Hierarchical Queue-Length-Aware Power Control for Real-Time Applications over Wireless Networks

Xiaochen Li, Xihua Dong and Dapeng Wu
Department of Electrical and Computer Engineering
University of Florida, Gainesville, FL 32611
Correspondence author: Prof. Dapeng Wu, wu@ece.ufl.edu,
http://www.wu.ece.ufl.edu

Abstract

We consider the problem of optimal power control for quality-of-service-assured wireless communication. The quality of service (QoS) measures of our consideration are a triplet of data rate, delay, and delay bound violation probability. Our target is to develop power control laws that can provide delay guarantees for real-time applications over wireless networks. The power control laws that aim at optimizing certain physical-layer performance measures, usually adapt the transmission power based on the channel gain; we call these "channel-gain-based" (CGB) power control (PC). In this paper, we show that CGB-PC laws achieve poor link-layer delay performance. To improve the performance, we propose a novel scheme called hierarchical queue-length-aware (HQLA) power control. The key idea is to combine the best features of the two PC laws, i.e., a given CGB-PC law and the clear-queue PC law; here, the clear-queue PC is defined as a PC law that uses a transmission power just enough to empty the queue at the link layer. We analyze our proposed HQLA-PC scheme by the matrix-geometric method. The analysis agrees well with the simulation results. More importantly, our results show that the proposed HQLA power control scheme is superior to the corresponding CGB-PC in both average power consumption and effective capacity.

Index Terms

Delay-constrained communications, power control, queuing theory, matrix-geometric method, effective capacity

I. Introduction

Future wireless networks are expected to support real-time applications, such as streaming multimedia, online games or remote medical monitoring and diagnosis, which require quality of service (QoS) guarantees. For the delay sensitive real-time services, two QoS requirements, reliability and transmission delay, are of great importance. Reliability accounts for the quality

This work was supported in part by an Intel gift, the US National Science Foundation under grant DBI-0529012 and CNS-0643731, and the US Office of Naval Research under grant N000140810873.

of the information transmitted. In physical layer, it is referred to as bit error rate (BER). Each application has its own BER requirement. Many physical layer technologies such as coding and interleaving have been invented to improve the BER performance. Another QoS requirement, the transmission delay, comes from the link layer, which affects the "real-time" property and user experiences. For good user experiences, such as watching a streaming video, the packet delay should be kept small and consistent so that the user can enjoy a smooth screen.

The success of these real-time applications critically depends on how much spectrum efficiency and power efficiency these network can achieve. This paper focuses on power efficiency. The transmission power is an important physical layer resource for a wireless communication system. On one hand, for a particular user, the larger the transmission power, the better performance can be expected. On the other hand, one user's strong signal will interfere with other users' on going links and degenerate their performance. Also for mobile terminals, lower transmission power leads to longer standby and talking time. In this paper, we study the problem of determining a PC law that minimizes transmission power while satisfying the required QoS, especially, delay guarantee, for a given spectrum.

Existing PC laws fall into two categories, namely, CGB PC and queue-length-aware (QLA) PC. CGB-PC adapts the transmission power based on the channel gain (physical-layer information), with the aim of optimizing certain physical-layer performance measures such as physical-layer data rate, BER, or signal-to-noise ratio (SNR). An example of CGB-PC is time-domain water filling (TDWF) proposed by Goldsmith and Varaiya [1], which maximizes the physical-layer data rate (Shannon's ergodic capacity), subject to the average power constraint. Since the constraint only comes from the physical layer, in the optimal solution, the physical layer information, i.e. the instantaneous channel gain, uniquely determines the transmission power. The price for the TDWF PC strategy to achieve the high capacity is that the packets may experience an arbitrary long delay (coding delay). To find the upper bound of capacity with finite coding delay, people studied the delay-constrained capacity of fading channels [2] and the power control schemes which maximize those capacities. Existing delay-constrained capacity notions include outage capacity [3], delay-limited capacity [4], and expected capacity [5]. The optimal causal power control scheme which maximizes the expected capacity subject to average power constraint is also TDWF [5]. An optimal non-causal power control scheme is studied in [6], where the channel gains of all future fading states are assumed to be known at the beginning of the transmission. The optimal power control that maximizes outage capacity under non-causal channel state information, is studied in [7]; The optimal power control that maximizes outage capacity under causal channel state information, is studied in [5].

QLA-PC adapts the transmission power based on the queue-length (link-layer information) and possibly the channel gain (physical-layer information), with the aim of optimizing certain link-layer performance measures such as link-layer data rate and delay bound violation probability. In this paper, we show that CGB-PC laws achieve poor link-layer delay performance. To improve the performance, QLA-PC is needed. Existing QLA-PC laws [8] aim at minimizing the transmission power under the constraint on the average delay. But average-delay guarantee may not satisfy the requirements of delay-sensitive applications; e.g., using a handheld device to watch mobile TV over WiMax, requires certain delay bound violation probability, which cannot be specified by

average delay since average delay cannot specify the (tail) probability distribution function; e.g., for a given delay bound (say, 1 second), two systems with the same average delay of 500 ms could have quite different delay bound violation probabilities, e.g., 40% vs. 0.1%. Different from guaranteeing average delay, this paper considers *statistical delay guarantees* or statistical QoS [9], i.e., *the triplet of data rate, delay bound, and delay bound violation probability (DBVP)*, which is more general and challenging than guaranteeing average delay. That the statistical QoS constraint is satisfied means that for a constant source data rate, the probability that the transmission delay of a packet exceeding the delay bound is smaller than DBVP.

Under the statistical QoS constraint, the maximum capacity achieved by the TDWF PC strategy is no longer achievable. Wu and Negi [9] studied the capacity of a wireless link under this constraint, namely the effective capacity, which describes the maximum constant data rate that the system can sustain under the statistical QoS constraints. Tang and Zhang [10] studied the optimal PC scheme that maximize the effective capacity subject to a average power constraint. They still assumed that the transmission power is controlled by the instantaneous channel gain and did not consider the queue length.

In fact, since for many real-time applications, the data source is generated on the fly, e.g. video captured by a webcamera, the instantaneous channel capacity provided by the CGB PC scheme may be greater than the backlog in the transmitter buffer. Under this scenario, if the transmitter still schedules that power to transmit, it will be idle after the buffer is cleared. As suggested in the lazy scheduling [11], we can always find a more energy/power efficient strategy by preventing the transmitter from idle. On the other words, if we keep the average power unchanged, the effective capacity will be increased. We show later in the simulation that the effective capacity achieved by their PC can be exceeded.

In this paper, we focus on QLA-PC laws that aim at minimizing the transmission power under the constraint on statistical delay guarantees. Specifically, we propose a novel scheme called HQLA PC. The key idea is to combine the best features of the two PC laws, i.e., a given CGB-PC law and the clear-queue (CQ) PC law; here, the CQ-PC is defined as a PC law that uses a transmission power just large enough to empty the queue (i.e., transmit all the buffered bits in the link-layer queue). The HQLA PC scheme is not optimal in terms of maximizing the effective capacity yet it is practically usable and dramatically increases the effective capacity.

