Queue Proportional Scheduling via Geometric Programming in Fading Broadcast Channels

Kibeom Seong, *Student Member, IEEE*, Ravi Narasimhan, *Senior Member, IEEE*, and John M. Cioffi, *Fellow, IEEE*

***Abstract*—For fading broadcast channels (BC), a throughput optimal scheduling policy called queue proportional scheduling (QPS) is presented via geometric programming (GP). QPS finds a data rate vector such that the expected rate vector over all fading states is proportional to the current queue state vector and is on the boundary of the ergodic capacity region of a fading BC. Utilizing the degradedness of BC for each fading state, QPS is formulated as a geometric program that can be solved with efficient algorithms. The GP formulation of QPS is also extended to orthogonal frequency-division multiplexing (OFDM) systems in a fading BC. The throughput optimality of QPS is proved, and it is shown that QPS can arbitrarily scale the ratio of each user's average queueing delay. Throughput, delay, and fairness properties of QPS are numerically evaluated in a fading BC and compared with other scheduling policies such as the well-known maximum weight matching scheduling (MWMS). Simulation results for Poisson packet arrivals and exponentially distributed packet lengths demonstrate that compared with MWMS, QPS provides a significant decrease in average queueing delay and has more desirable fairness properties.**

***Index Terms*—Broadcast channels (BC), channel capacity, convex optimization, cross-layer resource allocation, fairness, geometric programming (GP), orthogonal frequency-division multiplexing (OFDM), queueing analysis, queueing delay, scheduling.**

I. Introduction

OPTIMAL allocation of communication resources, such as the transmit power and data rate, is a central problem in multiuser communication systems. With perfect channel state information (CSI) at both the transmitter and receivers, each user's transmit power and rate can be determined based on the channel capacity region. This information-theoretic approach to resource allocation, which ignores the randomness in packet arrivals and queueing, cannot guarantee stability of queueing systems. In [1], the network capacity region is defined as a set of all packet arrival rate vectors for which it is possible to keep every queue length finite. For bursty input traffic, it is generally difficult to estimate the packet arrival rates. Thus, resource allocation solely based on CSI is unable to update rate allocation properly according to the dynamics of the input traffic. As a result, even for a packet arrival rate vector within the network capacity region, some users' queue backlogs may become unacceptably large, causing long queueing delay as well as buffer overflow.

To account for queueing parameters, a cross-layer approach to resource allocation has been recently proposed in [2]–[5] and the references therein. These works show that consideration of both CSI and queue state information (QSI) allows the entire network capacity region to be achieved in fading broadcast and multiple-access channels (MACs). A scheduling policy that achieves the network capacity region is called *throughput optimal*. One well-known throughput optimal scheduling algorithm is maximum weight matching scheduling (MWMS) that maximizes the inner product of the queue state vector and the achievable rate vector [6], [7]. This MWMS policy is proved to be throughput optimal for both fading broadcast channels (BC) and MAC [1], [2]. Recent applications of MWMS can be also found in OFDM downlink systems [8] and MIMO downlink systems [4], [5]. For the fading MAC, [9] shows that MWMS actually minimizes the average queueing delay over all users if symmetric channels and equal packet arrival rates are assumed. This property is a consequence of the polymatroidal structure of the MAC capacity region [10]. However, there are no such structural properties in the fading BC capacity region so that even with symmetry assumptions, MWMS cannot guarantee the minimum average queueing delay.

This paper presents another throughput optimal scheduling policy called queue proportional scheduling (QPS), which has more desirable delay and fairness properties than MWMS in a fading BC. Given the current queue state, QPS allocates a data rate vector such that the expected rate vector averaged over all fading states is proportional to the current queue state vector and is on the boundary of the ergodic capacity region of a fading BC. Utilizing the degradedness of BC for each fading state, QPS is formulated via geometric programming (GP) [11]. GP is a special case of convex optimization problems for which very efficient interior point methods have been developed [12]. Also, the GP formulation of QPS is extended to orthogonal frequency-division multiplexing (OFDM) which has received much attention as a promising modulation technique for next-generation wireless communication systems supporting high data rate services.

Reference [13] introduced the minimum draining time (MDT) policy, which was shown to be throughput optimal and shown to minimize the draining time of the queue backlogs in a fluid model with no further arrivals. Our work was performed independent of [13], and it turns out that QPS has the properties of the MDT policy. We present another approach for proving the throughput optimality of QPS, which is different from [13]. Also, using the new proof, QPS is shown to have the capability of arbitrarily scaling the ratio of each user's average queueing delay. This fairness property of QPS is desirable for satisfying

Manuscript received September 15, 2005; revised April 15, 2006. This work was supported in part by a Stanford Graduate Fellowship. This paper was presented in part at the IEEE International Conference on Communications, Istanbul, Turkey, June 2006, and in part at the IEEE Vehicular Technology Conference, Spring 2006, Melbourne, Australia, May 2006.

K. Seong and J. M. Cioffi are with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: kseong@stanford.edu; cioffi@stanford.edu).

R. Narasimhan is with the Department of Electrical Engineering, University of California, Santa Cruz, CA 95064 USA (e-mail: ravi@soe.ucsc.edu).

Digital Object Identifier 10.1109/JSAC.2006.879404

different *quality-of-service* (QoS) requirement of each user [14]. From derived GP formulations, the queueing delay for Poisson packet arrivals and exponentially distributed packet lengths is simulated under various scheduling policies. Numerical results corroborate the throughput optimality of QPS and indicate that QPS provides significantly smaller average queueing delay than MWMS. Moreover, it is observed that with the QPS policy, the fairness in terms of average queueing delay can be guaranteed for any arrival rate vectors. Compared with other scheduling policies, QPS has larger number of variables and constraints, which increases computational complexity. In order to further reduce the complexity of QPS, we present a method to approximate the ergodic capacity region of a fading BC utilizing the hypersphere. This method is shown to make the complexity of QPS comparable to other policies while allowing only a small increase in the average queueing delay.

The organization of this paper is as follows. Section II describes the model of fading broadcast channels and queueing systems. Together with the introduction to GP, the QPS policy is presented and formulated via GP in Section III. Section IV provides GP formulations of three well-known scheduling policies: MWMS, best channel highest possible rate (BCHPR), and longest queue highest possible rate (LQHPR). In Section V, the throughput optimality of QPS is proved, and its fairness and delay properties are investigated. Section VI presents the hypersphere approximation of the ergodic capacity region of a fading BC. Numerical results and discussion are given in Section VII, and finally, Section VIII provides concluding remarks.

Notation: Vectors are bold-faced. $\mathbb{R}^n$ denotes the set of real n-vectors and $\mathbb{R}_+^n$ denotes the set of nonnegative real n-vectors. Given two column vectors $\mathbf{x}$ and $\mathbf{y}$ of length n, $\sum_{i=1}^n x_i y_i$ is expressed as an inner product $\mathbf{x} \cdot \mathbf{y}$. The curled inequality symbol $\succeq$ (and its strict form $\succ$) is used to denote the component-wise inequality between vectors: $\mathbf{x} \succeq \mathbf{y}$ means $x_i \geq y_i$, $i = 1, 2, \cdots, n$. A column vector with all entries being 1 is denoted as $\mathbf{1}$; the length of $\mathbf{1}$ will be clear from context. $\mathbb{E}_x$ denotes expectation over the random variable x.

II. System Model

This section presents the models of fading broadcast channels and queueing systems that are used in this paper. The overall system is summarized in Fig. 1.

A. Fading Broadcast Channels

In this paper, a block fading channel is assumed where the fading state is constant over one scheduling period and each scheduling period undergo es independent and identically distributed (i.i.d.) fading. Also, both the transmitter and receivers are assumed to have perfect knowledge of CSI. It is known that the capacity region of a Gaussian BC can be achieved by utilizing superposition coding at the transmitter in conjunction with successive interference cancellation at each receiver [15]. With this optimal scheme, one user can remove the interference caused by other users' messages encoded earlier. Consider a Gaussian broadcast channel with a single transmitter sending independent messages to K users over two-sided bandwidth $2W$. It is assumed that the transmitter has a peak power constraint of

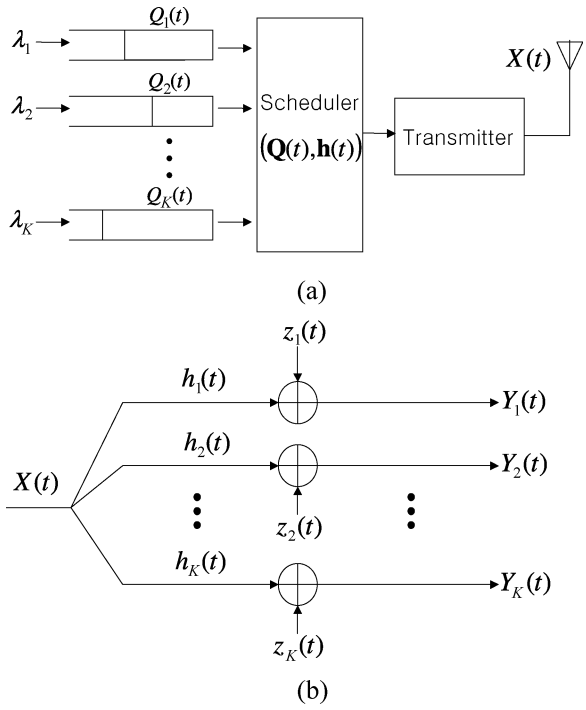

Fig. 1. (a) Block diagram of the queueing system and scheduler. (b) Fading broadcast channel models.

