}, \varepsilon)$ vs. time diversity index $D_{max}/T_c$ for the three schemes. It is clear that the larger the index $D_{max}/T_c$ is, the higher degree of time diversity the channel possesses.

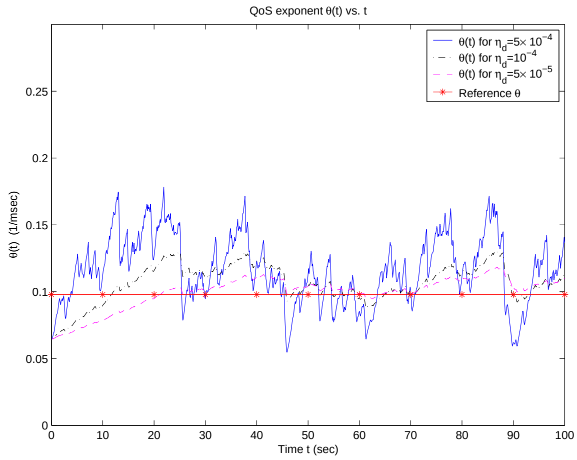
### 6.2.2 Accuracy of Exponentially Smoothed Estimate of $\theta$

This experiment is to show the accuracy of the exponentially smoothed estimate of $\theta$ via (13) and (14).

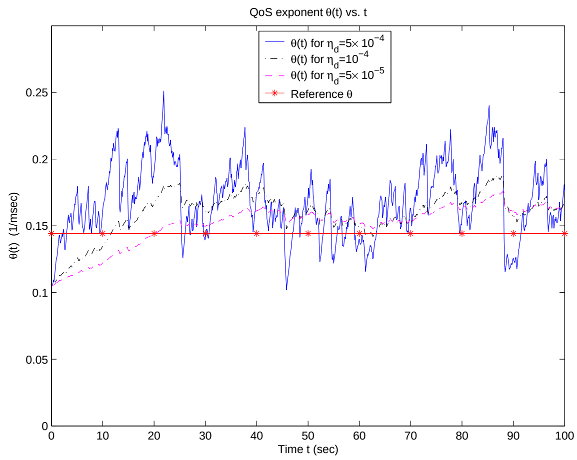
We do experiments with three constant mobile speeds, *i.e.*, $v_s = 5$, 10, and 15 m/s, respectively. For each mobile speed, we do simulations under three values of smoothing factor $\eta_d$ in (13), *i.e.*, $\eta_d = 5 \times 10^{-4}$, $10^{-4}$, and $5 \times 10^{-5}$, respectively. Since the objective is




(a)




(b)




(c)



Figure 4: $\hat{\theta}(t)$ vs. time $t$ for speed $v_s$ = (a) 5 m/s, (b) 10 m/s, and (c) 15 m/s.

to test the accuracy of the estimator of $\theta$, we do not use power control and scheduling; that is, we keep both the power and the channel allocation constant during the simulations.

We set the following parameters: source data rate $r_s = 50$ kb/s, $B_c = 300$ kHz, $P_0 = 20$ dBm, $E[g(t)] = -120$ dB, $\sigma_n^2 + P_I(t) = -100$ dBm, and $\Gamma_{link} = 6.3$ dB. Assume that the transmitter has a perfect knowledge about the channel power gain $g(t)$.

Figure 4 shows the estimate $\hat{\theta}(t)$ vs. time $t$ under different mobile speed $v_s$ and different smoothing factor $\eta_d$. The reference $\theta$ in the figure is obtained by the estimation algorithm in [6] at $t = 10^5$. It can be observed that for $\eta_d = 5 \times 10^{-5}$, the estimate $\hat{\theta}(t)$ gives the best agreement with the reference $\theta$, as compared to other values of $\eta_d$; for $\eta_d = 5 \times 10^{-4}$, the estimate $\hat{\theta}(t)$ reaches the reference $\theta$ in the shortest time (within 2 seconds), as compared to other values of $\eta_d$.