To analytically analyze the performance of the proposed scheme, we model the queue length as a Markov chain and numerically calculated the steady state queue length distribution by means of matrix-geometric method [12]. From the steady state distribution,the average power can be obtained. Both the numerical results and the simulation results show that the proposed HQLA PC scheme reduces the average power greatly, comparing to the CGB PC scheme, for the same statistical QoS constraint. And for the same average power, the HQLA PC scheme significantly increases the effective capacity.

The remainder of this paper is organized as follows. Section II describes the system structure and the Markov chain model of the system. Section III introduces the proposed HQLA PC scheme and analyze its performance in terms of the resulting queue length distribution and the average power. Section IV presents the simulation results. Section V concludes the paper.

II. System Description

The abbreviations used in this paper are listed in Table V. We consider a node to node transmission over wireless fading channels. The wireless channel in question is modeled as a time-slotted flat fading channel. The slot has a fixed duration of T_s sec, which is assumed to be small enough that the channel gains are constant, yet large enough that ideal channel codes can achieve Shannon's capacity over that duration. The channel gain of each slot is assumed to be i.i.d. Rayleigh distributed. Notice that the i.i.d. assumption is only for the analytical analysis purpose, the proposed power control scheme can be applied to any channel conditions.

A. Structure of Data Source and Transmitter

The structure of data source and transmitter is illustrated in Fig. 1. A data source generates bit stream at a constant rate $\tilde{\mu}$ bits/sec. The bit stream is first stored in a source buffer where data processing is performed. This procedure may involve block processing such as frequency domain video coding, block coding or formatting the bits stream into a packet. In general, the output of the source buffer is no longer consecutive. Assume that the output of the source buffer is blocked data with a constant rate of one block per slot. Each block contains μ bits, where μ is an integer and $\mu \geq \tilde{\mu}T_s$. The equality takes place when there is no extra data added in the data processing procedure and a proper chosen T_s such that $\tilde{\mu}T_s$ is an integer. The later condition is easy to meet for large $\tilde{\mu}$. The blocked data is fed into the transmission buffer at the beginning of each slot and is served (transmitted) in a first-in-first-out (FIFO) fashion. We assume that the capacity of the transmission buffer is infinite, therefore all the bits can be served eventually. Since for the transmitter, only the block size μ matters, we will use μ to denote the arrival rate and discard $\tilde{\mu}$ in the following discussions.

In the model described above, the transmission delay is simply the waiting time in the transmission buffer. For a constant arrival rate, the transmission delay (count in slots) is equivalent to the queue length (count in bits) up to a scalar and a fraction. In fact, denote q the queue length in the buffer when a certain bit is transmitted, the transmission delay of this bit ranges from $\lfloor q/\mu \rfloor$ to $\lfloor q/\mu \rfloor + 1$, where $\lfloor x \rfloor$ finds the largest integer that is not greater than x. By studying the queue length distribution of the system , the statistical measure of the transmission delay can be obtained. In the following subsections, we demonstrate that the queue length forms a Markov chain, and in subsection III-B the steady state queue length distribution is obtained by matrix-geometric method.

B. Markov Chain Model

According to the time-slotted flat fading assumption, the channel gain remains unchanged during one slot. The transmitter will adapt the modulation and coding scheme to achieve the Shannon's capacity. Therefore the queue length within one slot may have various behaviors depending on the particular scheme used. The random process of continuous time queue length $q(t)$ may not have Markovian property. However if we focus on the queue length at the beginning of each slot, and the following two conditions are satisfied, 1) the channel gain of each slot is

i.i.d. 2) the transmission power is a function of the current slot channel gain and queue length, we will have a Markov chain, as described following.

Let $q(n)$ denote the queue length at the beginning of the nth slot, or on the other words, it is the number of bits waiting in the transmission buffer before the arrival of the new block and the transmission of that slot; $s(n)$ denotes the number of bits that will be transmitted during the nth slot. As illustrated in Fig. 2. the queue length update function is

$$q(n+1) = [q(n) + \mu - s(n)]^+, \tag{1}$$

where $[x]^+$ is the Lindley's operator which means $max(x, 0)$.

$s(n)$ depends on the bandwidth which is fixed, the current channel gain $g(n)$ and the transmission power $P(n)$. As indicated in assumption 2), $P(n)$ is a function of $g(n)$ and the current slot queue length $q(n)$. Therefore $s(n)$ is also a function of $g(n)$ and $q(n)$. Consequently, $q(n+1)$ can be also written as a function of $g(n)$ and $q(n)$

$$q(n+1) = h(g(n), q(n)). \tag{2}$$

The relationship between $q(n+1)$, $g(n)$, $q(n)$ is illustrated in Fig. 3. Notice that $q(n)$ is the queue length at the beginning of the nth slot, it is independent of $g(n)$. Since $\{g(n)\}$ are i.i.d by assumption 1), if $q(n)$ is known, the value of $q(n+1)$ is uniquely determined by $g(n)$ and is irrelative to the previous queue length $q(k)$, $k < n$. Therefore under the two assumptions mentioned above, $q(n)$ forms a discrete-time Markov chain. Furthermore, if we require that $s(n)$ be an integer, which is almost always true in practice, $q(n)$ forms a discrete-time, discrete-state Markov chain.

III. Hierarchical queue-length-aware power control scheme

In this section we introduce the proposed HQLA PC scheme, and analyze the steady state queue length distribution, from which the average power can be obtained.

As mentioned in section I, for real-time applications, since the data to be transmitted is generated on the fly, it is possible that the capacity provided by the CGB PC scheme exceeds the backlog in the buffer. To schedule the transmission power more efficiently, we need to design a PC scheme which considers both the channel gain and the queue length. Without loss of generality, denote the transmission power as $\tilde{P}(n) = \tilde{P}(g(n), q(n))$. The optimal PC scheme should maximize the throughput while satisfies the statistical QoS constraint.

The statistical QoS requirement is satisfied means that for a certain constant arrival rate μ, the probability that the delay bound D_{max} is violated is not greater than the DBVP

$$\text{Prob}[\ D(\infty) \geq D_{max}\ ] \leq \epsilon, \tag{3}$$

Where ϵ is the user specified DBVP, and $D(\infty)$ is the transmission delay when the queue enters the steady state (the transient process is neglected). The optimization problem can be expressed

as

$$\begin{aligned} \max_{\tilde{P}(g(n),q(n))} \quad & \mu \\ s.t. \quad & \text{Prob}[\ D(\infty) \geq D_{max}\ ] \leq \epsilon \\ & \text{average power} = \bar{P}, \end{aligned} \tag{4}$$

where $\bar{P}$ is the average power constraint. In practice, another optimization problem is also important: to support a given throughput, what is the optimal power control scheme which has the lowest average power. By alternating the objective function and the power constraint in (4), we have the second optimization problem,

$$\begin{aligned} \min_{\tilde{P}(g(n),q(n))} \quad & \text{average power} \\ s.t. \quad & \text{Prob}[\ D(\infty) \geq D_{max}\ ] \leq \epsilon \\ & \mu = \mu_0, \end{aligned} \tag{5}$$

where μ_0 is the throughput constraint.