P on each transmission. At time t, the received signal of user i is expressed as

$$Y_i(t) = h_i(t)X(t) + z_i(t), \quad i = 1, \cdots, K \tag{1}$$

where the transmitted signal $X(t)$ is composed of K independent messages, the complex channel gain of user i is denoted by $h_i(t)$ and $z_i(t)$'s are i.i.d. zero-mean Gaussian band-limited noises with power N_0W. As in [16], the channel gain can be combined with the noise component by defining an effective noise $\tilde{z}_i(t) = z_i(t)/h_i(t)$. Then, the equivalent received signal is given by

$$\tilde{Y}_i(t) = X(t) + \tilde{z}_i(t), \quad i = 1, \cdots, K \tag{2}$$

where the power of $\tilde{z}_i(t)$ conditioned on the channel gain is defined as $n_i(t) = N_0W/|h_i(t)|^2$. Without loss of generality, $W = 1$ is assumed throughout this paper for simplicity. The effective noise power $\mathbf{n} = [n_1\ n_2 \cdots n_K]^T$ is used to denote a fading state. The ergodic capacity region of a fading BC is the set of all long-term average rate vectors achievable in a fading BC with arbitrarily small probability of error. A power control policy $\mathcal{P}$ over all possible fading states is defined as a function that maps from any fading state $\mathbf{n}$ to each user's transmit power $P_i(\mathbf{n})$. Let Ω denote the set of all power policies satisfying the sum power constraint P which is given by

$$\Omega = \left\{ \mathcal{P} : \sum_{i=1}^{K} P_i(\mathbf{n}) \leq P, \text{ for all } \mathbf{n} \right\}. \tag{3}$$

For each fading state, the channel is a degraded Gaussian broadcast channel, where the capacity region is achieved by encoding a message of the user with smaller effective noise power later. With this optimal ordering, the capacity of user i for a fading state $\mathbf{n}$ is given by

$$R_i\left(\mathbf{P}(\mathbf{n})\right) = \log_2\left(1 + \frac{P_i(\mathbf{n})}{n_i + \sum_{k=1}^{K} P_k(\mathbf{n}) 1[n_i > n_k]}\right) \quad (4)$$

where $\mathbf{P}(\mathbf{n}) = [P_1(\mathbf{n})\ P_2(\mathbf{n}) \cdots P_K(\mathbf{n})]^T$ and $1[\cdot]$ is the indicator function, which equals 1 if its argument is satisfied; 0 otherwise. Then, the capacity region of a Gaussian BC for the fading state $\mathbf{n}$ and transmit power P is

$$C(\mathbf{n}, P) = \left\{R_i : R_i \le R_i\left(\mathbf{P}(\mathbf{n})\right),\ i = 1, 2, \cdots, K, \text{ where } \sum_i P_i(\mathbf{n}) = P\right\}. \quad (5)$$

Let $C_{\mathrm{BC}}(\mathcal{P})$ denote the set of achievable rates averaged over all fading states for a power policy $\mathcal{P}$

$$C_{\mathrm{BC}}(\mathcal{P}) = \left\{R_i : R_i \le \mathbb{E}_{\mathbf{n}}\left[R_i\left(\mathbf{P}(\mathbf{n})\right)\right],\ i = 1, \cdots, K\right\}. \quad (6)$$

With the sum power constraint P and perfect CSI at the transmitter and receivers, the ergodic capacity region of a fading BC is given by [16]

$$C_{\mathrm{erg}}(P) = \bigcup_{\mathcal{P} \in \Omega} C_{\mathrm{BC}}(\mathcal{P}) \quad (7)$$

where the region $C_{\mathrm{erg}}(P)$ is convex.

B. Queueing Systems

The queueing system and scheduler are modeled by the following: K data sources generate packets according to independent Poisson arrival processes $\{A_i(t), i = 1, \cdots, K\}$, which are stationary counting processes with $\lim_{t\to\infty} A_i(t)/t = a_i < \infty$, and $\mathrm{var}(A_i(t+T) - A_i(t)) < \infty$ for $T < \infty$. Packet lengths in bits $\{B_i\}$ are i.i.d. exponentially distributed with $\mathbb{E}[B_i] = \gamma_i < \infty$ and $\mathbb{E}[B_i^2] < \infty$. We assume packet lengths are independent of packet arrival processes; thus, user i's average arrival rate in bits is given by $\lambda_i = a_i\gamma_i$. The transmitter has K output queues assumed to have infinite capacity. Packets from source i enter queue i and wait until they are served to receiver i. The scheduling period is denoted by T_s, and without loss of generality, we assume $T_s = 1$. At time t, the fading state is represented as $\mathbf{n}(t) = [n_1(t)\ n_2(t) \cdots n_K(t)]^T$, and the queue state vector is $\mathbf{Q}(t) = [Q_1(t)\ Q_2(t) \cdots Q_K(t)]^T$, where $Q_i(t)$ denotes the number of bits waiting to be sent to user i. The allocated rate vector at time t is represented as $\mathbf{R}(\mathbf{n}(t), \mathbf{Q}(t)) = [R_1(\mathbf{n}(t), \mathbf{Q}(t)) \cdots R_K(\mathbf{n}(t), \mathbf{Q}(t))]^T$, which is determined by the scheduler based on both fading and queue states. For simplicity, $\mathbf{R}(t)$ and $\mathbf{R}(\mathbf{n}(t), \mathbf{Q}(t))$ are interchangeably used throughout this paper. $\mathbf{R}(t)$ is achievable only when it is within the capacity region $C(\mathbf{n}(t), P)$ defined in (5).

A time interval $[t, t+1)$ with $t = 0, 1, 2, \cdots$ is denoted by the *time slot* t. It is assumed that the rate allocation is determined at the beginning of each time slot, and it remains unchanged until the new time slot begins. Since $W = 1$ and $T_s = 1$ are assumed, $\mathbf{R}(t)$ for $t = 0, 1, 2, \cdots$ is equivalent to a vector denoting the number of bits supported by each user in the time slot t. Define $Z_i(t)$ as the number of arrived bits at user i's queue in the time slot t. Then, after a scheduling period, user i's queue state vector is equal to $Q_i(t+1) = \max\{Q_i(t) - R_i(t), 0\} + Z_i(t)$. In this paper, each scheduling policy has an explicit constraint of $\mathbf{R}(t) \preceq \mathbf{Q}(t)$; thus, $\max\{\cdot, 0\}$ operation can be simply removed. We adopt the stability definition of queueing systems given in [1]. Therefore, with the overflow function defined by $g(M) = \limsup_{t\to\infty}(1/t)\int_0^t 1[Q_i(\tau) > M]d\tau$, queue i is said to be stable if $g(M) \to 0$ as $M \to \infty$. An arrival rate vector $\boldsymbol{\lambda}$ is stabilizable if there exists a feasible power-and-rate-allocation policy that keeps all queues stable. A set of stabilizable arrival rate vectors forms the network capacity region [1], and a scheduling method that achieves the entire network capacity region is called throughput optimal.

III. Queue Proportional Scheduling via Geometric Programming (GP)

In this section, QPS is introduced and formulated via GP. First, the next section presents brief introduction of GP.

A. Geometric Programming (GP)

GP is a special form of convex optimization problems for which very efficient algorithms have been developed [12], and a variety of GP applications can be found in communication systems [18]. GP uses monomial and posynomial functions. A monomial function has the form of $h_j(\mathbf{x}) = c_j x_1^{a_{j,1}} x_2^{a_{j,2}} \cdots x_n^{a_{j,n}}$, where $\mathbf{x} \succ 0$, $c_j > 0$ and $a_{j,l} \in \mathbb{R}$. A posynomial is a sum of monomials $f_i(\mathbf{x}) = \sum_k c_{ik} x_1^{a_{ik,1}} x_2^{a_{ik,2}} \cdots x_n^{a_{ik,n}}$. Then, GP takes the following form:

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \le 1 \\ & h_j(\mathbf{x}) = 1 \end{aligned} \quad (8)$$

where f_0 and f_i are posynomials and h_j are monomials. Although this is not a convex optimization problem, with a change of variables: $y_i = \log x_i$ and $b_{ik} = \log c_{ik}$, we can convert it into a convex form as the following:

$$\begin{aligned} \text{minimize} \quad & p_0(\mathbf{y}) = \log \Sigma_k \exp\left(\mathbf{a}_{0k}^T \mathbf{y} + b_{0k}\right) \\ \text{subject to} \quad & p_i(\mathbf{y}) = \log \Sigma_k \exp\left(\mathbf{a}_{ik}^T \mathbf{y} + b_{ik}\right) \le 0 \\ & q_j(\mathbf{y}) = \mathbf{a}_j^T \mathbf{y} + b_j = 0. \end{aligned} \quad (9)$$

The solution of this problem can be easily found by using very efficient GP algorithms that are well-developed [18].

B. Queue Proportional Scheduling via GP

At time slot t, MWMS, a well-known throughput optimal scheduling policy, assigns the following data rate vector:

$$\mathbf{R}_{\mathrm{MWMS}}\left(\mathbf{n}(t), \mathbf{Q}(t)\right) = \arg\max_{\mathbf{r}} \mathbf{Q}'(t)^T \mathbf{r} \quad \text{such that } \mathbf{r} \in C\left(\mathbf{n}(t), P\right) \quad (10)$$

where $\mathbf{r} = [r_1\ r_2 \cdots r_K]^T$, $\mathbf{n}(t)$ is the fading state vector at time t, and $\mathbf{Q}'(t) = [\beta_1 Q_1(t) \cdots \beta_K Q_K(t)]^T$. β_i is the user i's priority weight which is set to 1 for all users if everyone has the

same priority. This algorithm tends to allocate higher data rate to the user with longer backlog or better channel conditions. By jointly considering queue and channel states, MWMS is shown to achieve the entire network capacity region [6].

