Queue Proportional Scheduling in Gaussian Broadcast Channels

Kibeom Seong
Dept. of Electrical Engineering
Stanford University
Stanford, CA 94305 USA
Email: kseong@stanford.edu

Ravi Narasimhan
Dept. of Electrical Engineering
University of California
Santa Cruz, CA 95064 USA
Email: ravi@soe.ucsc.edu

John M. Cioffi
Dept. of Electrical Engineering
Stanford University
Stanford, CA 94305 USA
Email: cioffi@stanford.edu

Abstract— **Queue Proportional Scheduling (QPS) assigns each user a data rate proportional to the number of packets (or bits) in that user's queue. This paper analyzes stability, delay and fairness properties of QPS in a Gaussian broadcast channel (BC). QPS is shown to achieve throughput optimality, and guarantee fairness as well as different priorities among users in terms of average queuing delay. One well known throughput optimal policy for broadcast channels is Maximum Weight Matching Scheduling (MWMS) that maximizes the inner product of the queue state vector and the achievable rate vector. Simulation results with Poisson packet arrivals and exponentially distributed packet lengths demonstrate that QPS provides a significant decrease in average queuing delay compared to MWMS in a Gaussian BC.**

I. Introduction

Optimal allocation of communication resources such as the transmit power and data rate is a central problem in multi-user 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 queuing, cannot guarantee stability of queuing 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 the bursty input traffic, it is generally quite difficult to estimate the packet arrival rates. Thus, resource allocation solely based on CSI is unable to properly update rate allocation 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 queuing delay as well as more frequent packet loss.

To account for queuing parameters, a cross-layer approach to resource allocation has been recently proposed in [2], [3], [4] 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 broadcast and multiple-access channels. 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 [5][6]. This MWMS policy is proved to be throughput optimal for both Gaussian broadcast channels (BC) and multiple-access channels (MAC) [1], [2]. Recent applications of MWMS can be also found in OFDM downlink systems [7] and MIMO downlink systems [3], [4]. For the Gaussian MAC, [8] shows that MWMS actually minimizes the average queuing delay if symmetric channels and equal packet arrival rates are assumed. This property is a consequence of the polymatroidal structure of Gaussian MAC capacity region [9]. However, for the Gaussian BC, there are no such structural properties in the capacity region so that even with symmetry assumptions, MWMS cannot guarantee the minimum average queuing delay.

On the other hand, [10] proposes a new throughput optimal scheduling policy in Gaussian broadcast channels. It is a type of minimum draining time policy introduced in [11]. At each scheduling period, this algorithm allocates each user a data rate on the boundary of capacity region such that the ratio of each user's rate to the queue length is identical for every user. In other words, it assigns the rate vector that is proportional to the queue state vector as well as on the boundary of the capacity region. This paper calls the above scheduling policy Queue Proportional Scheduling (QPS). While MWMS has been widely applied and studied, properties of QPS are relatively unknown except its throughput optimality. This paper further investigates delay and fairness properties of QPS in a Gaussian BC.

In [10], a fluid model is utilized to prove throughput optimality of QPS. In this paper, we present another proof for throughput optimality without considering fluid models. Though this model is simple to analyze, our direct approach provides some insights on fairness property of QPS in terms of average queuing delay, and it eventually reveals that QPS has a capability of arbitrarily scaling the ratio of each user's average queuing delay. Moreover, this paper numerically evaluates and compares average queuing delays of both MWMS and QPS with Poisson packet arrivals and exponentially distributed packet lengths. In a Gaussian BC, QPS is demonstrated to provide a substantial decrease in average queuing delay compared to MWMS.

The organization of this paper is as follows: Section II

This work was supported by a Stanford Graduate Fellowship.

1-4244-0355-3/06/$20.00 (c) 2006 IEEE

describes the model of Gaussian broadcast channels and queuing systems. QPS is introduced in Section III along with the description of the conventional MWMS. In Section IV, throughput optimality of QPS is proved and Section V presents delay and fairness properties of QPS. Numerical results and discussion are given in Section VI and Section VII provides the 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. $1[\cdot]$ is the indicator function which equals 1 if its argument is satisfied, 0 otherwise.