The above two optimization problems can be related by a two objective optimization problem. The two objectives are throughput μ and average power $\bar{P}$. Denote $\{\mu^*, \bar{P}^*\}$ the optimal solution. The optimality is in a Pareto sense, which means that there does not exist any other pair of $\{\mu, \bar{P}\}$, such that the following two conditions are satisfied simultaneously,

$$\begin{cases} \mu \geq \mu^* \\ \bar{P} \leq \bar{P}^*. \end{cases} \tag{6}$$

$\{\mu^*, \bar{P}^*\}$ is not unique. Actually it forms a curve, called Pareto curve, in a two-dimensional plane spanned by μ and $\bar{P}$. If the Pareto curve is continuously increasing (Intuitively, this is true in our problem since higher average power is allowed, higher throughput can be expected, and vice versa. However we are not intent to provide proof on this since it is beyond the scope of this paper), the two optimization problems (4) and (5) are equivalent, i.e. if $\tilde{P}(g(n), q(n))$ is optimal to (4), it is also optimal to (5).

Both of the two optimization problems are not easy to solve. The transmission power is averaged over both the channel gain and the queue length. However, for most queueing problems, the closed form of the steady state queue length distribution is not available. In this paper, we propose a PC scheme which is not optimal yet still superior to the CGB PC scheme in terms of the throughput or average power.

The proposed HQLA PC scheme consists of two components, the CGB PC and the CQ PC . The CGB PC part could be any PC scheme which determines the transmission power only according to the current slot channel gain, i.e. TDWF. The CQ PC finds the minimum power needed to clear the current slot queue. The two parts work independently. After each of the two parts determines the transmission power, the smaller one is chosen. The hierarchical structure is easy to be implemented or upgraded from the existing CGB PC system. In practice, there is always a peak power constraint. A third component, the peak power component, should be

added to the HQLA PC scheme. At each power control cycle, the smallest one of the three components is chosen.

In subsection III-A, we discuss the proposed HQLA PC scheme in detail. In subsection III-B we analyze the steady state queue length distribution of the proposed scheme. The average power is given in subsection III-C. Subsection III-D discusses the effective capacity with power control. The HQLA with peak power constraint is discussed in subsection III-E.

A. Hierarchical Queue-Length-Aware Power Control Scheme

Denote B the bandwidth of the channel and $f(g(n))$ the CGB PC scheme. By Shannon's capacity formula, the instantaneous channel capacity provided by the channel gain based PC scheme is

$$s(n) = \lfloor BT_s \log_2(1 + f(g(n))g(n)) \rfloor. \tag{7}$$

From (1), $s(n)$ should not exceed $q(n) + \mu$ because the maximum number of bits that can be transmitted during one slot is the number of bits remained in the buffer plus the new arrival of that slot. Therefore

$$\lfloor BT_s \log_2(1 + f(g(n))g(n)) \rfloor \leq q(n) + \mu. \tag{8}$$

When (8) is violated, the actual transmission time T_{actual} will be smaller than the slot length T_s; while if the transmission time is set to be T_s, the transmission power can be reduced to

$$P_0(g(n), q(n)) = \frac{2^{\frac{q(n)+\mu}{BT_s}} - 1}{g(n)}. \tag{9}$$

One can also use other pair of $\{T, P(n)\}$, $T_{actual} \leq T \leq T_s$, $P_0(g(n), q(n)) \leq P(n) \leq f(g(n))$ to schedule the transmission under this situation. Obviously the pair $\{T_s, P_0(g(n), q(n))\}$ minimizes the power. Choosing transmission time equal to the slot time T_s can be also viewed as the lazy scheduling within one slot, which minimizes the total energy of a transmission with an arbitrary arrival pattern and a deadline constraint. Lazy scheduling always schedules the current work load evenly between the current time and the deadline. In our case all the work loads come at the beginning of the slot and the deadline is T_s sec after that. Therefore the pair $\{T_s, P_0(g(n), q(n))\}$ minimizes both the power and the energy within one slot. These lead to the HQLA PC strategy: when the power determined by $f(g(n))$ can not clear the queue, keep $P(n) = f(g(n))$ unchanged otherwise reduce the power from $f(g(n))$ to $P_0(g(n), q(n))$

$$\tilde{P}(n) = \min \left[f(g(n)), \frac{2^{\frac{q(n)+\mu}{BT_s}} - 1}{g(n)} \right]. \tag{10}$$

B. Steady State Queue Length Distribution

In the queueing theory's perspective, the system has group arrivals with a fixed group size μ and a fixed arrival rate $1/T_s$. The service facility has multiple servers. Each of them has a fixed service rate $1/T_s$ and serves (transmits) one bit per time. The number of server, which equals $s(n)$, varies every T_s sec and synchronizes with the arrival. It is different from the general G/G/m queue model because although the arrival and departure process can be said to have general interval pattern, the number of server is not deterministic.

As described in subsection II-B, the sequence $q(n)$ forms a discrete-time discrete-state Markov chain. Since we assume the buffer has infinite capacity, the dimension of the states space is infinite. Denote the infinite dimension row vector $\mathbf{x} = [x_0, x_1, \dots]$ the steady state distribution of the queue where

$$x_i = \lim_{n\to\infty} \text{Prob}[\ q(n) = i\ ].\quad i = 0, 1, 2 \dots . \tag{11}$$

In an irreducible and aperiodic homogeneous Markov chain, the steady state distribution always exists and is independent of the initial state probability distribution. Either $x_i = 0$ for all i, where there exists no stationary distribution, or $x_i > 0$ for all i and the value of x_i are uniquely determined through the equilibrium equation [13, page 29]

$$\mathbf{x} = \mathbf{xP}, \tag{12}$$

$$\sum_{i=0}^{\infty} x_i = 1, \tag{13}$$

where the infinite dimension matrix $\mathbf{P}$ is the one step transition probability matrix, with its element $p_{i,j}$ the one step transition probability from state i to state j.

In the following discussion we show that the Markov chain $q(n)$ is homogeneous under the proposed HQLA PC scheme and the i.i.d. channel assumption. Then the transition probability matrix $\mathbf{P}$ is calculated from the marginal distribution of $g(n)$. The irreducible and aperiodic property can be easily obtained from matrix $\mathbf{P}$.

By contradiction, suppose the Markov chain is not homogeneous. Denote $p_{i,j}(n)$ the one step transition probability from state i to state j at slot n,

$$\begin{aligned} p_{i,j}(n) &= \text{Prob}[\ q(n+1) = j | q(n) = i\ ], \quad i \geq 0, j \geq 0 \\ &= \text{Prob}[\ [i + \mu - s(n)]^+ = j\ ] \\ &= \text{Prob}[\lfloor BT_s \log_2(1 + \tilde{P}(n)g(n)) \rfloor = i + \mu - j]. \end{aligned} \tag{14}$$

The last equality holds because $\tilde{P}(n)$ guarantees that $i + \mu - s(n) \geq 0$. From (10), $\tilde{P}(n)$ is a function only of $g(n)$ when the current queue length $q(n)$ is known to be i. Therefore the marginal distribution of the channel gain $g(n)$ uniquely determines $p_{i,j}(n)$. When $g(n)$ is i.i.d., $p_{i,j}(n) = p_{i,j}$ and $\mathbf{P}(n) = \mathbf{P}$ are irrelevant to the slot index n. Hence the Markov chain $q(n)$ is homogeneous.