In contrast, the proposed QPS algorithm allocates the following data rate vector at time slot t:

$$\mathbf{R}_{\mathrm{QPS}}(\mathbf{n}(t),\mathbf{Q}(t)) \in C(\mathbf{n}(t),P) \text{ such that}$$
$$\mathbb{E}_{\mathbf{n}(t)}[\mathbf{R}_{\mathrm{QPS}}(\mathbf{n}(t),\mathbf{Q}(t))] = \mathbf{Q}'(t)\left(\max_{\mathbf{Q}'(t)x\in C_{\mathrm{erg}}(P)} x\right) \quad (11)$$

where x is a scalar. Assuming equal priority on each user, $\mathbf{Q}'(t) = \mathbf{Q}(t)$. Then, the average rate vector under the QPS policy, $\mathbb{E}_{\mathbf{n}(t)}[\mathbf{R}_{\mathrm{QPS}}(\mathbf{n}(t),\mathbf{Q}(t))]$ is proportional to the queue state vector and also lies on the boundary surface of the ergodic capacity region. As shown in [17], each boundary point of $C_{\mathrm{erg}}(P)$ in a fading BC is a solution to the optimization problem $\max_{\mathbf{r}} \boldsymbol{\mu}\cdot\mathbf{r}$, where $\mathbf{r} \in C_{\mathrm{erg}}(P)$ for some $\boldsymbol{\mu} \in \mathbb{R}_+^K$. When such $\boldsymbol{\mu}$ is given, $\mathbf{R}_{\mathrm{QPS}}(\mathbf{n}(t),\mathbf{Q}(t))$ is a solution to the optimization problem $\max_{\mathbf{r}} \boldsymbol{\mu}\cdot\mathbf{r}$, where $\mathbf{r} \in C(\mathbf{n}(t),P)$ for any fading state $\mathbf{n}(t)$. Therefore, the data rate vector assigned by QPS at time slot t can be expressed as

$$\mathbf{R}_{\mathrm{QPS}}(\mathbf{n}(t),\mathbf{Q}(t)) = \arg\max_{\mathbf{r}} \boldsymbol{\mu}^T\mathbf{r} \quad \text{such that } \mathbf{r}\in C(\mathbf{n}(t),P). \quad (12)$$

Under the QPS policy, $\boldsymbol{\mu}$ is determined based on the current queue state vector, as well as the ergodic capacity region of a fading BC. However, as shown in (10), MWMS only considers the queue state vector in deriving the weight vector.

Fig. 2 illustrates two distinct expected rate vectors supported by MWMS and QPS for the queue state vector $\mathbf{Q}(t)$. A two user Rayleigh-fading BC is considered, where $P = 2$, user 1's average signal-to-noise ratio (SNR) is 13 dB and user 2's average SNR is 7 dB. Each user's average SNR is defined as the average received SNR when the total transmit power is allocated to that user. Since $W = T_s = 1$ is assumed, bits per second per Hertz (bps/Hz) is equivalent to bits/scheduling period. Thus, the ergodic capacity region in Fig. 2 represents the set of vectors denoting each user's expected number of bits served in one scheduling period. Also, note that with $W = T_s = 1$, the network capacity region is the same as the ergodic capacity region. From Fig. 2, as the queue state changes, MWMS is expected to exhibit more variations in the average rate vector compared with QPS. According to queueing theory, lower variance in service rate or arrival rate provides smaller queueing delay [19]. Therefore, QPS is expected to have smaller average queueing delay than MWMS, as demonstrated in Section VII.

By utilizing the degradedness of BC for each fading state, the rate allocation of QPS can be formulated via GP. Assume that the M most recent fading states are sampled, which are denoted by $\{\mathbf{n}^{(1)},\cdots,\mathbf{n}^{(M)}\}$. To reduce the correlation among samples, the sampling period needs to be determined in consideration of fading coherence time. In this paper, the sampling period is simply assumed equal to one scheduling period because of i.i.d. block fading over each scheduling time. Without loss of generality, $\mathbf{n}^{(M)}$ is assumed to denote the current fading state $\mathbf{n}(t)$. Then, consider a family of M parallel Gaussian broadcast

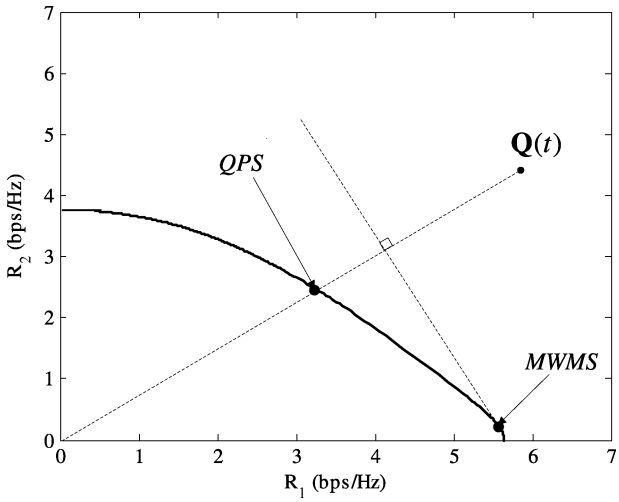

Fig. 2. Ergodic capacity region of a two user Rayleigh-fading BC, and expected rate vectors of QPS and MWMS when the queue state vector is $\mathbf{Q}(t)$ ($P = 2$, user 1's average $\mathrm{SNR} = 13$ dB and user 2's average $\mathrm{SNR} = 7$ dB).

channels, such that in the mth component channel, user i has effective noise variance $n_i^{(m)}$, rate $R_i^{(m)}$, and power $P_i^{(m)}$. Note that each BC has a power constraint of P. At time slot t, QPS allocates the data rate vector $\mathbf{R}_{\mathrm{QPS}}(\mathbf{n}^{(M)},\mathbf{Q}(t))$ that is a solution of the following optimization problem:

$$\frac{1}{M}\sum_{m=1}^{M}\mathbf{R}_{\mathrm{QPS}}\left(\mathbf{n}^{(m)},\mathbf{Q}(t)\right) = \mathbf{Q}(t)\left(\max_{\mathbf{Q}(t)x\in C_{\mathrm{erg}}(P)} x\right)$$
$$\mathbf{R}_{\mathrm{QPS}}\left(\mathbf{n}^{(m)},\mathbf{Q}(t)\right) \in C\left(\mathbf{n}^{(m)},P\right) \text{ for all } m. \quad (13)$$

From (4) and (5), the capacity region of the mth Gaussian BC is given by

$$C\left(\mathbf{n}^{(m)},P\right) = \left\{ R_{\pi_m(i)}^{(m)} : R_{\pi_m(i)}^{(m)} \le \log_2\left(1+\frac{\alpha_{\pi_m(i)}^{(m)}P}{n_{\pi_m(i)}^{(m)}+\sum_{j<i}\alpha_{\pi_m(j)}^{(m)}P}\right), i=1,2,\cdots,K \text{ where } \sum_i \alpha_{\pi_m(i)}^{(m)} = 1 \right\} \quad (14)$$

where $\pi_m(\cdot)$ is the permutation such that $n_{\pi_m(1)}^{(m)} < n_{\pi_m(2)}^{(m)} < \cdots < n_{\pi_m(K)}^{(m)}$, and $\alpha_{\pi_m(i)}^{(m)}$ is the fraction of the total transmit power used for user $\pi_m(i)$'s signal in the mth Gaussian BC. When $\mathbf{R}^{(m)}$ is on the boundary of the capacity region, solving the $\alpha_{\pi_m(i)}^{(m)}$'s in terms of the rate vector $\mathbf{R}^{(m)}$ yields the following equations:

$$\sum_{i=1}^{l}\alpha_{\pi_m(i)}^{(m)}P = \sum_{i=1}^{l}\left(n_{\pi_m(i)}^{(m)} - n_{\pi_m(i-1)}^{(m)}\right) \times \exp\left(\log 2\sum_{j=i}^{l}R_{\pi_m(j)}^{(m)}\right) - n_{\pi_m(l)}^{(m)}, \quad l = 1,\cdots,K \quad (15)$$

where $n^{(m)}_{\pi_m(0)} \equiv 0$. As shown in [16], (14) is equivalent to

$$C\left(\mathbf{n}^{(m)}, P\right) = \left\{ R^{(m)}_{\pi_m(i)} : \sum_{i=1}^{K} \left(n^{(m)}_{\pi_m(i)} - n^{(m)}_{\pi_m(i-1)}\right) \times \exp\left(\log 2 \sum_{j=i}^{K} R^{(m)}_{\pi_m(j)}\right) - n^{(m)}_{\pi_m(K)} \le P \text{ and } R^{(m)}_{\pi_m(i)} \ge 0,\ i = 1, 2, \cdots, K \right\}. \quad (16)$$

Using this relation, (13) can be converted into

$$\begin{aligned} \text{minimize} \quad & \log\left(\exp(-x)\right) \\ \text{subject to} \quad & \log\left(\exp\left(-R^{(m)}_i\right)\right) \le 0, \quad \forall\, i \text{ and } m \\ & \log\left(\exp\left(-Q_i(t)\right)\exp\left(R^{(M)}_i\right)\right) \le 0, \quad \forall\, i \\ & \log \sum_{i=1}^{K} \left(\frac{n^{(m)}_{\pi_m(i)} - n^{(m)}_{\pi_m(i-1)}}{P + n^{(m)}_{\pi_m(K)}}\right) \\ & \times \exp\left(\log 2 \sum_{j=i}^{K} R^{(m)}_{\pi_m(j)}\right) \le 0, \quad \forall\, m \\ & \mathbf{Q}(t)x - \frac{1}{M}\sum_{m=1}^{M} \mathbf{R}^{(m)} = 0 \end{aligned} \quad (17)$$

where the second set of constraints is added to avoid allocating redundant power to some users with short queue lengths. If the optimization variable is defined as $\mathbf{y} = [x\ (\mathbf{R}^{(1)})^T \cdots (\mathbf{R}^{(M)})^T]^T \in \mathbb{R}_+^{(KM+1)}$, (17) is a standard geometric program with the globally optimal solution $\mathbf{y}^* = [x^*\ (\mathbf{R}^{*(1)})^T \cdots (\mathbf{R}^{*(M)})^T]^T$. Then, the data rate vector supported under the QPS policy is $\mathbf{R}_{\mathrm{QPS}}(\mathbf{n}^{(M)}, \mathbf{Q}(t)) = \mathbf{R}^{*(M)}$, and the corresponding power allocation can be obtained by solving (15) for $m = M$. This GP formulation of QPS can be extended to OFDM systems, as discussed in the next subsection.