Hence, for the admission control in Algorithms 2 and 3, which requires quick estimate of $\theta$, we suggest to use $\eta_d = 5 \times 10^{-4}$; the estimate takes less than 2 seconds, which is tolerable in practice. For Algorithms 1 and 4, we also suggest to use $\eta_d = 5 \times 10^{-4}$ since we want the estimate $\hat{\theta}(t)$ to be more adaptive to time-varying mobile speed $v_s(t)$ and the resulting queueing behavior.

### 6.2.3 Performance under a Time-varying Mobile Speed

This experiment is to evaluate the performance of Algorithm 1 under a time-varying mobile speed, *i.e.*, under non-stationary small scale fading. Our objective is to see whether the power control and the scheduler in Algorithm 1 can achieve the required QoS.

For the channel gain process $g(t)$, we only simulate small scale fading; that is, there are no large scale path loss and shadowing. We set $E[g(t)] = -100$ dB. Other parameters are listed in Table 1.

Figure 5 shows the speed $v_s(t)$ of an MS vs. time $t$, which is the mobility pattern used

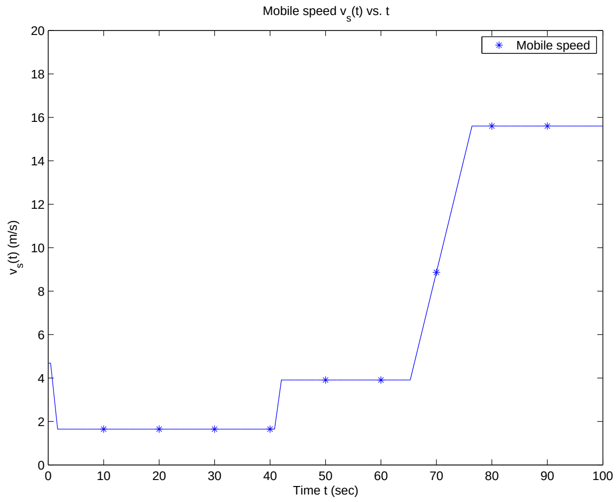



Figure 5: Speed behavior $v_s(t)$ of a mobile station in downtown.

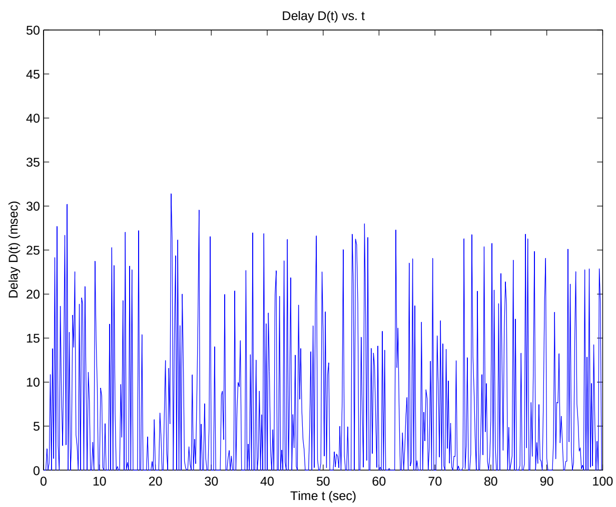



Figure 6: Delay $D(t)$ vs. time $t$.

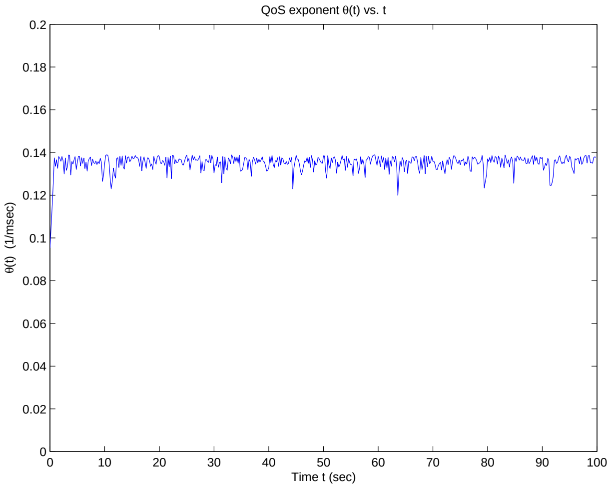



Figure 7: $\hat{\theta}(t)$ vs. time $t$ for varying mobile speed.