Obviously $p_{i,j} = 0$ for $j > i + \mu$. When $j = 0$, the queue is cleared. According to (10), $f(g(n)) \geq P_0(g(n), q(n))$ because $P_0(g(n), q(n))$ is the minimum power needed to clear the queue. Therefore

$$\begin{aligned} p_{i,0} &= \text{Prob}[f(g(n)) \geq P_0(g(n), q(n))] \\ &= \text{Prob}[f(g(n))g(n) \geq 2^{\frac{i+\mu}{BT_s}} - 1]. \end{aligned} \tag{15}$$

When $0 < j \leq i + \mu$, $f(g(n)) < P_0(g(n), q(n))$, the transition probability is given by

$$p_{i,j} = \text{Prob}[2^{\frac{i+\mu-j}{BT_s}} - 1 \leq f(g(n))g(n) < 2^{\frac{i+\mu-j+1}{BT_s}} - 1]. \tag{16}$$

Define a set of μ by μ matrix $\{\mathbf{A}_i\}$, $\{\mathbf{B}_i\}$, $i = 0, 1, 2, \ldots$. The kth row and lth column element of matrix $\mathbf{A}_i$ and $\mathbf{B}_i$ are

$$(\mathbf{A}_i)_{k,l} = p_{k+i\mu,l+\mu}, \tag{17}$$

$$(\mathbf{B}_i)_{k,l} = p_{k+i\mu,l}, \tag{18}$$

$$0 \leq k,l \leq \mu - 1.$$

Notice that when $j > 0$, the transition probability is uniquely determined by the difference of the two states $j - i$. Therefore $p_{i,j} = p_{i+k,j+k}$, $k \geq 0$. The matrix $\mathbf{P}$ has the repetitive structure of the form

$$\mathbf{P} = \begin{bmatrix} \mathbf{B}_0 & \mathbf{A}_0 & \mathbf{0} & \mathbf{0} & \ldots \\ \mathbf{B}_1 & \mathbf{A}_1 & \mathbf{A}_0 & \mathbf{0} & \ldots \\ \mathbf{B}_2 & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \ldots \\ \mathbf{B}_3 & \mathbf{A}_3 & \mathbf{A}_2 & \mathbf{A}_1 & \ldots \\ . & . & . & . & \ldots \\ . & . & . & . & \ldots \\ . & . & . & . & \ldots \end{bmatrix}. \tag{19}$$

All the states are connected hence the Markov chain is irreducible. If the chain starts at state i, it can return to state i after arbitrary steps. Therefore the chain is aperiodic.

Since the Markov chain $q(n)$ is irreducible, aperiodic and homogenous, we can solve for $\mathbf{x}$ by (12) and (13). If the transition probability matrix $\mathbf{P}$ has the form in (19), the eigenvector problem (12) can be solved by the matrix-geometric method up to a scalar, which can be obtained by the normalization function (13). Partition the infinite dimension row vector $\mathbf{x}$ into a set of $1 \times \mu$ row vectors $\mathbf{x}_i = [x_{i\mu}, x_{i\mu+1}, ..., x_{(i+1)\mu-1}]$, $i \geq 0$. According to the matrix-geometric method,

$$\mathbf{x}_i = \mathbf{x}_0 \mathbf{R}^i, \tag{20}$$

where $\mathbf{R}$ is the solution of

$$\mathbf{R} = \sum_{k=0}^{\infty} \mathbf{R}^k \mathbf{A}_k, \tag{21}$$

and $\mathbf{x}_0$ is the solution of

$$\mathbf{x}_0 = \mathbf{x}_0 \sum_{k=0}^{\infty} \mathbf{R}^k \mathbf{B}_k. \tag{22}$$

Both equation (21) and (22) can be solved in an iterative way, see [14] for detail discussions. Finally apply the normalization constraint $\sum_{i=0}^{\infty} x_i = 1$ to get the steady state queue length distribution.

C. Average Power

In this subsection, we calculate the average power of the HQLA PC scheme on 1) constant power (CONST) $f_c(g(n)) \equiv \bar{P}$; 2) TDWF, $f_w(g(n)) = [C^{-1} - g(n)^{-1}]^+$, where the cutoff threshold C is properly chosen such that the average power $\mathbf{E}_{g(n)}[f_w(g(n))] = \bar{P}$, where $\mathbf{E}_{g(n)}(x)$ averages x over $g(n)$.

Assuming channel gain $g(n)$ is i.i.d. Rayleigh distributed with probability density function $f_{ch}(g) = \lambda e^{-\lambda g}$. From (15) and (16), the transition probability for CONST $\{p^c_{i,j}\}$ is:

$$\begin{cases} p^c_{i,0} = e^{-\frac{\lambda}{\bar{P}}(2^{\frac{i+\mu}{BT_s}}-1)} \\ p^c_{i,j} = e^{-\frac{\lambda}{\bar{P}}(2^{\frac{i+\mu-j}{BT_s}}-1)} - e^{-\frac{\lambda}{\bar{P}}(2^{\frac{i+\mu-j+1}{BT_s}}-1)} \\ \qquad 0 < j \le i+\mu \\ p^c_{i,j} = 0, \quad j > i+\mu. \end{cases} \tag{23}$$

For TDWF, there are two situations that account to the event $q(n+1) = q(n) + \mu$. One is the same as CONST, the transmission power is so small that $s(n)$ is smaller than one; another situation is $g(n) < C$, $\tilde{P}(n) = f(g(n)) = 0$. The channel gain is smaller than the cutoff threshold, the transmitter will not transmit at the current slot at all. The transition probability for TDWF $\{p^w_{i,j}\}$ is:

$$\begin{cases} p^w_{i,0} = e^{-\lambda C 2^{\frac{i+\mu}{BT_s}}} \\ p^w_{i,j} = e^{-\lambda C 2^{\frac{i+\mu-j}{BT_s}}} - e^{-\lambda C 2^{\frac{i+\mu-j+1}{BT_s}}} \\ \qquad 0 < j < i+\mu \\ p^w_{i,i+\mu} = 1 - e^{-\lambda C 2^{\frac{1}{BT_s}}} \\ p^w_{i,j} = 0, \quad j > i+\mu. \end{cases} \tag{24}$$

The average power is

$$\mathbf{E}_{g,q}(\tilde{P}) = \sum_{q=0}^{\infty} x_q \int_0^{\infty} f_{ch}(g)\tilde{P}(g,q)dg, \tag{25}$$

where $\mathbf{E}_{g,q}(x)$ averages x over both g and q. The slot index n in (25) is omitted because at steady state, both channel gain and queue length are stationary. (25) can be numerically evaluated.