C. Extension to OFDM Systems

In a fading BC with intersymbol interference (ISI), the ISI can be completely removed by exploiting OFDM techniques with sufficient number of tones, i.e., the frequency response can be made flat within each tone. Consider OFDM systems with K users and L tones. On each tone, the channel is equivalent to a fading BC without ISI, which becomes a degraded Gaussian BC for the fixed fading state. Therefore, by extending the results from Section III-B, QPS for OFDM systems in a fading BC can be also converted into GP. At tone l, M sampled fading state vectors are denoted by $\{\mathbf{n}^{(l,1)}, \cdots, \mathbf{n}^{(l,M)}\}$, where $\mathbf{n}^{(l,m)} = [n_1^{(l,m)} \cdots n_K^{(l,m)}]^T$. For the mth sampled fading state, $n_i^{(l,m)}$, $R_i^{(l,m)}$, and $P_i^{(l,m)}$ denote the effective noise variance, rate, and power on user i's tone l, respectively. Without loss of generality, the Mth sample is assumed to denote the current fading state. Also, a total power constraint of P is imposed on each transmission of OFDM symbols. Define $\pi_{l,m}(\cdot)$ as the permutation such that $n^{(l,m)}_{\pi_{l,m}(1)} < n^{(l,m)}_{\pi_{l,m}(2)} < \cdots < n^{(l,m)}_{\pi_{l,m}(K)}$. By carefully applying above updates to (16) and (17), QPS in OFDM systems can be converted into the following GP:

$$\begin{aligned} \text{minimize} \quad & \log(\exp(-x)) \\ \text{subject to} \quad & \log\left(\exp\left(-R_i^{(l,m)}\right)\right) \le 0, \quad \forall\, i,\ l, \text{ and } m \\ & \log\left(\exp\left(-Q_i(t)\right)\exp\left(\sum_{l=1}^{L} R_i^{(l,M)}\right)\right) \le 0, \ \forall i \\ & \log \sum_{l=1}^{L}\sum_{i=1}^{K} \left(\frac{n^{(l,m)}_{\pi_{l,m}(i)} - n^{(l,m)}_{\pi_{l,m}(i-1)}}{P + \sum_{s=1}^{L} n^{(s,m)}_{\pi_{s,m}(K)}}\right) \\ & \times \exp\left(\log 2 \sum_{j=i}^{K} R^{(l,m)}_{\pi_{l,m}(j)}\right) \le 0, \quad \forall\, m \\ & \mathbf{Q}(t)x - \frac{1}{M}\sum_{m=1}^{M}\sum_{l=1}^{L} \mathbf{R}^{(l,m)} = 0 \end{aligned} \quad (18)$$

where $n^{(l,m)}_{\pi_{l,m}(0)} \equiv 0$. If the optimization variable is defined as $\mathbf{y} = [x\ (\mathbf{R}^{(1,1)})^T \cdots (\mathbf{R}^{(L,M)})^T]^T \in \mathbb{R}_+^{(KLM+1)}$, (18) is a standard geometric program with the globally optimal solution $\mathbf{y}^* = [x^*\ (\mathbf{R}^{*(1,1)})^T \cdots (\mathbf{R}^{*(L,M)})^T]^T$. Consequent rate allocation on tone l under the QPS policy is $\mathbf{R}^{(l)}_{\mathrm{QPS}}(\mathbf{n}^{(l,M)}, \mathbf{Q}(t)) = \mathbf{R}^{*(l,M)}$ for $l = 1, \cdots, L$, and the corresponding power allocation can be obtained by applying (15) on each tone with $m = M$.

IV. Other Scheduling Policies via GP

This section provides GP formulations of three other scheduling methods in a fading BC: MWMS, BCHPR, and LQHPR.

A. Maximum Weight Matching Scheduling (MWMS) via GP

At time slot t, the rate allocation under MWMS can be found by solving (10), which is the weighted sum-rate maximization problem over $C(\mathbf{n}(t), P)$ considering the queue state vector $\mathbf{Q}(t)$ as the weight vector. Utilizing (16), MWMS can be formulated as the following GP:

$$\begin{aligned} \text{minimize} \quad & \log\left(\exp\left(-\mathbf{Q}(t)^T\mathbf{r}\right)\right) \\ \text{subject to} \quad & \log\left(\exp(-r_i)\right) \le 0, \quad \forall\, i \\ & \log\left(\exp\left(-Q_i(t)\right)\exp(r_i)\right) \le 0, \quad \forall\, i \\ & \log \sum_{i=1}^{K} \left(\frac{n_{\pi(i)}(t) - n_{\pi(i-1)}(t)}{P + n_{\pi(K)}(t)}\right) \\ & \times \exp\left(\log 2 \sum_{j=i}^{K} r_{\pi(j)}\right) \le 0 \end{aligned} \quad (19)$$

where $\pi(\cdot)$ is the permutation such that $n_{\pi(1)}(t) < n_{\pi(2)}(t) < \cdots < n_{\pi(K)}(t)$. Let $\mathbf{r}^*$ be the solution of (19), then $\mathbf{R}_{\mathrm{MWMS}}(\mathbf{n}(t), \mathbf{Q}(t)) = \mathbf{r}^*$.

B. Best Channel Highest Possible Rate (BCHPR) via GP

Under the BCHPR policy, a user with the better channel condition takes higher priority in resource allocation. Also, user i is served only if some transmit power remains after clearing queue backlogs of users with higher priorities than user i. This algorithm is equivalent to allocating a data rate vector that minimizes the l_1-norm distance from the current queue state vector. The l_1-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as $||\mathbf{x}||_1 = |x_1| + \cdots + |x_n|$. At time slot t, the BCHPR policy supports the rate vector $\mathbf{R}_{BCHPR}(\mathbf{n}(t), \mathbf{Q}(t))$ that is a solution of the following optimization problem:

$$\min ||\mathbf{Q}(t) - \mathbf{r}||_1 \text{ subject to } \mathbf{r} \in C(\mathbf{n}(t), P). \tag{20}$$

With the constraint of $\mathbf{r} \preceq \mathbf{Q}(t)$, the solution of the above problem is unaffected by $\sum_{i=1}^{K} Q_i(t)$. After removing this summation from the objective, (20) can be converted into the following GP:

$$\begin{aligned} \text{minimize} \quad & \log\left(\exp(-\mathbf{1}^T\mathbf{r})\right) \\ \text{subject to} \quad & \log\left(\exp(-r_i)\right) \le 0, \quad \forall\, i \\ & \log\left(\exp\left(-Q_i(t)\right)\exp(r_i)\right) \le 0, \quad \forall\, i \\ & \log \sum_{i=1}^{K} \left(\frac{n_{\pi(i)}(t) - n_{\pi(i-1)}(t)}{P + n_{\pi(K)}(t)} \right) \\ & \times \exp\left(\log 2 \sum_{j=i}^{K} r_{\pi(j)} \right) \le 0. \end{aligned} \tag{21}$$

Let $\mathbf{r}^*$ be the solution of (21), then $\mathbf{R}_{BCHPR}(\mathbf{n}(t), \mathbf{Q}(t)) = \mathbf{r}^*$. When $\mathbf{Q}(t) \succeq \mathbf{r}$ for any $\mathbf{r} \in C(\mathbf{n}(t), P)$, the BCHPR policy solely depends on channel conditions. At each scheduling time, it allocates total power to the single user with the best channel condition, which is a sum-rate maximizing scheme in a fading BC [20].

C. Longest Queue Highest Possible Rate (LQHPR) via GP

LQHPR schedules a data rate vector such that the longest queue length is minimized, which is equivalent to selecting a rate vector minimizing the l_∞-norm distance from the current queue state vector. The l_∞-norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is defined as $||\mathbf{x}||_\infty = \max\{|x_1|, \cdots, |x_n|\}$. Hence, at time slot t, the LQHPR policy assigns the rate vector $\mathbf{R}_{LQHPR}(\mathbf{n}(t), \mathbf{Q}(t))$ that is a solution of the following optimization problem:

$$\min ||\mathbf{Q}(t) - \mathbf{r}||_\infty \text{ subject to } \mathbf{r} \in C(\mathbf{n}(t), P). \tag{22}$$

Let x denote the upper bound on $||\mathbf{Q}(t) - \mathbf{r}||_\infty$ such that $-x\mathbf{1} \prec \mathbf{Q}(t) - \mathbf{r} \prec x\mathbf{1}$. Then, the above equation can be represented as

$$\begin{aligned} \text{minimize} \quad & \log\left(\exp(x)\right) \\ \text{subject to} \quad & \log\left(\exp(-r_i)\right) \le 0, \quad \forall\, i \\ & \log\left(\exp\left(-Q_i(t)\right)\exp(-x + r_i)\right) \le 0, \quad \forall\, i \\ & \log\left(\exp\left(Q_i(t)\right)\exp(-x - r_i)\right) \le 0, \quad \forall\, i \\ & \log \sum_{i=1}^{K} \left(\frac{n_{\pi(i)}(t) - n_{\pi(i-1)}(t)}{P + n_{\pi(K)}(t)} \right) \\ & \times \exp\left(\log 2 \sum_{j=i}^{K} r_{\pi(j)} \right) \le 0. \end{aligned} \tag{23}$$

Define the optimization variable as $\mathbf{y} = [x\ \mathbf{r}^T]^T$ then, (23) is a standard geometric program with the globally optimal point $\mathbf{y}^* = [x^*\ \mathbf{r}^{*T}]^T$. The data rate vector supported under LQHPR is $\mathbf{R}_{\text{LQHPR}}(\mathbf{n}(t), \mathbf{Q}(t)) = \mathbf{r}^*$.