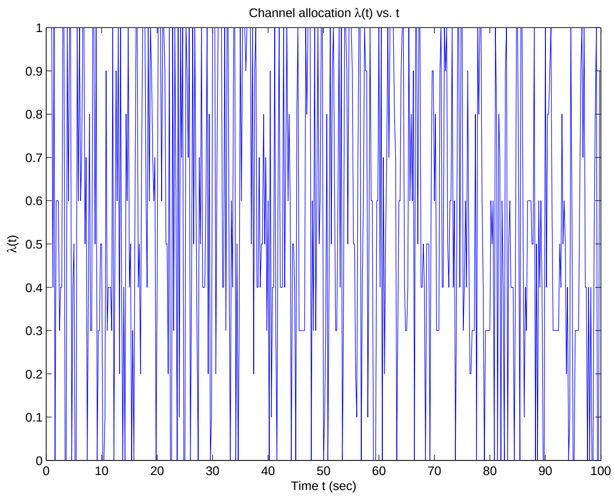



Figure 8: Channel allocation $\lambda(t)$ vs. time $t$.

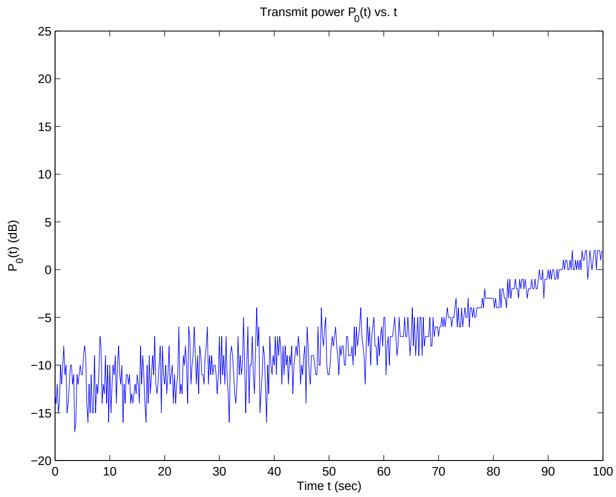



Figure 9: Transmit power $P_0(t)$ vs. time $t$ for large scale path loss.

in the simulation. Figure 6 depicts the delay $D(t)$ vs. time $t$. The simulation gives zero delay-bound violation for $D_{max} = 50$ ms, and hence the required QoS is met. Figure 7 plots QoS exponent $\hat{\theta}(t)$ vs. time $t$. Since we set $\gamma_{inc} = 1$ and $\gamma_{dec} = 1$, the resulting QoS exponent $\hat{\theta}(t)$ fluctuates around the required $\rho = -\log_e \varepsilon / D_{max} = 0.1382$. Figures 6 and 7 demonstrates the effectiveness of our scheduler in utilizing QoS exponent $\hat{\theta}(t)$ and the delay $D(t)$ for QoS provisioning.

Figure 8 illustrates how the channel allocation $\lambda(t)$ varies with time $t$. The average channel usage is 0.54.

In summary, Algorithm 1 can achieve the required QoS, under non-stationary small scale fading.

### 6.2.4 Performance under Large Scale Path Loss

This experiment is to evaluate the performance of Algorithm 1 under large scale path loss. We would like to see how the scheduler and the power control coordinate under large scale path loss.