D. Effective Capacity with Power Control

Effective capacity is defined as the maximum arrival rate that the system can sustain for a given $\{D_{max}, \epsilon\}$. It is a dual problem of effective bandwidth, where the departure process is constant and the arrival process is random. The validity of effective bandwidth requires that the arrival process be stationary [15, Page 291]. Analogously, the validity of effective capacity requires that the virtual departure process (the capacity provided by the channel) be stationary. Notice that the HQLA PC does not change $s(n)$ comparing to its corresponding CGB PC. And for CGB PC, the departure process $s(n)$ is determined by the channel gain $g(n)$. If $\{g(n)\}$ is stationary, $\{s(n)\}$ is also stationary. Therefore, the effective capacity is valid for CGB PC and HQLA PC (except for channel inversion PC where $P \propto 1/g$, $s(n)$ becomes a deterministic value, not random value).

Theoretically, for CGB PC scheme, the effective capacity can be calculated by [9]

$$\alpha(u) = -\frac{1}{u} \lim_{t\to\infty} \frac{1}{t} \log \mathbf{E}[e^{-u\sum_{i=0}^{t} s(i)}], \tag{26}$$

where u is the QoS exponent which relates the D_{max} and ϵ. In fact, using the theory of large deviations, it can be shown that the probability of steady state queue length $q(\infty)$ exceeding the threshold Q_{max} satisfies [16]

$$\text{Prob}[\ q(\infty) \geq Q_{max}\ ] \simeq e^{-uQ_{max}}. \tag{27}$$

From II-A, $D(\infty) \in [\lfloor q(\infty)/\mu \rfloor, \lfloor q(\infty)/\mu \rfloor + 1]$, If the fraction part can be neglected, $D(\infty) \simeq q(\infty)/\mu$, (27) can be also written as

$$\text{Prob}[\ D(\infty) \geq D_{max}\ ] \simeq e^{-u\mu D_{max}}, \tag{28}$$

where $D_{max} = Q_{max}/\mu$. From (28) and (3), let $e^{-u\mu D_{max}} = \epsilon$, the statistical QoS requirement can be satisfied. And $u = -\log(\epsilon)/D_{max}\mu$.

E. Peak Power Constraint

When there is a peak power constraint P_{peak}, the transmission power is

$$\tilde{P}(n) = \min\left[P_{peak}, \min[f(g(n)), \frac{2^{\frac{q(n)+\mu}{BT_s}} - 1}{g(n)}]\right]. \tag{29}$$

Define a combined CGB PC scheme $f_{peak}(g(n)) = \min[P_{peak}, f(g(n))]$, $\tilde{P}(n)$ can be rewritten as

$$\tilde{P}(n) = \min\left[f_{peak}(g(n)), \frac{2^{\frac{q(n)+\mu}{BT_s}} - 1}{g(n)}\right]. \tag{30}$$

The conclusion and result in subsection III-B and III-C can be directly applied to this situation, provided that the transition probability matrix $\mathbf{P}$ is re-calculated according to $f_{peak}(g(n))$.

The peak power constraint limits the QoS providing capability of the system. Define the achievable D_{max} region as the set of D_{max} that a certain PC scheme can support under the peak power constraint, for a fixed μ and ϵ.

1) The lower bound of D_{max}: Obviously, the achievable D_{max} has a minimum value. For HQLA/CONST (We use the notation "HQLA/X" to describe the specific PC scheme, where X represents the CGB PC), it is obtained when $f(g(n)) = \bar{P} = P_{peak}$. For HQLA/TDWF, it is obtained when $C \to 0$. Notice that when $C \to 0$, $\tilde{P}(n) = f_{peak}(g(n)) = P_{peak}$ except for $g(n) = 0$. Hence $\tilde{P}(n)g(n) = P_{peak}g(n)$ for all $g(n)$. HQLA/TDWF is equivalent to HQLA/COSNT. They have the same queue length distribution and the average power. To obtain the lower bound of D_{max}, first calculate the queue length distribution of the PC scheme mentioned above. Then find the queue length which has the violation probability nearest to ϵ. The lower bound of D_{max} equals this queue length divided by μ.

2) The lower bound of average power: The average power decreases as D_{max} increases. The lower bound of average power is achieved as $D_{max} \to \infty$. When the average power is smaller than the lower bound, the queue will be unstable. We can obtain this lower bound by the effective capacity at $D_{max} \to \infty$.

Notice that the HQLA PC scheme with CGB PC part $f(g(n))$ actually has the same queue length distribution as $f(g(n))$. Therefore the effective capacity calculated by (26) is also true for HQLA PC scheme with the same $f(g(n))$. For i.i.d. channel gain,

$$\begin{aligned}\alpha(u) &= -\frac{1}{u}\lim_{t\to\infty}\frac{1}{t}\log \mathbf{E}[e^{-u\sum_{i=0}^{t}s(i)}] \\ &= -\frac{1}{u}\lim_{t\to\infty}\frac{1}{t}\log \mathbf{E}^t[e^{-us(1)}] \\ &= -\frac{1}{u}\log \mathbf{E}[e^{-us(1)}]. \end{aligned} \tag{31}$$

For fixed μ and ϵ, as $D_{max} \to \infty$, $u \to 0$,

$$\begin{aligned}\lim_{u\to 0}\alpha(u) &= -\frac{1}{u}\log\int_0^\infty e^{-uBT_s\log_2(1+f(g)g)}f_{ch}(g)dg \\ &= \int_0^\infty BT_s\log_2(1+f(g)g)f_{ch}(g)dg \\ &= \mathbf{E}_g[BT_s\log_2(1+f(g)g)]. \end{aligned} \tag{32}$$

Solve for

$$\mathbf{E}_g[BT_s\log_2(1+f(g)g)] = \mu \tag{33}$$

get the parameter of the CGB PC scheme, i.e. $\bar{P}$ for CONST and C for TDWF, for the critical status. If a smaller $\bar{P}$ or larger C is used, the queue will be unstable, and all the states are transient. At the critical status, the queue is recurrent null, the mean recurrence time is ∞ and there is no stationary distribution. Therefore as time goes by, the probability that the queue length return to zero is zero. Under this situation, the HQLA PC scheme is equivalent to the CGB PC scheme because the probability that the queue can be cleared is zero. Therefore $\mathbf{E}_{g,q}[\tilde{P}(g,q)] = \mathbf{E}_g[f(g)]$. The lower bound of average power is

$$\mathbf{E}_{g,q}[\tilde{P}(g,q)] = \mathbf{E}_g(f(g)) = \int_0^\infty f_{ch}(g)f(g)dg \tag{34}$$

where the parameter of $f(g)$ is obtained by (33).

IV. Simulation Results

A. Steady State Queue Length Distribution

In this subsection, we demonstrate the steady state queue length distributions obtained from both the analytical matrix-geometric approach and the computer simulation. Fig. 4 and Fig. 5 show the probability mass function for HQLA/CONST and HQLA/TDWF respectively. For all the simulations, $\lambda = 1$, $BT_s = 100$, $\mu = 50$ and $\bar{P} = 1$. The computer simulation has 10^6 runs. The analytical results match the simulation results very well. The queue length distribution behaves dramatically different for different CGB PC components. The distribution of HQLA/CONST decays smoothly and exponentially when the queue length is large. The distribution of HQLA/TDWF has peaks at queue length equal to $K\mu$, $K = 0, 2, \ldots$. And the amplitude of the peak decreases as K increases. This is because for HQLA/TDWF, when the channel gain is smaller than the cutoff threshold, the transmitter stops transmitting. When K such slots successively occur, the increment of the queue length will be $K\mu$. The probability decreases exponentially with K. Fig. 6 shows DBVP of HQLA/TDWF with the same simulation parameters. For large D_{max}, the violation probability is approximately exponentially distributed, which validates the applicability of effective capacity theory to the power controlled physical layer.