II. System model

Consider a Gaussian broadcast channel with a single transmitter sending independent messages to K users over two-sided bandwidth $2W$. At time t, the received signal of user i is expressed as

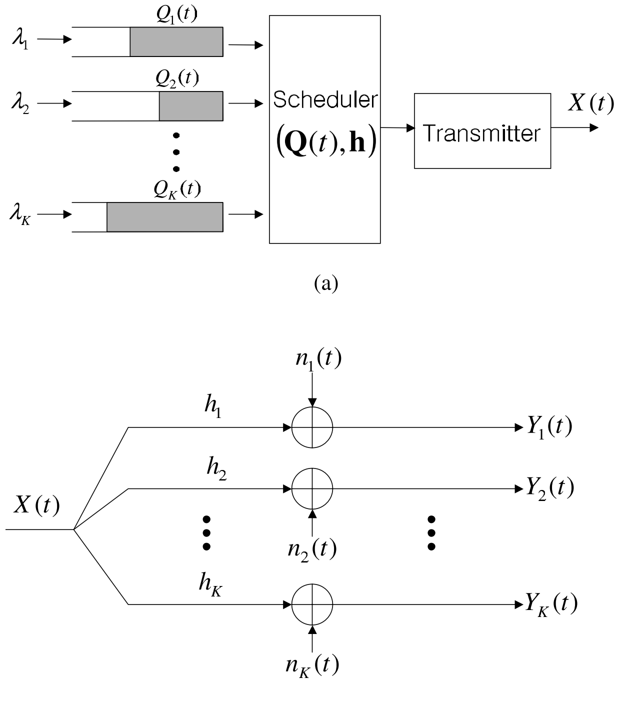
$$Y_i(t) = h_i X(t) + n_i(t), \quad i = 1, \cdots, K \tag{1}$$

where the transmitted signal $X(t)$ is composed of K independent messages, the complex channel gain of user i is denoted by h_i and $n_i(t)$'s are independent and identically distributed (i.i.d.) zero-mean Gaussian noise with power $N_0 W$. The transmitter has a total power constraint of P. This multi-user channel is a degraded broadcast channel whose capacity region is well known [12]. Without loss of generality, it can be assumed that $W = 1$ and $|h_1| \geq |h_2| \geq \cdots \geq |h_K|$. Then, the capacity region is defined as

$$C(P) = \left\{ R_i : R_i \leq \log \left(1 + \frac{\alpha_i |h_i|^2 P}{N_0 + \Sigma_{j<i} \alpha_j |h_i|^2 P} \right), \right. \\ \left. i = 1, 2, \cdots, K, \text{ where } \Sigma_i \alpha_i = 1 \right\} \tag{2}$$

where α_i is the fraction of total transmit power used for user i's signal. This capacity region is convex since time-sharing can be always performed, and each point is achieved by superposition coding along with the successive interference cancellation [12].

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 $\text{var}(A_i(t+T) - A_i(t)) < \infty$ for $T < \infty$. The packet lengths in bits $\{X_i\}$ are i.i.d. exponentially distributed and satisfy $E(X_i) = \mu_i < \infty$, and $E({X_i}^2) < \infty$. Packet lengths are assumed independent of packet arrival processes. User i's arrival rate in bits is given by $\lambda_i = a_i \mu_i$. 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, which is assumed 1 without loss of generality. Over each scheduling period, the achievable data rate vector should be within the capacity region $C(P)$ defined in (2). At time t, the number of bits waiting to be sent to user i is denoted by $Q_i(t)$. A time interval $[t, t+1)$, with $t = 0, 1, 2, \cdots$, is denoted by the *time slot* t, and $Z_i(t)$ is defined as the number of arrived bits at user i's queue during the time slot t. Then, after a scheduling period, user i's queue state vector



(b)

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

is equal to $Q_i(t+1) = \max\{Q_i(t) - R_i(t), 0\} + Z_i(t)$. The allocated rate vector at time slot t, $\mathbf{R}(t)$ is determined by the queue-aware scheduler based on both queue states and channel conditions. Fig. 1 summarizes the system model described above. This paper adopts the stability definition of queuing systems given in [1]. Thus, with the overflow function defined by $g(M) = \limsup_{t\to\infty} \frac{1}{t} \int_0^t 1_{[Q_i(\tau) > M]} \, d\tau$, 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. Also, [1] defines the network capacity region as a set of arrival rate vectors for which all queues can be stable. If a scheduling method achieves the entire network capacity region, it is called throughput optimal.