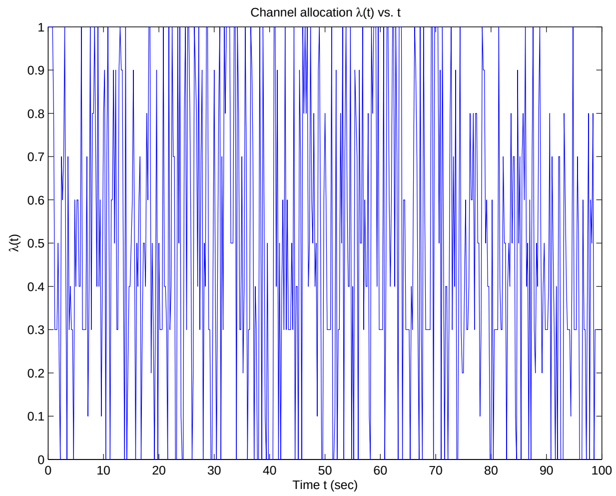



Figure 10: Channel allocation $\lambda(t)$ vs. time $t$ for large scale path loss.

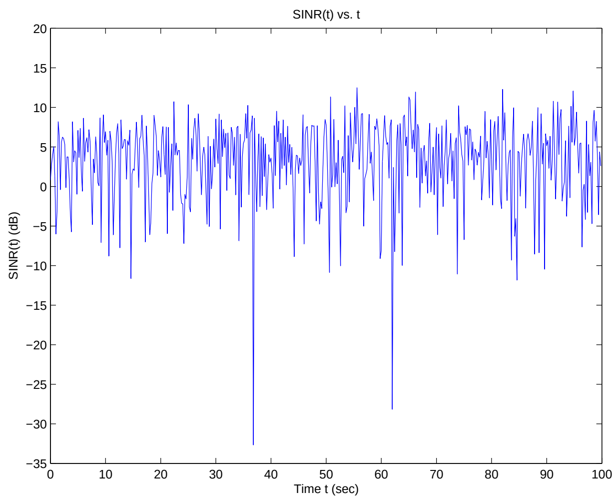



Figure 11: $SINR(t)$ vs. time $t$ for large scale path loss.

In the simulation, we use the same mobility pattern as shown in Figure 5 and generate large scale path loss according to Section 6.1.2. We do not simulate the shadowing effect here, which will be addressed in Section 6.2.5. The simulation parameters are listed in Table 1.

Figure 9 shows how the transmit power $P_0(t)$ evolves over time. The average transmit power is -7.4 dB. The power control is fast with a frequency of 1000 Hz, so that it can utilize time diversity. It is observed that as time elapses, the distance between the transmitter and the receiver increases and hence the expectation of the transmit power increases in order to mitigate the path loss.

Figure 10 depicts how the scheduler allocates the channel resource $\lambda(t)$ over time. The simulation gives zero delay-bound violation for $D_{max} = 50$ ms, and hence the required QoS is met. This demonstrates the good coordination between the power control and the scheduler; that is, the power control mitigates large scale path loss, while the scheduler utilizes time diversity in QoS provisioning. The average channel usage is 0.5. Figure 11 plots the received $SINR(t)$ vs. time $t$.

In summary, we observe the concerted efforts of the scheduler and the power control for QoS provisioning; the power control handles the effects of large scale path loss, while both the power control and the scheduler utilize time diversity. Different from ideal power control, our power control does not eliminate small scale fading, so that time diversity in small scale fading can be utilized.

#### 6.2.5 Performance under Shadowing

This experiment is to evaluate the performance of Algorithm 1 under shadowing.

In the first simulation, we use the same mobility pattern as shown in Figure 5 and generate large scale path loss and AR(1) shadowing process according to Section 6.1.2. The simulation parameters are listed in Table 1. Figure 12 depicts the transmit power $P_0(t)$ vs. time $t$. The simulation gives zero delay-bound violation for $D_{max} = 50$ ms, and hence the

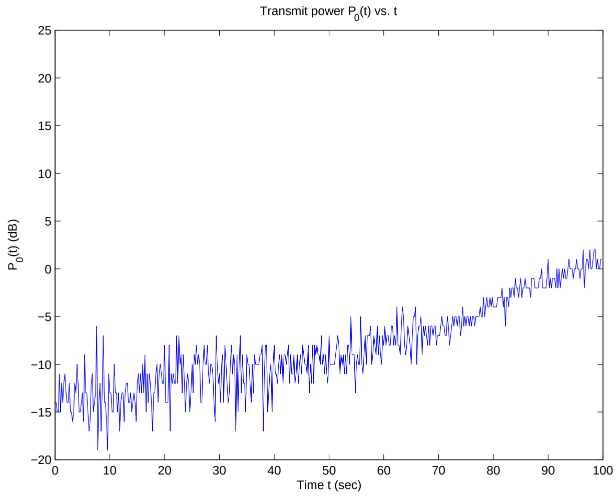



Figure 12: Transmit power $P_0(t)$ vs. time $t$ for AR(1) shadowing.