Since the analytical result matches the simulation result very well, in the following subsections, only the analytical (numerical) results will be shown.

B. Effective Capacity

Secondly we show the effective capacity improvements of the proposed PC scheme. In the simulation, HQLA/OPT uses the PC scheme proposed by Tang and Zhang [10] as the CGB part, which maximizes the effective capacity among all the CGB PC scheme. Fig. 7, Fig. 8 and Fig. 9 illustrate the effective capacity of the HQLA/CONST, HQLA/TDWF and HQLA/OPT respectively. In Fig. 7 and Fig. 8, the effective capacity of OPT is also illustrated as a reference. In the simulation, $\lambda = 1$, $BT_s = 100$, the average power $\mathbf{E}(\tilde{P}) = 1$. For the same average power, the effective capacity of HQLA/TDWF and HQLA/CONST PC scheme is significantly increased comparing to the corresponding channel gain based PC scheme. HQLA/OPT boosts the effective capacity for moderate u. For large u, all of the three HQLA PC schemes approach the OPT PC scheme. HQLA/TDWF and HQLA/OPT perform almost the same and are both superior to OPT.

C. Power Gain in 3G Environment

Fig. 10-12 shows the power gain in a typical 3G WCDMA environment [17]. In the simulation, "TCI" denotes the truncated channel inversion PC [1]. The performance gain is defined as the ratio of the average power required by the CGB PC scheme and that of the proposed HQLA PC scheme to fulfil the 3G QoS requirement,

$$G = \frac{\mathbf{E}_g(f(g))}{\mathbf{E}_{g,q}(\tilde{P}(g,q))}. \tag{35}$$

Two types of services are considered in the simulation, voice data and video data. The simulation parameters are listed as in Table V.

We simulate three typical moving speeds, 3 mph walking speed, 35 mph local driving speed and 70 mph highway speed. The movement of the terminal causes Doppler frequency shift f_m. In time domain, the coherent time is $T_c = \sqrt{9/16\pi}/f_m$. As an approximation, we assume that the channel gain within the coherent time T_c is highly correlated and out of the coherent time is uncorrelated. The transmitter adapts the transmission power every T_c sec. Therefore the slot time $T_s = T_c$ and $\mu = \lceil \tilde{\mu} T_s \rceil$ where $\lceil x \rceil$ finds the smallest integer that is not smaller than x. All of the three schemes have a power gain greater than one (i.e., 0 dB). The power gain for slow speed is much more significant than high speed. Voice data has a larger gain than video data. The power gain for HQLA/TDWF and HQLA/TCI is much higher than HQLA/CONST.

D. HQLA with Adaptive Modulation

In practice, $s(n)$ cannot take arbitrary value. And the idea Shannon's capacity is not achievable during a small slot duration. The power decision process must consider the specific modulation scheme and the target BER. For MQAM modulation, each symbol bears $M = 2^k$ bits, $k = 1, 2, 3...$ When the BER constraint is given, the minimum SINR required to achieve the target BER can be obtained by the BER performance of each modulation [18]. Denote α_k the minimum SINR of modulation scheme k, and B_k the number of bits that can be transmitted during one slot if modulation scheme k is applied. For convenient, let $\alpha_0 = B_0 = 0$, and $\alpha_{K+1} = B_{K+1} = \infty$ where K is the maximum modulation scheme. For CGB PC, the modulation scheme t_c is chose such that $\alpha_{t_c} \leq f(g(n))g(n) < \alpha_{t_c+1}$. For QLA PC, the modulation scheme t_p is chosen such that $B_{t_p-1} < q(n) + \mu \leq B_{t_p}$. The actual modulation scheme is $t = \min(t_c, t_p)$ and the transmission power is $\alpha_t / g(n)$.

Fig. 13 shows the power gain of HQLA/TCI over TCI PC for voice data transmission. The parameters are listed in Table V

The factor 1.1 in μ addresses the signaling overhead. $f(g(n)) = \min[\alpha_6/g(n), P_{max}]$ where P_{max} is the peak power constraint. The delay constraint is fulfilled by turning P_{max}. The receiver estimates and channel gain and sends an channel gain adjustment indicator back to the receiver which takes value of ± 1 dB. The transmitter updates the channel gain by the indicator at each slot and determines the modulation scheme. The power gain is significant at low speed and approaches unit at high speed. That is because at high speed, the coherent time becomes shorter, and the effective capacity increases. Therefore less P_{max} is needed to fulfill the QoS requirement. Accordingly, t_c is smaller and the possibility that t_p is smaller than t_c decreases.

E. Peak Power Constraint

The achievable D_{max} region and its corresponding average power for HQLA/TDWF and HQLA/CONST are shown in Fig. 14. In the simulation, $\lambda = 1$, $BT_s = 200$, $\mu = 120$, $P_{peak} = 1$, $\epsilon = 10^{-3}$. As indicated in subsection III-E, HQLA/TDWF and HQLA/CONST have the same

average power at the minimum achievable D_{max}, the two curves joint at the leftmost point. As $D_{max} \to \infty$, the average power approaches to the lower bound indicated by (34).

V. Conclusion

We study the problem of efficient PC to maximize the delay-throughput performance over wireless networks. A suboptimal yet practically usable HQLA PC scheme is proposed. The proposed HQLA PC scheme adapts the transmission power not only according to the instantaneous channel gain but also to the queue length in the buffer. This strategy can be applied to any CGB PC scheme such as the time-domain water filling. The transmission delay is analytically analyzed and the numerical solution is obtained by the matrix-geometric method. The proposed HQLA PC scheme increases the effective capacity. The effective capacity of HQLA/TDWF is much higher than the maximum effective capacity achieved by the optimal channel gain based PC scheme. We compared the average power of CONST and TDWF with and without HQLA PC scheme. Both the simulation results and the numerical results suggest a great average power saving under the typical 3G environment.

References

[1] A. J. Goldsmith and P. P. Varaiya, "Capacity of fading channels with channel side information," *IEEE Transactions on Information Theory*, vol. 43, no. 6, pp. 1986–1992, Nov. 1997.

[2] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2619–2692, Oct. 1998.

[3] T. Cover, "Broadcast channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 2–14, Jan 1972.

[4] S. Hanly and D. Tse, "Multiaccess fading channels. ii. delay-limited capacities," *IEEE Transactions on Information Theory*, vol. 44, no. 7, pp. 2816–2831, Nov 1998.

[5] R. Negi and J. M. Cioffi, "Delay-constrained capacity with causal feedback," *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2478–2494, Sept. 2002.