III. Queue proportional scheduling (QPS)

First, MWMS takes the following form in a Gaussian BC.

$$\mathbf{R}_{MWMS}(t) = \arg \max_{\mathbf{r} \in C(P)} \sum_{i=1}^{K} \alpha_i Q_i(t) r_i \tag{3}$$

where the achievable rate vector is $\mathbf{r} = [r_1 \; r_2 \; \cdots \; r_K]^T$ and $\mathbf{R}_{MWMS}(t)$ denotes the scheduled data rate vector at time slot t by employing MWMS. For user i, $Q_i(t)$ is the queue state at time t, r_i is the achievable rate, and α_i is the priority weight which is equal to 1 if all users have the same priority. From (3), this algorithm tends to allocate higher data rate

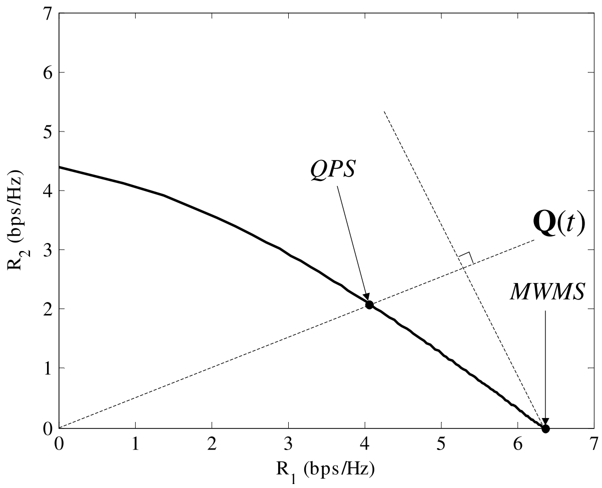


Fig. 2. Capacity region of two user Gaussian BC, and rate vectors of QPS and MWMS when the queue state vector is $\mathbf{Q}(t)$ (User 1's SNR=19dB and user 2's SNR=13dB).

to the user with longer queue or better channel conditions. By jointly considering queue and channel states, MWMS achieves the entire network capacity region. A *delay optimal* scheduling policy minimizes average queuing delay over all K users, which is defined as $\lim_{t\to\infty}\mathbb{E}[\frac{1}{K}\sum_{i=1}^{K}Q_i(t)]$ [8]. As mentioned in Section I, the delay optimal scheduling in broadcast channels is still unknown.

This section introduces Queue Proportional Scheduling (QPS), which has advantages over MWMS in the Gaussian BC in terms of delay and fairness. At each scheduling period, QPS assigns a maximum data rate vector which is proportional to the current queue state vector. Assuming equal priority for all users, the proposed algorithm can be formulated as follows.

$$\begin{aligned}\mathbf{R}_{QPS}(t) &= \mathbf{Q}(t)(\max x)\\ \text{subject to}\quad \mathbf{Q}(t)x &\in C(P) \quad \text{and} \quad x \le 1. \end{aligned}\tag{4}$$

At time slot t, $\mathbf{R}_{QPS}(t)$ is the rate vector scheduled by QPS and the queue state vector in bits is denoted by $\mathbf{Q}(t) = [Q_1(t)\ \ Q_2(t)\ \ \cdots\ \ Q_K(t)]^T$. (4) is a convex optimization problem with a globally optimal solution [13]. By utilizing the degradedness of a Gaussian BC, it is shown in [14] that (4) can be converted into geometric programming (GP) which is a special form of convex optimization problems with very efficient interior point methods.