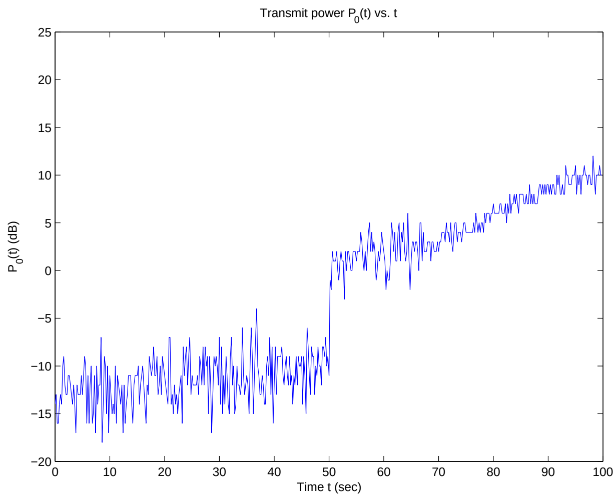



Figure 13: Transmit power $P_0(t)$ vs. time $t$ for the case of sudden shadowing.

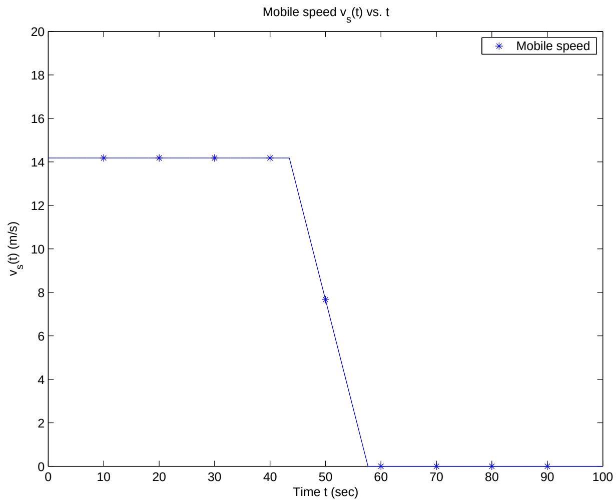



Figure 14: Speed behavior $v_s(t)$ of a mobile station in downtown.

required QoS is met. This demonstrates that the power control can mitigate both large scale path loss and shadowing effectively for QoS provisioning.

In the second simulation, we intentionally generate a shadowing of -10 dB at the 50-th second (which may happen when a car suddenly moves into the 'shadow' of a building) and see whether our power control can adapt and mitigate the shadowing effect. We use the same mobility pattern as shown in Figure 5 and generate large scale path loss according to Section 6.1.2. Figure 13 depicts the transmit power $P_0(t)$ vs. time $t$. It is observed that the power can quickly adapt to the sudden power change caused by the shadowing at the 50-th second in the figure. The simulation gives zero delay-bound violation for $D_{max} = 50$ ms, and hence the required QoS is met. Therefore, our time-diversity dependent power control can also mitigate the sudden shadowing effect.

In summary, Algorithm 1 is able to achieve good performance under shadowing.