V. Properties of the QPS Policy in the Fading BC

In this section, QPS is proved to be throughput optimal in a fading BC, and its fairness and delay properties are analyzed.

A. Throughput Optimality of QPS

The next theorem shows the convergence property of the expected queue state vector under QPS, which is crucial in showing throughput optimality and fairness properties.

Theorem 1: Under the QPS policy in a fading BC, as $t \to \infty$, the expected queue state vector conditioned on any initial queue state, converges to a vector proportional to the arrival rate vector.

Proof: Let $\mathbf{q_0} \in \mathbb{R}_+^K$ be the initial queue state vector, and denote the bit arrival rate vector by $\boldsymbol{\lambda} = [\lambda_1 \cdots \lambda_K]^T$, where $\lambda_1 > 0$. Consider time slot t when some queues have backlogs and let $\mathbf{Q}(t)$ be equal to $\mathbf{q_t} = [q_{t,1}\ q_{t,2} \cdots q_{t,K}]^T$. Without loss of generality, assume $q_{t,1} > 0$. Then, $\mathbf{q_t}$ can be represented as $\mathbf{q_t} = w(t)[\lambda_1, \lambda_2 + \Delta\lambda_2, \cdots, \lambda_K + \Delta\lambda_K]^T$, where $w(t) = q_{t,1}/\lambda_1$ and $\Delta\lambda_i \in \mathbb{R}$ such that $w(t)(\lambda_i + \Delta\lambda_i) = q_{t,i}$ for $i = 2, \cdots, K$. The expectation of $\mathbf{Q}(t+1)$ conditioned on $\mathbf{Q}(t) = \mathbf{q_t}$ becomes

$$\mathbb{E}\left[\mathbf{Q}(t+1) | \mathbf{Q}(t) = \mathbf{q_t}\right] = \mathbf{q_t} + \boldsymbol{\lambda} - \mathbb{E}\left[\mathbf{R}_{\text{QPS}}(t) | \mathbf{Q}(t) = \mathbf{q_t}\right]. \tag{24}$$

Under QPS, $\mathbb{E}[\mathbf{R}_{\text{QPS}}(t) | \mathbf{Q}(t) = \mathbf{q_t}] = r(t)(\mathbf{q_t}/w(t))$, where $r(t)$ equals $\max x$ subject to $x(\mathbf{q_t}/w(t)) \in C_{\text{erg}}(P)$. (24) can be converted into the following form:

$$\begin{aligned} \mathbb{E}\left[\mathbf{Q}(t+1) | \mathbf{Q}(t) = \mathbf{q_t}\right] &= (w(t) - r(t) + 1) \\ &\times \left[\lambda_1,\ \lambda_2 + \gamma(t)\Delta\lambda_2,\ \cdots,\ \lambda_K + \gamma(t)\Delta\lambda_K\right]^T \end{aligned} \tag{25}$$

where $\gamma(t) = 1 - 1/(w(t) - r(t) + 1)$. If $\mathbf{q_t} \in C_{\text{erg}}(P)$, then $w(t) = r(t)$; hence, $\gamma(t) = 0$ and $\mathbb{E}[\mathbf{Q}(t+1) | \mathbf{Q}(t) = \mathbf{q_t}] = \boldsymbol{\lambda}$. Otherwise, $w(t) > r(t)$ and $\gamma(t)$ is strictly less than 1. Let the angle between $\boldsymbol{\lambda} \in \mathbb{R}_+^K$ and $\mathbf{x} \in \mathbb{R}_+^K$ be denoted by $\theta_{\boldsymbol{\lambda}}(\mathbf{x})$ that is

$$\theta_{\boldsymbol{\lambda}}(\mathbf{x}) = \cos^{-1}\left(\frac{\boldsymbol{\lambda}^T\mathbf{x}}{||\boldsymbol{\lambda}||_2 ||\mathbf{x}||_2} \right), \quad 0 \le \theta_{\boldsymbol{\lambda}}(\mathbf{x}) \le \frac{\pi}{2}. \tag{26}$$

Since $\gamma(t) < 1$, $\theta_{\boldsymbol{\lambda}}(\mathbf{q_t}) \ge \theta_{\boldsymbol{\lambda}}(\mathbb{E}[\mathbf{Q}(t+1) | \mathbf{Q}(t) = \mathbf{q_t}])$. This paper assumes i.i.d. block fading and Poisson packet arrivals. Therefore, each user's queue state is the first-order Markov process, which allows the following relation to hold from Chapman–Kolmogorov equations [21]

$$\begin{aligned} &\mathbb{E}\left[\mathbf{Q}(t+1) | \mathbf{Q}(0) = \mathbf{q_0}\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\mathbf{Q}(t+1) | \mathbf{Q}(t)\right] | \mathbf{Q}(0) = \mathbf{q_0}\right] \text{ for } t = 1, 2, \cdots. \end{aligned} \tag{27}$$

Since $\theta_{\boldsymbol{\lambda}}(\mathbf{Q}(t)) \ge \theta_{\boldsymbol{\lambda}}(\mathbb{E}[\mathbf{Q}(t+1) | \mathbf{Q}(t)])$, the right-hand side (RHS) of (27) has a direction closer to $\boldsymbol{\lambda}$ than $\mathbb{E}[\mathbf{Q}(t) | \mathbf{Q}(0) = \mathbf{q_0}]$. Consequently, the following relation is obtained:

$$\begin{aligned} \theta_{\boldsymbol{\lambda}}\left(\mathbb{E}\left[\mathbf{Q}(t) | \mathbf{Q}(0) = \mathbf{q_0}\right]\right) &\ge \theta_{\boldsymbol{\lambda}}\left(\mathbb{E}\left[\mathbf{Q}(t+1) | \mathbf{Q}(0) = \mathbf{q_0}\right]\right) \\ &\text{for } t = 1, 2, \cdots. \end{aligned} \tag{28}$$

Define an infinite sequence $\theta_t = \theta_{\boldsymbol{\lambda}}(\mathbb{E}[\mathbf{Q}(t) | \mathbf{Q}(0) = \mathbf{q_0}])$ for $t = 1, 2, \cdots$. Since θ_t is monotonically decreasing and nonnegative, θ_t is a converging sequence. In the RHS of (27), $\mathbb{E}[\mathbf{Q}(t+1) | \mathbf{Q}(t)] = \mathbf{Q}(t) + \boldsymbol{\lambda} - \mathbb{E}[\mathbf{R}_{\text{QPS}}(t) | \mathbf{Q}(t)] = (1 - c)\mathbf{Q}(t) + \boldsymbol{\lambda}$,

where $c = \max r$ such that $r\mathbf{Q}(t) \in C_{\text{erg}}(P)$. Therefore, (27) can be expressed as

$$\mathbb{E}\left[\mathbf{Q}(t+1)|\mathbf{Q}(0)=\mathbf{q_0}\right] = (1-c)\mathbb{E}\left[\mathbf{Q}(t)|\mathbf{Q}(0)=\mathbf{q_0}\right] + \boldsymbol{\lambda}, \quad \text{for } t = 1, 2, \cdots. \tag{29}$$

By the convergence property, as $t \to \infty$, the angle between $\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0) = \mathbf{q_0}]$ and $\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(0) = \mathbf{q_0}]$ becomes zero. Therefore, to satisfy the equality in (29) when $t \to \infty$, the direction of these two vectors should converge to that of $\boldsymbol{\lambda}$. As a result, it can be concluded that $\lim_{t\to\infty} \theta_t = 0$, which completes the proof of the theorem. ∎

Based on Theorem 1, the throughput optimality of QPS can be proved by using Lyapunov stability analysis [1].

Theorem 2: In a fading BC, the QPS policy is throughput optimal.