For the queue state vector $\mathbf{Q}(t)$, Fig. 2 illustrates two distinct rate vectors supported by MWMS and QPS. Two user Gaussian BC is considered where user 1's average signal-to-noise ratio (SNR) is 19dB and user 2's average SNR is 13dB. Since both bandwidth and scheduling period are assumed 1, bps/Hz is equivalent to bits/slot. In other words, the rate region in Fig. 2 shows how many bits can be supported in each time slot. The given queue state vector satisfies $Q_2(t) = 0.5Q_1(t)$, which results in $R_{QPS}(t) = [4.1\ \ 2.05]^T$ and $R_{MWMS}(t) = [6.34\ \ 0]^T$. From Fig. 2, it can be anticipated that as the queue state changes, MWMS will exhibit more fluctuations in the supported rate vector compared to QPS. According to queuing theory, lower variance in service rate or arrival rate provide smaller queuing delay [15]. Therefore, QPS can be expected to have smaller average queuing delay than MWMS, which will be elaborated in next sections.

IV. Throughput optimality of QPS

In this section, QPS is proved to be a throughput optimal scheduling policy in the Gaussian BC. The next theorem shows convergence property of the expected queue state vector, which is crucial in proving throughput optimality as well as analyzing fairness properties of QPS.

Theorem 1: Under the QPS policy in a Gaussian 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. Consider time slot t 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 $q_{t,1} \neq 0$ and $\lambda_1 \neq 0$. 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

$$\mathbb{E}\left[\mathbf{Q}(t+1)|\,\mathbf{Q}(t) = \mathbf{q_t}\right] = \mathbf{q_t} + \boldsymbol{\lambda} - \mathbf{R}_{QPS}(t). \tag{5}$$

By definition of QPS, $\mathbf{R}_{QPS}(t) = r(t)\left(\mathbf{q_t}/w(t)\right)$ where $r(t)$ equals $\max x$ subject to $x\left(\mathbf{q_t}/w(t)\right) \in C(P)$ and $x \le w(t)$. (5) can be converted into the following form.

$$\begin{aligned}\mathbb{E}\left[\mathbf{Q}(t+1)|\,\mathbf{Q}(t) = \mathbf{q_t}\right] &= (w(t) - r(t) + 1) \times\\ &[\lambda_1,\ \lambda_2 + \gamma(t)\Delta\lambda_2,\ \cdots\ ,\lambda_K + \gamma(t)\Delta\lambda_K]^T \end{aligned}\tag{6}$$

where $\gamma(t) = 1 - 1/(w(t) - r(t) + 1)$. If $\mathbf{q_t} \in C(P)$, 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

$$\theta_{\boldsymbol{\lambda}}(\mathbf{x}) = \cos^{-1}\left(\frac{\boldsymbol{\lambda}^T\mathbf{x}}{\|\boldsymbol{\lambda}\|_2\|\mathbf{x}\|_2}\right),\quad 0 \le \theta_{\boldsymbol{\lambda}}(\mathbf{x}) \le \frac{\pi}{2}. \tag{7}$$

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 1st order Markov process, which allows the following relation to hold from Chapman-Kolmogorov equations [16].

$$\begin{aligned}&\mathbb{E}\left[\mathbf{Q}(t+1)|\,\mathbf{Q}(0) = \mathbf{q_0}\right] =\\ &\mathbb{E}\left[\mathbb{E}\left[\mathbf{Q}(t+1)|\,\mathbf{Q}(t)\right]|\mathbf{Q}(0) = \mathbf{q_0}\right] \quad \text{for} \quad t = 1, 2, \cdots \end{aligned}\tag{8}$$

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 (8) 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.

$$\begin{aligned}\theta_{\boldsymbol{\lambda}}\left(\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0) = \mathbf{q_0}]\right) \ge \theta_{\boldsymbol{\lambda}}\left(\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(0) = \mathbf{q_0}]\right)&\\ \text{for} \quad t = 1, 2, \cdots& \end{aligned}\tag{9}$$