[6] X. Liu and A. J. Goldsmith, "Optimal power allocation over fading channels with stringent delay constraints," in *Proc. IEEE ICC 2002*, vol. 3, 28 April–2 May 2002, pp. 1413–1418.

[7] G. Caire, G. Taricco, and E. Biglieri, "Optimum power control over fading channels," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1468–1489, July 1999.

[8] R. Berry and E. Yeh, "Cross-layer wireless resource allocation - fundamental performance limits for wireless fading channels," *IEEE Signal Processing Magazine, special issue on "Signal Processing for Networking,"*, vol. 21, no. 5, pp. 59–68, September 2004.

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

[10] J. Tang and X. Zhang, "Quality-of-service driven power and rate adaptation over wireless links," *IEEE Transactions on Wireless Communications*, vol. 6, no. 7, 2007.

[11] B. Prabhakar, E. Uysal Biyikoglu, and A. El Gamal, "Energy-efficient transmission over a wireless link via lazy packet scheduling," in *Proc. IEEE INFOCOM 2001*, vol. 1, Anchorage, AK, Apr. 2001, pp. 386–394.

[12] M. Neuts, *Matrix-Geometric solutions in stochastic models, an algorithmic approach*. Johns Hopkins University Press, 1981.

[13] L. Kleinrock, *Queueing Systems, Volume I: Theory*. John Wiley & Sons, 1975.

[14] M. Schwartz, *Broadband integrated networks*. Prentice Hall, 1996.

[15] C.-S. Chang, *Performance guarantees in communication networks*. Springer, 2000.

[16] C.-S. Chang and J. Thomas, "Effective bandwidth in high-speed digital networks," *IEEE Journal on Selected Areas in Communications*, Aug. 1995.