Proof: With $W = T_s = 1$, the network capacity region is equivalent to $C_{\text{erg}}(P)$. Thus, we need to show that for any $\boldsymbol{\lambda} \in \text{int}C_{\text{erg}}(P)$, where int $\mathcal{S}$ denotes the interior of a set $\mathcal{S}$, the queue lengths for all users can be kept finite. First, choose the Lyapunov function $L(\mathbf{Q}(t)) = \sum_{i=1}^{K} Q_i(t)$. The evolution of $L(\mathbf{Q}(t))$ in one scheduling interval is $L(\mathbf{Q}(t+1)) = \sum_{i=1}^{K} Q_i(t+1) = \sum_{i=1}^{K}(Q_i(t) + Z_i(t) - R_i(t))$. Conditioned on $\mathbf{Q}(t) = \mathbf{q_t}$, the expected drift of the Lyapunov function is

$$\mathbb{E}\left[L\left(\mathbf{Q}(t+1)\right) - L\left(\mathbf{Q}(t)\right)|\mathbf{Q}(t) = \mathbf{q_t}\right] = \sum_{i=1}^{K}\left(\lambda_i - \mathbb{E}\left[R_i(t)|\mathbf{Q}(t)=\mathbf{q_t}\right]\right). \tag{30}$$

To prove the throughput optimality of QPS, it is required to show that as queue lengths grow sufficiently large, (30) becomes strictly negative for any $\boldsymbol{\lambda} \in \text{int}C_{\text{erg}}(P)$ [22]. By Theorem 1, in the stationary regime, $\mathbb{E}[\mathbf{Q}(t)] = w(t)\boldsymbol{\lambda}$ for some $w(t) \geq 0$. Thus, $\mathbf{Q}(t)$ can be represented as $\mathbf{Q}(t) = \mathbb{E}[\mathbf{Q}(t)] + \mathbf{N}(t) = w(t)\boldsymbol{\lambda} + \mathbf{N}(t)$, where $\mathbf{N}(t) = [N_1(t)\cdots N_K(t)]^T$ and $\mathbb{E}[N_i(t)] = 0$ for $i = 1, \cdots, K$. As $w(t)$ increases, $\mathbf{Q}(t) = w(t)(\boldsymbol{\lambda} + \mathbf{N}(t)/w(t)) \to w(t)\boldsymbol{\lambda}$ with probability 1, which results in $\mathbb{E}[\mathbf{R}_{\text{QPS}}(t)|\mathbf{Q}(t) = \mathbf{q_t}] \to \mathbb{E}[\mathbf{R}_{\text{QPS}}(t)|\mathbf{Q}(t) = w(t)\boldsymbol{\lambda}]$ with probability 1. $\mathbb{E}[\mathbf{R}_{\text{QPS}}(t)|\mathbf{Q}(t) = w(t)\boldsymbol{\lambda}] = \boldsymbol{\lambda}(\max r)$ such that $\boldsymbol{\lambda} r \in C_{\text{erg}}(P)$. If $\boldsymbol{\lambda} \in \text{int}C_{\text{erg}}(P)$, then $\max r > 1$. Thus, when $\|\mathbf{q_t}\|_\infty$ grows sufficiently large, the Lyapunov drift in (30) becomes strictly negative for any $\boldsymbol{\lambda} \in \text{int}C_{\text{erg}}(P)$. ∎

B. Fairness and Delay Properties of QPS

This section shows that for any arrival rates, QPS can arbitrarily scale the ratio of each user's average queueing delay. Also, it is shown that without new packet arrivals, QPS minimizes the expected time to empty all the backlogs. First, the next theorem shows that QPS has a capability of guaranteeing fairness among users in terms of average queueing delay.

Theorem 3: In a fading BC under the QPS policy, as $t \to \infty$, each user's average queueing delay becomes equalized.

Proof: From Theorem 1, the average queue state vector becomes proportional to the arrival rate vector as $t \to \infty$. By Little's theorem [23], the average queue length is the same as a product of the arrival rate and average queueing delay. Therefore, with QPS policy, each user's average queueing delay is equalized after the convergence. ∎

In general, QPS can satisfy a different QoS for each user in terms of average queueing delay. This property is shown in the following corollary to Theorem 3.

Corollary 1: Let $\boldsymbol{\beta}$ denote the priority vector on average queueing delay. For example, $\beta_1 = 2\beta_2$ means that the average delay of user 1 should be half of user 2's average delay. This priority can be satisfied with the QPS policy by replacing $\mathbf{Q}(t)$ with the modified queue state vector $\mathbf{Q}'(t) = [\beta_1 Q_1(t)\ \beta_2 Q_2(t)\cdots\beta_K Q_K(t)]^T$.

Proof: From Theorem 1, the average of a modified queue state vector $\mathbf{Q}'(t)$ converges to $\boldsymbol{\lambda}x$ for some $x \in \mathbb{R}_+$. Thus, user i's average queue length converges to $(\lambda_i x)/\beta_i$, and by Little's theorem, user i's average queueing delay becomes x/β_i. ∎

One reasonable way of choosing the priority vector $\boldsymbol{\beta}$ is to find a vector proportional to each user's maximum achievable average rate when no other users transmit.

The next theorem shows that without new packet arrivals, QPS minimizes the expected time to empty all the queue backlogs.

Theorem 4: Let the initial queue state vector be $\mathbf{Q}(0) = \mathbf{q_0} \in \mathbb{R}_+^K$, and assume that there are no more packet arrivals after $t = 0$. Then, in a fading BC, the QPS policy presuming the constant queue state vector of $\mathbf{q_0}$ for all $t \geq 0$ minimizes the expected time until all the queue backlogs are cleared.

Proof: Let $\mathbb{E}[T_X]$ denote the expected time until a scheduling algorithm X empties all the queue backlogs $\mathbf{q_0}$. The total supported data vector is $\mathbf{q_0}$. Thus, given $\mathbb{E}[T_X]$, the average data vector allocated per each scheduling period can be expressed as $\mathbb{E}[\mathbf{R}_X] = \mathbf{q_0}/\mathbb{E}[T_X]$. Since $C_{\text{erg}}(P)$ is convex, $\mathbb{E}[\mathbf{R}_X] \in C_{\text{erg}}(P)$ is always satisfied. Therefore, $\mathbb{E}[T_X]$ is minimized by assigning $\mathbf{R}_{\text{opt}}(\mathbf{n}(t), \mathbf{Q}(t)) \in C(\mathbf{n}(t), P)$ at time slot t such that

$$\mathbb{E}_{\mathbf{n}(t)}\left[\mathbf{R}_{\text{opt}}\left(\mathbf{n}(t), \mathbf{Q}(t)\right)\right] = \mathbf{q_0}\left(\max_{\mathbf{q_0} r \in C_{\text{erg}}(P)} r\right). \tag{31}$$

From the definition of QPS, it can be easily seen that $\mathbf{R}_{\text{opt}}(\mathbf{n}(t), \mathbf{Q}(t))$ is equal to $\mathbf{R}_{\text{QPS}}(\mathbf{n}(t), \mathbf{q_0})$, which completes the proof of the theorem. ∎

In actual systems with random packet arrivals, the property in Theorem 4 can be approximated by replacing $\mathbf{q_0}$ with the current queue state vector $\mathbf{Q}(t)$. Therefore, at each scheduling time, QPS tries to minimize the expected time to empty current queue backlogs. This property of QPS results in low average queueing delay, which will be demonstrated to be much smaller than MWMS in a fading BC.

VI. Hypersphere Approximation of the Ergodic Capacity Region of a Fading BC

At each scheduling time, QPS solves (17) which has $KM+1$ optimization variables and $KM+2K+M$ constraints. In order to capture the fading statistics, QPS requires the number of sampled fading states, $M \gg 1$. Even though GP can be efficiently solved and the constraint matrix of (17) is sparse, $M \gg 1$ implies that the computational complexity of QPS can be higher than other scheduling polices such as MWMS, which has K variables and $2K+1$ constraints. The expected rate vector under QPS is a boundary point of the ergodic capacity region that is proportional to the current queue state vector. The rate allocation satisfying this condition can be obtained by solving (12) with a proper weight vector $\boldsymbol{\mu}$ determined from the current queue state vector and ergodic capacity region. With the QPS policy, $\boldsymbol{\mu}$ is a normal vector of the tangent plane, which is drawn at the boundary point of $C_{\text{erg}}(P)$ supported by QPS.

Thus, if the boundary surface of $C_{\text{erg}}(P)$ can be characterized with a simple function, finding $\boldsymbol{\mu}$ becomes much easier, and the computational complexity of QPS becomes comparable to other scheduling policies.

This section proposes a simple method to approximate the boundary surface of $C_{\text{erg}}(P)$ by utilizing a hypersphere. By allowing a small increase in the average queueing delay, this hypersphere approximation method solves the complexity issue of QPS. First, $K+1$ boundary points on $C_{\text{erg}}(P)$ are sampled to characterize the K-dimensional hypersphere. K points correspond to each user's average rate when total transmit power is allocated to that user. They are equivalent to the intercept of each user's rate axis with $C_{\text{erg}}(P)$. The remaining point is the maximum average sum-rate vector achieved by transmitting only to the best user at each scheduling period. The next lemma provides the uniqueness of K-dimensional hypersphere constructed by using these $K+1$ rate vectors.

Lemma 1: In a fading BC with K users, there exists a unique K-dimensional hypersphere characterized with each user's maximum average rate vector and the maximum average sum-rate vector.

Proof: Let user i's maximum average rate vector be denoted by $\mathbf{x}_i = a_i \mathbf{e_i} \in \mathbb{R}_+^K$, where $\mathbf{e_i}$ is a unit vector whose ith element is 1 and the others are 0's. Also, denote the maximum average sum-rate vector by $\mathbf{x}_s \in \mathbb{R}_+^K$. In a fading BC, the sum rate is maximized by allocating full power to the best user. When excluding the trivial case, where the best user is always identical, $\mathbf{x}_s$ exists outside the $K-1$ dimensional hyperplane that passes through $\mathbf{x}_i$'s for $i = 1, \cdots, K$. The center of the K-dimensional hypershere is denoted by $\mathbf{x}_c \in \mathbb{R}^K$. Then, $\|\mathbf{x}_c - \mathbf{x}_s\|_2 = \|\mathbf{x}_c - \mathbf{x}_i\|_2$ for $i = 1, \cdots, K$. Therefore, the following linear equation is obtained:

$$A\mathbf{x}_c = \mathbf{b} \text{ where } A = \begin{bmatrix} 2(\mathbf{x}_1 - \mathbf{x}_s)^T \\ 2(\mathbf{x}_2 - \mathbf{x}_s)^T \\ \vdots \\ 2(\mathbf{x}_K - \mathbf{x}_s)^T \end{bmatrix} \in \mathbb{R}^{K \times K}$$

$$\text{and } \mathbf{b} = \begin{bmatrix} \mathbf{x}_1^T\mathbf{x}_1 - \mathbf{x}_s^T\mathbf{x}_s \\ \mathbf{x}_2^T\mathbf{x}_2 - \mathbf{x}_s^T\mathbf{x}_s \\ \vdots \\ \mathbf{x}_K^T\mathbf{x}_K - \mathbf{x}_s^T\mathbf{x}_s \end{bmatrix} \in \mathbb{R}^{K \times 1}. \quad (32)$$

A is nonsingular since every row of A is independent of each other. Thus, $\mathbf{x}_c$ has a unique solution, which is $\mathbf{x}_c = A^{-1}\mathbf{b}$. ■

Let the weight vector for QPS acquired from the hypersphere approximation be denoted by $\boldsymbol{\mu}'$. We can easily find a boundary point of the hypersphere that is proportional to the current queue state vector. If this boundary point is $\mathbf{x_b}$, $\boldsymbol{\mu}' = \mathbf{x_b} - \mathbf{x_c}$.