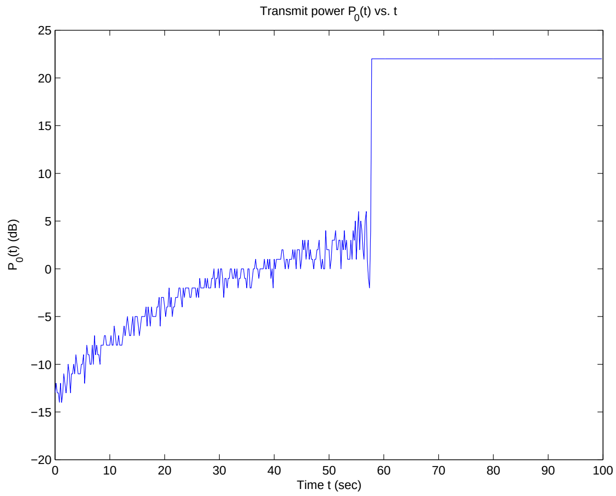



Figure 15: Transmit power $P_0(t)$ vs. time $t$ for the case of very low mobility.

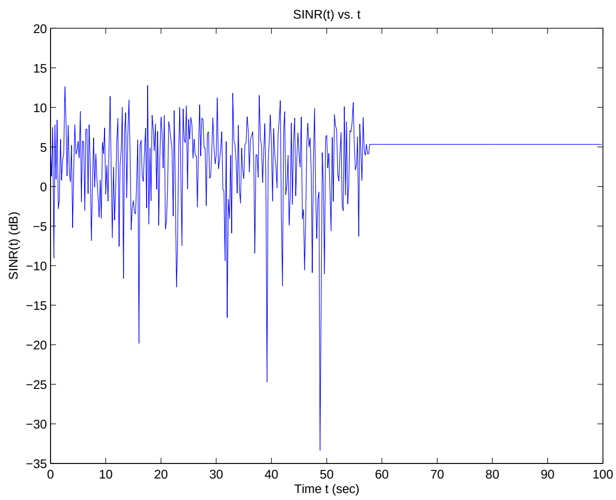



Figure 16: $SINR(t)$ vs. time $t$ for the case of very low mobility.

### 6.2.6 Performance under Very Low Mobility

This experiment is to evaluate the performance of Algorithm 1 under very low mobility, especially when the mobile speed is zero (due to red lights or traffic jams). Since our effective capacity approach and the scheduler require time diversity, they are not applicable to the case where the mobile speed is zero. Note that the effective capacity is zero when the mobile speed is zero. Hence, we rely on the power control to provide the required QoS.

Figure 14 shows the mobility pattern used in the simulation. We generate large scale path loss but do not simulate the shadowing effect. We assume perfect estimation of $SINR(t)$. The simulation parameters are listed in Table 1.

Figure 15 shows how the transmit power $P_0(t)$ varies over time. Figure 16 plots the received $SINR(t)$ vs. time $t$; this demonstrates that the power control converts the channel to an AWGN channel when the speed is zero between 57-th second and 100-th second. The simulation gives zero delay-bound violation for $D_{max} = 50$ ms, and hence the required QoS is met.

In summary, our power control can mitigate the effect of very low mobility and Algorithm 1 is able to guarantee the required QoS.

### 6.2.7 Admission Control

This experiment is to investigate whether our admission control in Algorithms 2 and 3 can be done quickly and accurately.

We use previous results in Sections 6.2.3 to 6.2.6. Define QoS exponent average $\hat{\theta}_{avg}(t) = \frac{1}{t+1}\sum_{\tau=0}^{t}\hat{\theta}(\tau)$. Figure 17 plots $\hat{\theta}_{avg}(t)$ vs. $t$ for the four cases, namely, a time-varying mobile speed, large scale path loss, shadowing with the AR(1) model, and very low mobility, which we investigated in Sections 6.2.3 to 6.2.6. We set the threshold $\theta_{th} = 0.9 \times \rho$. Since we are only concerned with the quickness of the estimation, we only plot the first ten seconds of