Define an infinite sequence $\theta_t \in \mathbb{R}_+$ such that $\theta_t = \theta_{\boldsymbol{\lambda}}(\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0) = \mathbf{q_0}])$ for $t = 1, 2, \cdots$. Since θ_t is monotonically decreasing and $\theta_t \geq 0$, θ_t is a converging sequence. In the RHS of (8), $\mathbb{E}[\mathbf{Q}(t+1)|\mathbf{Q}(t)] = \mathbf{Q}(t)+\boldsymbol{\lambda}-\mathbf{R}_{QPS}(t) = (1-c)\mathbf{Q}(t) + \boldsymbol{\lambda}$ where $c = \max r$ such that $r\mathbf{Q}(t) \in C(P)$ and $r \leq 1$. Therefore, (8) can be expressed as

$$\begin{aligned}&\mathbb{E}\left[\mathbf{Q}(t+1)|\,\mathbf{Q}(0) = \mathbf{q_0}\right] = \\ &\quad (1-c)\mathbb{E}\left[\mathbf{Q}(t)\,|\mathbf{Q}(0) = \mathbf{q_0}\,\right] + \boldsymbol{\lambda}, \quad \text{for} \quad t = 1, 2, \cdots \end{aligned} \tag{10}$$

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 (10), 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, throughput optimality of QPS policy can be proved by using Lyapunov stability analysis [1].

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

Proof: Assuming $T_s = 1$, the network capacity region is equivalent to $C(P)$. Thus, it needs to be shown that for any $\boldsymbol{\lambda} \in \text{int } C(P)$ 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 $L(\mathbf{Q}(t+1)) = \sum_{i=1}^{K} Q_i(t+1) = \sum_{i=1}^{K}(\max\{Q_i(t) - R_i(t), 0\} + Z_i(t))$. QPS always satisfies $Q_i(t) \geq R_i(t)$ for $i = 1, \cdots, K$. Therefore, $\max\{\ , 0\}$ operation can be ignored. Without loss of generality, assume that the initial queue state vector is $\mathbf{Q}(0) = \mathbf{q_0}$ where $\|\mathbf{q_0}\|_\infty$ is sufficiently small. Then, conditioned on $\mathbf{Q}(t) = \mathbf{q_t}$, the expected drift of the Lyapunov function is

$$\begin{aligned}&\mathbb{E}\left[L(\mathbf{Q}(t+1)) - L(\mathbf{Q}(t))|\mathbf{Q}(t) = \mathbf{q_t}\right] = \\ &\qquad\qquad \sum_{i=1}^{K} \left(\lambda_i - (R_i(t)|\mathbf{Q}(t) = \mathbf{q_t})\right). \end{aligned} \tag{11}$$

To prove throughput optimality of QPS, it is required to show that as $\|\mathbf{q_t}\|_\infty \to \infty$, (11) becomes strictly negative for any $\boldsymbol{\lambda} \in \text{int } C(P)$. Since $\|\mathbf{q_0}\|_\infty$ is sufficiently small and the bit arrival process has finite mean and variance, $\|\mathbf{q_t}\|_\infty \to \infty$ also implies $t \to \infty$. Thus, from Theorem 1, as $\|\mathbf{q_t}\|_\infty \to \infty$, $\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0) = \mathbf{q_0}]$ converges to $x\boldsymbol{\lambda}$ for some $x \geq 0$. In general, $\mathbf{Q}(t)$ can be represented as $\mathbf{Q}(t) = \mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0) = \mathbf{q_0}] + \mathbf{N}(t) = (w(t)\boldsymbol{\lambda} + \boldsymbol{\epsilon}(t)) + \mathbf{N}(t)$ where $\mathbf{N}(t)$ denotes the deviation of $\mathbf{Q}(t)$ from its conditional mean value, and $\boldsymbol{\epsilon}(t)$ is the deviation of $\mathbb{E}[\mathbf{Q}(t)|\mathbf{Q}(0) = \mathbf{q_0}]$ from the direction of $\boldsymbol{\lambda}$. Since $\mathbf{R}_{QPS}(t) \in C(P)$ and the bit arrival process has finite mean and variance, $\|\mathbf{N}(t)\|_\infty$ is finite. Also, from Theorem 1, $\boldsymbol{\epsilon}(t)$ vanishes as $\|\mathbf{q_t}\|_\infty \to \infty$. Consequently, as $\|\mathbf{q_t}\|_\infty \to \infty$, $\mathbf{Q}(t) = w(t)(\boldsymbol{\lambda}+(\boldsymbol{\epsilon}(t)+\mathbf{N}(t))/w(t)) \to w(t)\boldsymbol{\lambda}$ with probability 1.