VII. Numerical Results and Discussion

This section presents simulation results with Poisson packet arrivals and exponentially distributed packet lengths to demonstrate stability, delay, and fairness properties of the QPS algorithm. In the simulation, the average packet length for each user, the scheduling period, and the signal bandwidth are all equal to 1. Also, the average queue length over all users is defined as $\lim_{t\to\infty} \mathbb{E}[(1/K)\sum_{i=1}^{K} Q_i(t)]$. First, Fig. 3 demonstrates the effect of the number of sampled fading states, M on the average queue length under QPS. A Rayleigh-fading BC presented in Fig. 2 is considered, where $P = 2$, user 1's average SNR $= 13$ dB, and user 2's average SNR $= 7$ dB. Also, the bit arrival rate of user 2 is assumed to be the half of user 1's. Thus, the bit arrival rate vector can be represented as $\boldsymbol{\lambda} = \lambda_1[1\ 0.5]^T$. From Fig. 2, $\boldsymbol{\lambda} \in \text{int}\ C_{\text{erg}}(P)$ if and only if $\lambda_1 < 3.9$. The average queue lengths are evaluated for different values of λ_1 when $M = 2$, 5, 10, and 20. Fig. 3 shows that as M increases, larger throughput and smaller average queue length can be achieved with QPS. About 10% throughput loss is observed with $M = 2$ compared with the maximum achievable throughput. However, this loss quickly vanishes with larger M, which becomes much less than 1% for $M = 5$. Also, it is shown that for $M > 10$, the additional decrease in average queue length is quite small, which suggests that about ten independent fading samples are sufficient in using QPS.

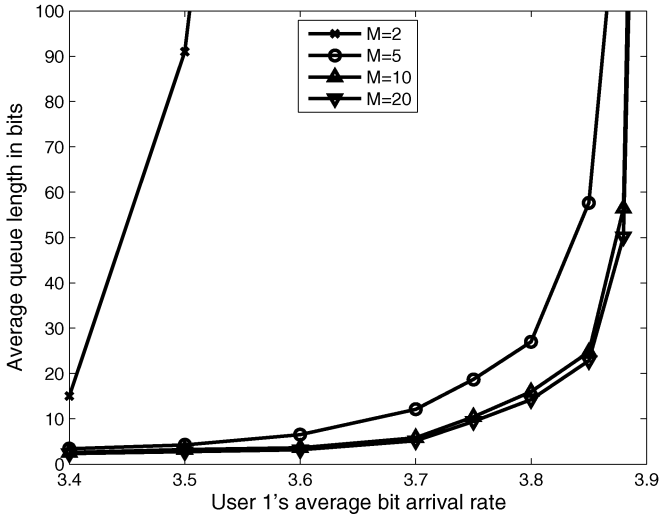

Fig. 3. Average queue length under QPS versus user 1's bit arrival rate for $M = 2$, 5, 10, and 20 ($K = P = 2$, user 1's average SNR $= 13$ dB, and user 2's average SNR $= 7$ dB, $\lambda_2 = 0.5\lambda_1$).

In Figs. 4 and 5, average queue lengths are evaluated for different values of λ_1 when $K = 2$ and $K = 10$, respectively. In both figures, $M = 10$ and five scheduling algorithms are compared: QPS, QPS with hypersphere approximation, MWMS, BCHPR, and LQHPR. For the two user case in Fig. 4, the channel and input traffic conditions are assumed to be the same as in Fig. 3. Fig. 4 shows that the average queue length of QPS is about 30% smaller than that of MWMS for any $\lambda_1 < 3.9$. Since MWMS is a throughput optimal policy, this observation corroborates the throughput optimality of QPS. LQHPR and BCHPR, which are not throughput optimal, have about 12% and 5% throughput loss, respectively. QPS using the hypersphere approximation of $C_{\text{erg}}(P)$ slightly increases the average queue length compared to QPS. However, its average queue length is still much smaller than MWMS. Simulation results with ten users are presented in Fig. 5. $P = 10$ and user i's average SNR is equal to $20 - (i-1)$ (dB) for $i = 1, \cdots, 10$. Also, the bit arrival rate is identical for all users. QPS is observed to provide about a 40%–50% reduction in average queue length compared to MWMS, a larger difference than in the two user case. The throughput loss of LQHPR and BCHPR is around 30% and 10%, respectively, which is also much greater than in Fig. 4. Accuracy of the hypersphere approximation is

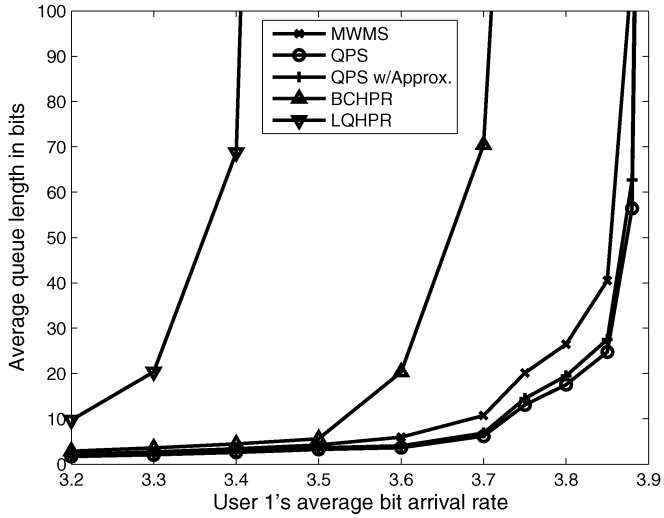

Fig. 4. Average queue length versus user 1's bit arrival rate under five scheduling policies ($K = P = 2$, $M = 10$, user 1's average $\mathrm{SNR} = 13$ dB, and user 2's average $\mathrm{SNR} = 7$ dB, $\lambda_2 = 0.5\lambda_1$).

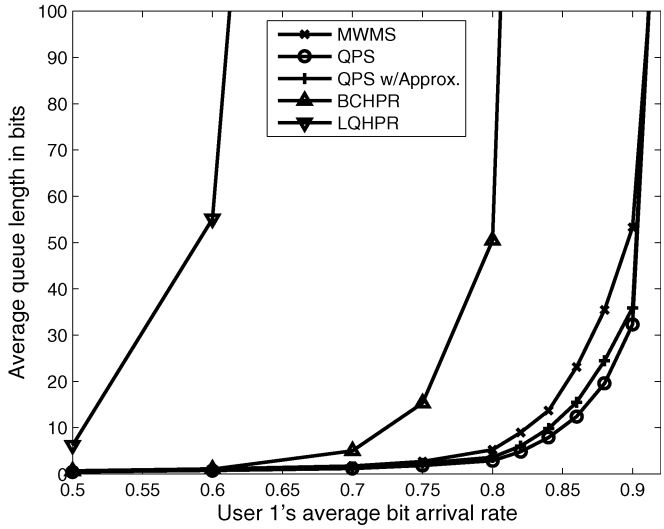

Fig. 5. Average queue length versus user 1's bit arrival rate under five scheduling policies ($K = P = 10$, $M = 10$, user i's average SNR (dB) $= 20 - (i-1)$, and $\lambda_i = \lambda_1$ for $i = 1, \cdots, 10$).

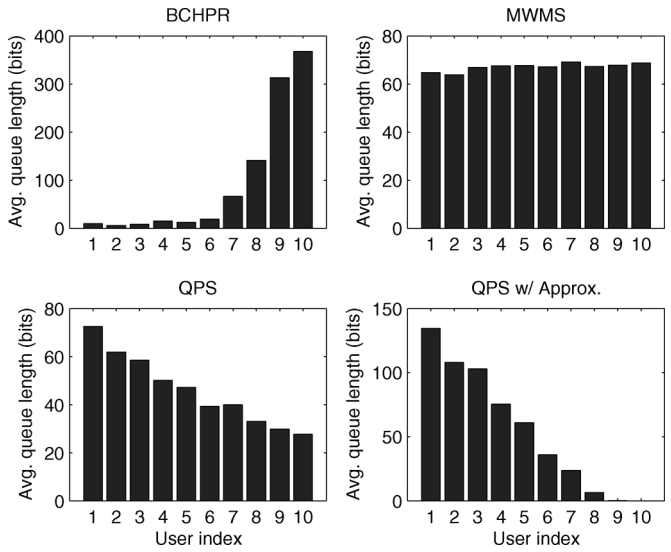

Fig. 6. Each user's average queue length under four scheduling policies ($K = P = 10$, $M = 10$, user i's average SNR (dB) $= 20 - 0.5(i-1)$, and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

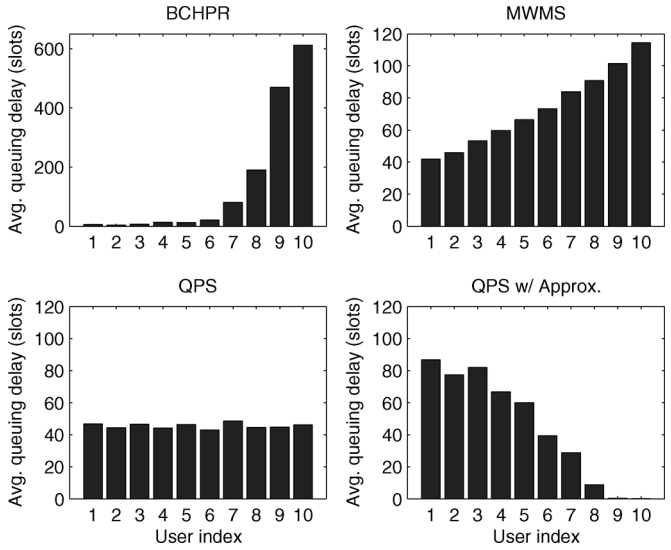

Fig. 7. Each user's average queueing delay under four scheduling policies ($K = P = 10, M = 10$, user i's average SNR (dB) $= 20 - 0.5(i-1)$, and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

somewhat lower than in the two user case, but this method still gives about a 30% decrease in the average queue length compared to MWMS.