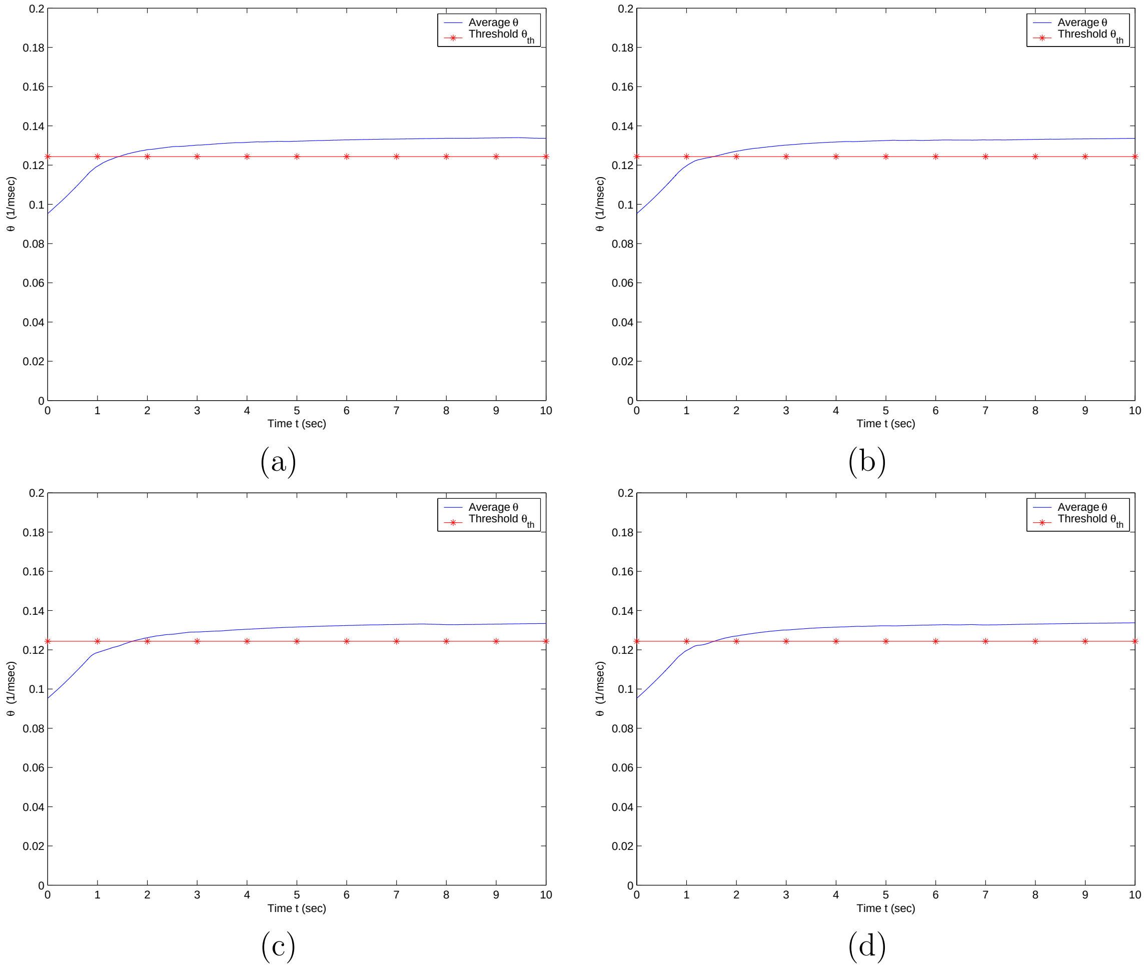



Figure 17: QoS exponent average $\hat{\theta}_{avg}(t)$ vs. $t$ for (a) time-varying mobile speed, (b) large scale path loss, (c) shadowing, and (d) very low mobility.

the simulations. Figure 17 shows that $\hat{\theta}_{avg}(t)$ is roughly an increasing function of $t$. Hence, $\hat{\theta}_{avg}(t)$ is a reliable QoS measure for admission control purpose. Moreover, the figure shows that for all the four cases, the system can obtain a reasonably accurate $\hat{\theta}_{avg}(t) \geq \theta_{th}$ within two seconds. Therefore, the system can make a quick and accurate admission decision.