Under the QPS policy, $\lim_{\|\mathbf{q_t}\|_\infty \to\infty}(\mathbf{R}(t)|\mathbf{Q}(t) = \mathbf{q_t}) = \lim_{w(t)\to\infty}(\mathbf{R}(t)|\mathbf{Q}(t) = w(t)\boldsymbol{\lambda}) = \boldsymbol{\lambda}(\max r)$ such that $\boldsymbol{\lambda}r \in C(P)$ and $r \leq w(t)$. If $\boldsymbol{\lambda} \in \text{int } C(P)$, then $\max r > 1$. Hence, when $\|\mathbf{q_t}\|_\infty \to \infty$, the Lyapunov drift in (11) becomes strictly negative for any $\boldsymbol{\lambda} \in \text{int } C(P)$. ∎

V. 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 queuing delay. Also, it is shown that without new packet arrivals, QPS empties all the backlogs faster than any other scheduling policies, which indicates QPS is a type of minimum draining time policies in [11]. First, the next theorem shows that QPS has a capability of guaranteeing fairness among users in terms of average queuing delay.

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

Proof: From Theorem 1, the expected queue state vector becomes proportional to the arrival rate vector as $t \to \infty$. By Little's theorem [17], an average queue length is the same as a product of the arrival rate and average queuing delay. Therefore, with the QPS policy, each user's average queuing delay is equalized after the convergence. ∎

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

Corollary 1: Let $\boldsymbol{\alpha}$ denote the priority vector on average queuing delay. For example, $\alpha_1 = 2\alpha_2$, 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 $\mathbf{Q}'(t) = [\alpha_1 Q_1(t)\ \alpha_2 Q_2(t)\ \cdots\ \alpha_K Q_K(t)]^T$.

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 $(\lambda_i x)/\alpha_i$, and by Little's theorem, user i's average queuing delay becomes x/α_i. ∎

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

The following theorem proves that QPS guarantees the minimum draining time without new packet arrivals.

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 Gaussian BC, no other scheduling methods clear up the backlogs faster than QPS.

Proof: Let T_X denote the time until a scheduling algorithm X empties all the queue backlogs. Over the time interval $[0,\ T_X]$, the total supported data vector in bits is $\mathbf{q_0}$. Thus, the average data vector allocated per each scheduling period is given by $\mathbf{R}_{X,avg} = \mathbf{q_0}/T_X$. Since $C(P)$ is convex, $\mathbf{R}_{X,avg} \in C(P)$ is always satisfied. Therefore, T_X is minimized by assigning $\mathbf{R}_{opt} = (\max r)\mathbf{q_0}$ such that $r\mathbf{q_0} \in C(P)$ for every scheduling period. Under the QPS policy, without new packet arrivals, the direction of the queue state vector is preserved over time since the scheduled rate vector is always proportional to the queue state vector. Therefore, by definition, QPS allocates a data rate vector $\mathbf{R}_{QPS} = (\max r)\mathbf{q_0}$ such that $r\mathbf{q_0} \in C(P)$ at each scheduling time. It can be easily seen that $\mathbf{R}_{opt} = \mathbf{R}_{QPS}$ and this 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. Therefore, at each scheduling time, QPS tries to empty the current queue backlogs as fast as possible. This property of QPS results in low average queuing delay, which will be demonstrated to be much smaller than MWMS in a Gaussian BC.

VI. 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, average packet length for each user, scheduling period, and signal bandwidth are all equal to 1, and noise power is 0.1. In Fig. 3 and Fig. 4, the average queue length is evaluated for different values of λ_1 with two users and ten users, respectively. Four scheduling algorithms are compared in both figures: QPS, MWMS, Longest Queue Highest Possible Rate (LQHPR) and Best Channel Highest Possible Rate (BCHPR) [18]. LQHPR allocates full power to a user with the longest queue. Under the BCHPR policy, a user with the better channel condition takes higher priority in resource allocation, user i is served only if some transmit power remains after clearing queue backlogs of users with higher priorities than user i.