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

[18] A. J. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

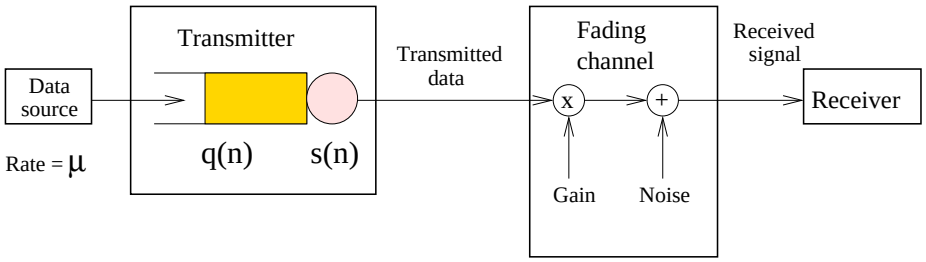

Fig. 1: Structure of data source and transmitter

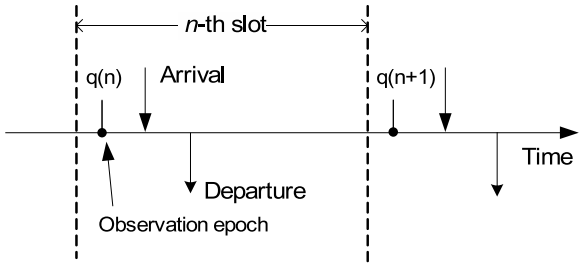

Fig. 2: Update of queue length

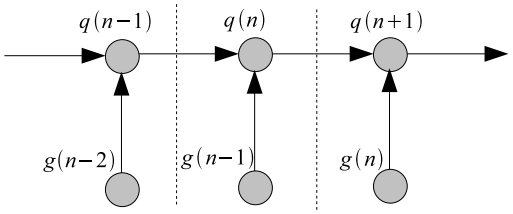

Fig. 3: Markov property of queue length

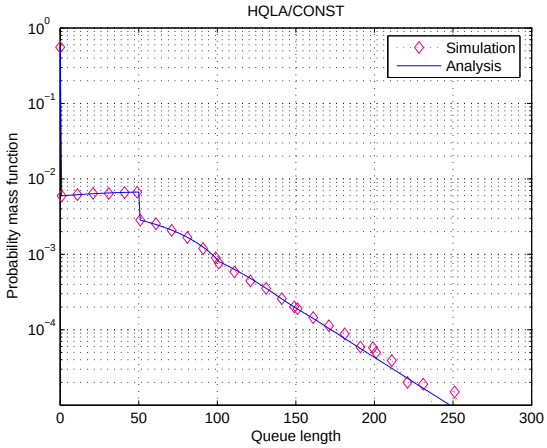

Fig. 4: Probability mass function of HQLA/CONST

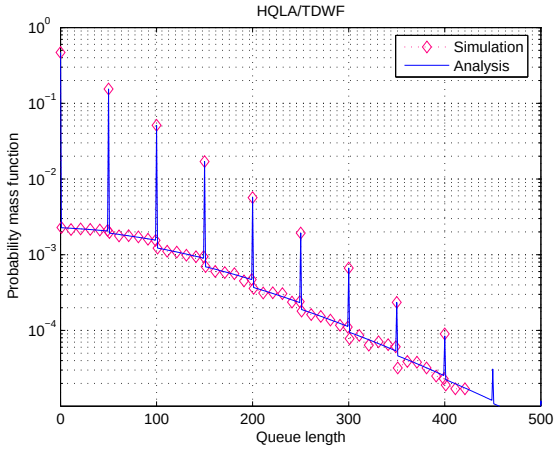

Fig. 5: Probability mass function of HQLA/TDWF

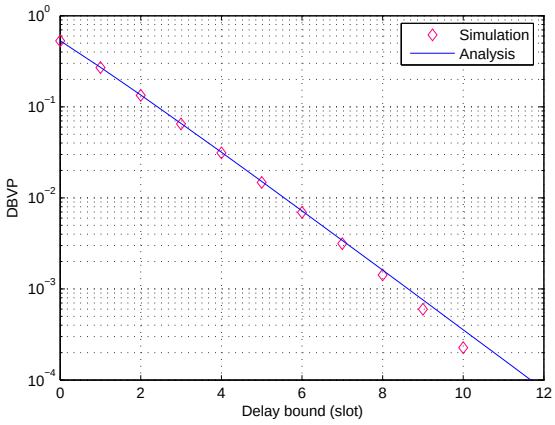

Fig. 6: DBVP of HQLA/TDWF

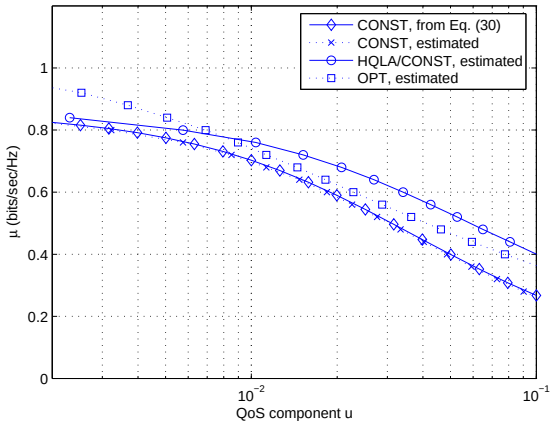

Fig. 7: Effective capacity of HQLA/CONST

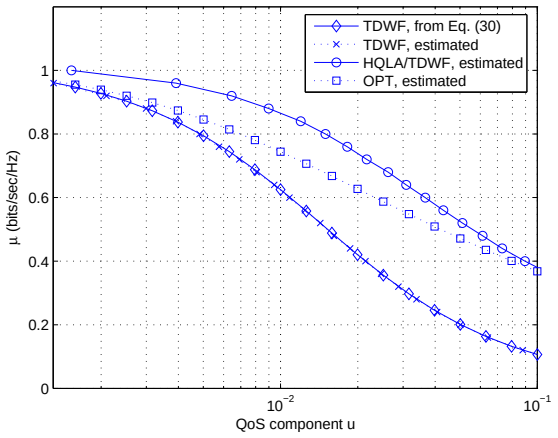

Fig. 8: Effective capacity of HQLA/TDWF

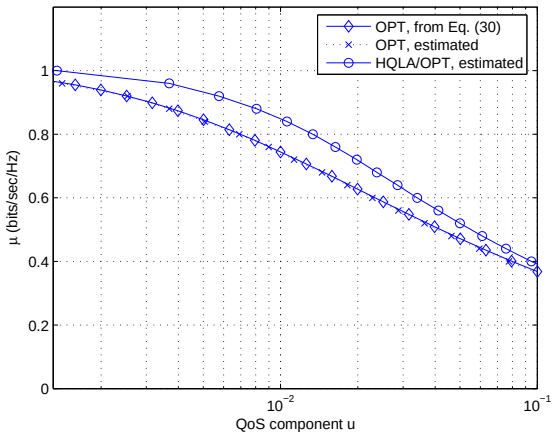

Fig. 9: Effective capacity of HQLA/OPT

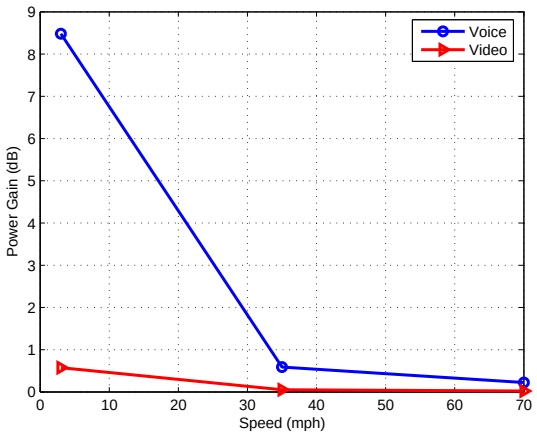

Fig. 10: Power gain of HQLA/CONST over CONST PC

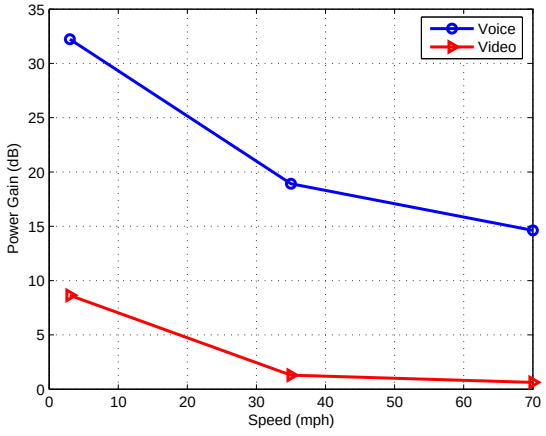

Fig. 11: Power gain of HQLA/TDWF over TDWF PC

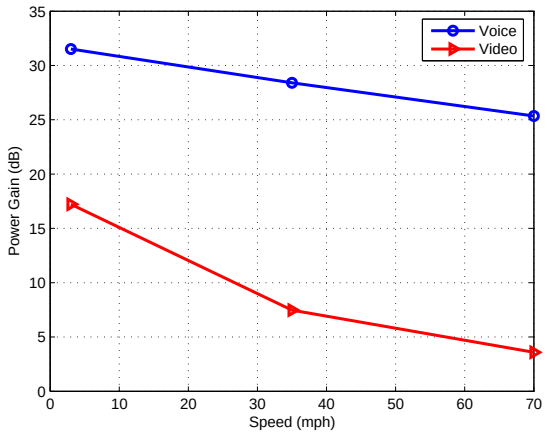

Fig. 12: Power gain of HQLA/TCI over TCI PC

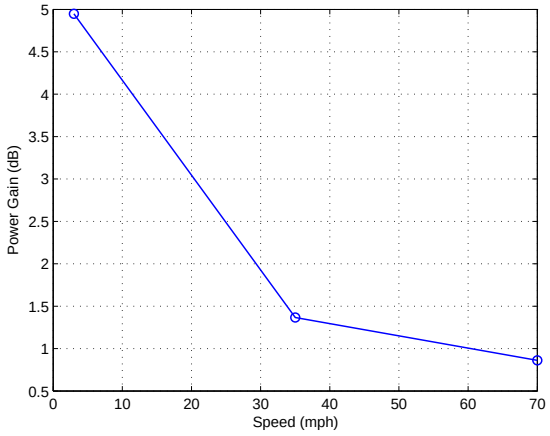

Fig. 13: Power gain of HQLA/TCI over TCI PC with adaptive modulation, voice data

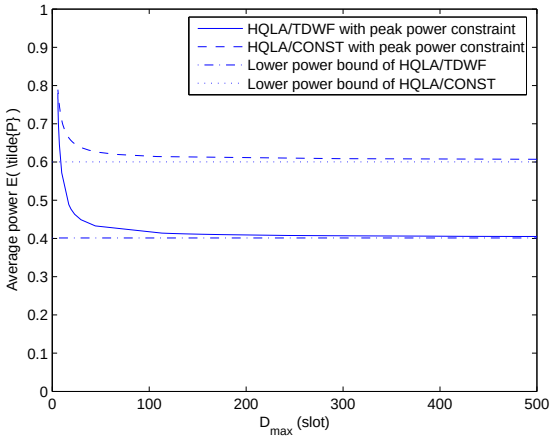

Fig. 14: D_{max} vs. average power with fixed μ and ϵ

BER	bit error rate
CGB	power-gain-based
CONST	constant power control
CQ	clear queue
DBVP	delay bound violation probability
HQLA	hierarchical queue-length-aware
OPT	optimal channel-gain-based power control
PC	power control
QLA	queue length aware
QoS	quality of service
SNR	signal to noise ratio
TCI	truncated channel inversion power control
TDWF	time domain water filling power control

TABLE I: Abbreviations

	$\widetilde{\mu}$	D_{max}	ϵ
Voice	12.2Kbps	50msec	10^{-3}
Data	144Kbps	50msec	10^{-3}
Bandwidth	3.84MHz		
Carrier Frequency	1.9GHz		

TABLE II: Parameters for 3G environment simulation

T_s	2/3msec
Spreading Factor	64
Coding Rate	1/3
K	6
B_k	$2560k/SF$
μ	$\lceil 1.1\tilde{\mu}R_cT_s \rceil$

TABLE III: Parameters for 3G environment simulation with adaptive modulation