# 7 Concluding Remarks

In this paper, we addressed an important issue in QoS provisioning for wireless networks, that is, robustness against large scale fading and non-stationary small scale fading, which can cause severe QoS violations. Equipped with the time-diversity dependent power control proposed in this paper and the effective capacity approach [6], we designed power control and scheduling mechanisms, which are robust in QoS provisioning against time-varying large scale path loss, shadowing, non-stationary small scale fading, and very low mobility. With these mechanisms, we proposed QoS provisioning algorithms for downlink and uplink transmissions, respectively; our QoS provisioning algorithms include channel estimation, power control, dynamic channel allocation, and adaptive transmission. The nice features of our QoS provisioning schemes are 1) power efficiency, 2) simplicity in QoS provisioning, 3) robustness against large scale fading and non-stationary small scale fading. Simulation results demonstrated the effectiveness of our proposed algorithms in providing QoS guarantees under various channel conditions.

# Acknowledgment

This work was supported by the National Science Foundation under the grant ANI-0111818.

# References


[1] C. Bettstetter, "Smooth is better than sharp: a random mobility model for simulation of wireless networks," in *Proc. 4th ACM International Workshop on Modeling, Analysis, and Simulation of Wireless and Mobile Systems (MSWiM)*, Rome, Italy, July 2001.

[2] Kalyanmoy Deb, "Multi-objective optimization using evolutionary algorithms," John Wiley & Sons, 2001.

[3] M. Gudmundson, "Correlation model for shadow fading in mobile radio systems," *IEE Electronics Letters*, vol. 27, no. 23, pp. 2145–2146, Nov. 1991.

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

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

[6] 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.

[7] 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`.

[8] D. Wu and R. Negi, "Downlink scheduling in a cellular network for quality of service assurance," *IEEE Transactions on Vehicular Technology*, vol. 53, no. 5, pp. 1547–1557, Sept. 2004.

[9] D. Wu and R. Negi, "Utilizing multiuser diversity for efficient support of quality of service over a fading channel," *IEEE Transactions on Vehicular Technology*, vol. 54, no. 3, pp. 1198–1206, May 2005.

[10] D. Wu, "QoS provisioning in wireless networks," to appear in *Wireless Communications and Mobile Computing*, Wiley.

# Biography for Dapeng Wu

Dapeng Wu received B.E. in Electrical Engineering from Huazhong University of Science and Technology, Wuhan, China, in 1990, M.E. in Electrical Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 1997, and Ph.D. in Electrical and Computer Engineering from Carnegie Mellon University, Pittsburgh, PA, in 2003.

Since August 2003, he has been with Electrical and Computer Engineering Department at University of Florida, Gainesville, FL, as an Assistant Professor. His research interests are in the areas of networking, communications, multimedia, signal processing, and information and network security. He received the IEEE Circuits and Systems for Video Technology (CSVT) Transactions Best Paper Award for Year 2001.

Currently, he is an Associate Editor for IEEE Transactions on Wireless Communications, IEEE Transactions on Circuits and Systems for Video Technology, IEEE Transactions on Vehicular Technology, and International Journal of Ad Hoc and Ubiquitous Computing. He is also a guest-editor for IEEE Journal on Selected Areas in Communications (JSAC), Special Issue on Cross-layer Optimized Wireless Multimedia Communications. He served as Program Chair for IEEE/ACM First International Workshop on Broadband Wireless Services and Applications (BroadWISE 2004); and as a technical program committee member of over 30 conferences. He is Vice Chair of Mobile and wireless multimedia Interest Group (MobIG), Technical Committee on Multimedia Communications, IEEE Communications Society. He is a member of the Best Paper Award Committee, Technical Committee on Multimedia Communications, IEEE Communications Society.

# Biography for Rohit Negi

Rohit Negi received the B.Tech. degree in Electrical Engineering from the Indian Institute of Technology, Bombay, India in 1995. He received the M.S. and Ph.D. degrees from Stanford

University, CA, USA, in 1996 and 2000 respectively, both in Electrical Engineering. He has received the President of India Gold medal in 1995.

Since 2000, he has been with the Electrical and Computer Engineering department at Carnegie Mellon University, Pittsburgh, PA, USA, where he is an Associate Professor. His research interests include signal processing, coding for communications systems, information theory, networking, cross-layer optimization and sensor networks.