The fairness properties of QPS, QPS with hypersphere approximation, MWMS, and BCHPR with ten users are illustrated in Figs. 6 and 7. $P = 10$, $M = 10$, user i's average SNR is equal to $20 - 0.5(i-1)$ (dB), and $\lambda_i = 1.55(0.9)^{i-1}$ for $i = 1, \cdots, 10$. First, Fig. 6 demonstrates each user's average queue length in bits for the above four scheduling policies. It is observed that fairness among users is not satisfied under the BCHPR, which provides intolerably long average queueing delay for users with worse channel conditions. MWMS is shown to approximately equalize every user's average queue length. Since each user has a different arrival rate, by Little's theorem, MWMS provides smaller average queueing delay for the user with higher bit arrival rate. On the other hand, each user's average queue length under QPS is shown to be proportional to the bit arrival rate vector so that average queueing delay of every user is equalized. Therefore, under the QPS policy, fairness among users is guaranteed in terms of average queueing delay. QPS with hypersphere approximation also shows a similar tendency with QPS, but some deviation from the arrival rate vector is observed because of the approximation error. Fig. 7 presents each user's average queueing delay in slots, which indicates that QPS equalizes every user's average queueing delay.

VIII. Conclusion

In fading broadcast channels, QPS is shown to provide more desirable delay and fairness properties than MWMS, a well-known throughput optimal scheduling policy. The GP formulation of QPS, which is also applicable to OFDM systems, is presented, where GP is a special form of convex optimization problems with well-developed efficient algorithms.

The throughput optimality of QPS is proved, and it is shown that QPS can arbitrarily scale the ratio of each user's average queueing delay, which is essential in satisfying different QoS requirement of each user. Numerical results demonstrate that QPS provides significantly smaller average queueing delay compared with MWMS for any arrival rate vector within the network capacity region.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for valuable comments and for informing us about the work in [13].

REFERENCES

[1] M. J. Neely, E. Modiano, and C. E. Rohrs, "Power allocation and routing in multibeam satellites with time-varying channels," *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 138–152, Feb. 2003.

[2] R. Barry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Process. Mag.*, pp. 59–68, Sep. 2004.

[3] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Commun. Mag.*, pp. 150–154, Feb. 2001.

[4] H. Viswanathan and K. Kumaran, "Rate scheduling in multiple antenna downlink," in *Proc. 39th Annu. Allerton Conf. Commun., Control, Comput.*, Allerton, IL, Oct. 2001, pp. 747–756.

[5] C. Swannack, E. Uysal-Biyikoglu, and G. Wornell, "Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel," in *Proc. 42nd Annu. Allerton Conf. Commun., Control, Comput.*, Allerton, IL, Oct. 2004, pp. 440–449.

[6] L. Tassiulas and A. Ephremides, "Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks," *IEEE Trans. Autom. Control*, vol. 37, pp. 1936–1948, Dec. 1992.

[7] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proc. IEEE INFOCOM*, San Francisco, CA, Mar. 1996, vol. 1, pp. 296–302.

[8] G. Song, Y. Li, L. Cimini, Jr, and H. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Atlanta, GA, 2004, pp. 1939–1944.

[9] E. Yeh and A. Cohen, "Throughput and delay optimal resource allocation in multi-access fading channels," in *Proc. Int. Symp. Inf. Theory*, Yokohama, Japan, 2003, p. 245.

[10] D. Tse and S. Hanly, "Multi-access fading channels: Part I," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 2796–2831, 1998.

[11] K. Seong, R. Narasimhan, and J. Cioffi, "Cross-layer resource allocation via geometric programming in fading broadcast channels," in *Proc. IEEE Veh. Technol. Conf.–Spring*, Melbourne, Australia, May 2006.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2003.

[13] R. Leelahakriengkrai and R. Agrawal, "Scheduling in multimedia wireless networks," in *Proc. 17th Int. Teletraffic Congr.*, Salvador da Bahia, Brazil, Dec. 2001, pp. 285–298.

[14] K. Seong, R. Narasimhan, and J. Cioffi, "Queue proportional scheduling in Gaussian broadcast channels," in *Proc. IEEE Int. Conf. Commun.*, Istanbul, Turkey, Jun. 2006.

[15] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1997.

[16] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels-Part I: Ergodic capacity," *IEEE Trans. Inf. Theory*, vol. 47, pp. 1083–1102, Mar. 2001.

[17] D. Tse, "Optimal power allocation over parallel Gaussian broadcast channels," in *Proc. Int. Symp. Inf.*, Ulm, Germany, Jun. 1997, p. 27.

[18] M. Chiang, "Geometric programming for communication systems," *Foundations and Trends in Commun. Inf. Theory*, vol. 2, no. 1, pp. 1–156, Aug. 2005.

[19] S. Asmussen, *Applied Probability and Queues*. New York: Springer, 2000.

[20] P. Viswanath, D. Tse, and R. Laroia, "Opportunistic beamforming using dumb antennas," *IEEE Trans. Inf. Theory*, vol. 48, pp. 1277–1294, Jun. 2002.

[21] S. Ross, *Stochastic Processes*. New York: Wiley, 1996.

[22] J. Walrand, *An Introduction to Queueing Networks*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[23] D. Bertsekas and R. Gallager, *Data Networks*. Englewood Cliffs, NJ: PrenticeHall, 1992.

Kibeom Seong (S'05) received the B.S. and M.S. degrees in electrical engineering from Seoul National University, Seoul, Korea, in 1997 and 1999, respectively. He is currently working towards the Ph.D. degree at Stanford University, Stanford, CA.

From 1999 to 2002, he served as a Faculty Member in the Department of Electrical Engineering, Korea Military Academy, Seoul. His research interests include communication theory, multiuser information theory, and dynamic resource management in wireless and wireline communication systems.

Ravi Narasimhan (S'96–M'99–SM'05) received the B.S. degree (Highest Honors) in electrical engineering and the Certificate of Distinction from the University of California, Berkeley, in 1995, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1996 and 2000, respectively.

From 2000 to 2004, he was involved in research and development for next-generation wireless systems at Marvell Semiconductor, Inc., Sunnyvale, CA, most recently as Senior Engineering Design Manager in the Signal Processing Department. In July 2004, he joined the faculty in the Electrical Engineering Department, University of California, Santa Cruz. He is also an active consultant for the wireless industry. His research interests include multiple-input–multiple-output (MIMO) systems, multicarrier modulation, and multiuser communication theory.

Dr. Narasimhan is a member of Phi Beta Kappa, Sigma Xi and Golden Key National Honor Society. He received the Warren Y. Dere Memorial Prize from University of California, Berkeley, in 1995. He secured the first rank in the Ph.D. qualifying examination in electrical engineering at Stanford University. He also received the Best Student Paper Award for U.S. at the IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC), Boston, MA, September 1998. His biography was selected for publication in *Who's Who in America* and *Who's Who in Science and Engineering*.

John M. Cioffi (S'77–M'78–SM'90–F'96) received the B.S. degree from the University of Illinois at Urbana–Champaign, Urbana, in 1978, and the Ph.D. degree from Stanford University, Stanford, CA, in 1984, both in electrical engineering.

He was with Bell Laboratories from 1978 to 1984, and IBM Research from 1984 to 1986. He has been a Professor of Electrical Engineering with Stanford University since 1986. He founded Amati Communications Corporation in 1991 (purchased by Texas Instruments in 1997), and was Officer/Director from 1991 to 1997. He is currently Chairman and founder of ASSIA Inc., a company responsible for the introduction and use of Dynamic Spectrum Management by several large telephone companies. He has served on the Board of Directors of public companies Amati, Marvell, and Integrated Telecom Express. He currently is on the Board of Directors of Teknovus, Teranetics, ClariPhy, and ASSIA. He is on the Advisory Boards of Wavion and Amicus. He has published over 280 papers and holds over 80 patents, most of which are widely licensed, including basic patents on DMT, VDSL, and vectored transmission. His specific interests are in the area of high-performance digital transmission.

Dr. Cioffi is a member of the National Academy of Engineering (2001). He has been the recipient of various awards: International Marconi Prize (2006), Holder of Hitachi America Professorship in Electrical Engineering at Stanford University (2002), the IEEE Kobayashi Medal (2001), the IEEE Third Millennium Medal (2000), IEE JJ Tomson Medal (2000), University of Illinois Outstanding Alumnus (1999), Committee T1 Outstanding Achievement Award of the ANSI (1995), Outstanding Achievement Award from the American National Standards Institute for "contributions to ADSL" (10/95), NSF Presidential Investigator (1987–1992), IEEE Communications Magazine Best Paper Award (1991), the IEEE ISSLS Best Paper Award (2004), and the Faculty Development Award from IBM Research (1986–1988).