For the two user case in Fig. 3, a Gaussian BC channel presented in Fig. 2 is considered where the power constraint $P = 2$ and the channel gain vector is $\mathbf{h} = [2\ 1]^T$; thus, user 1's SNR=19dB, and user 2's SNR=13dB. Also, the bit arrival rate vector satisfies $\boldsymbol{\lambda} = \lambda_1[1\ 0.5]^T$. From Fig. 2, $\boldsymbol{\lambda} \in \text{int}\ C(P)$ if and only if $\lambda_1 < 4.1$. Fig. 3 demonstrates that the average queue length of QPS is about 30% smaller than that of MWMS for any $\lambda_1 < 4.1$. Since MWMS is a throughput optimal policy, this observation corroborates the throughput optimality of QPS. LQHPR and BCHPR, which are not throughput optimal, have much longer average queue lengths than MWMS. Simulation results with 10 users are shown in Fig. 4. The total transmit power is $P = 10$ and user i's channel gain $h_i = 2 - 0.1(i-1)$ and $\lambda_i = \lambda_1(0.9)^{i-1}$ for $i = 1, \cdots, 10$. QPS provides about 40-50% smaller average queue length than MWMS, which is a more prominent difference than in the two user case. BCHPR is also observed to have around 20% smaller average queue length than MWMS at small λ_1. However, as λ_1 approaches to the boundary of a network capacity region, the average queue length of BCHPR grows faster than MWMS.

The fairness properties of QPS, MWMS and BCHPR with 10 users are illustrated in Fig. 5. The simulation environment is identical with Fig. 4 and $\lambda_1 = 1.32$ is considered. Fig. 5 shows the arrival rate vector as well as each user's average queuing delay in slots for above three scheduling policies. It is observed that fairness among users is not satisfied under the BCHPR, which provides intolerably long average queue length for users with worse channel conditions. MWMS tends to equalize each user's average queue length. Since each user has a different arrival rate, by Little's theorem, MWMS provides smaller average queuing delay for the user with higher arrival rate. On the other hand, the average queue length of QPS is shown to be almost proportional to the arrival rate vector so that each user's average queuing delay is equalized. Therefore, under the QPS policy, fairness among users is guaranteed in terms of average queuing delay.

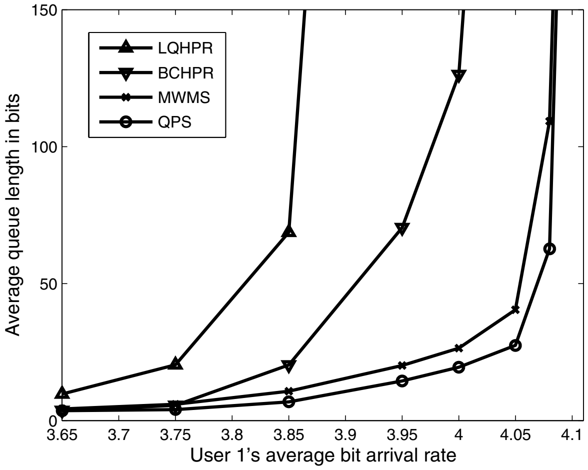


Fig. 3. Average queue length vs user 1's bit arrival rate under LQHPR, BCHPR, MWMS and QPS (2 users, user 1's SNR=13dB and user 2's SNR=7dB, $\lambda_2 = 0.5\lambda_1$).

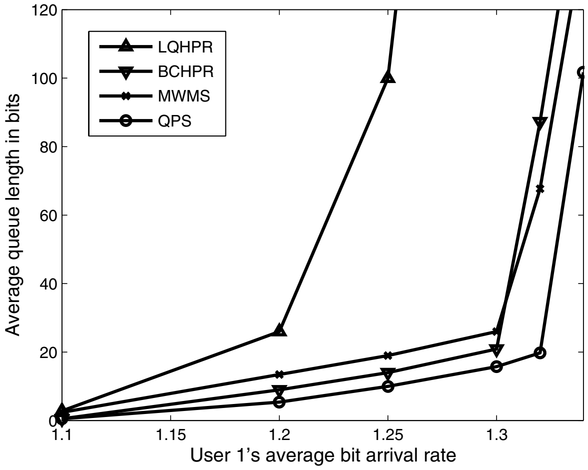


Fig. 4. Average queue length vs user 1's bit arrival rate under LQHPR, BCHPR, MWMS and QPS (10 users, user i's channel gain $h_i = 2-0.1(i-1)$ and $\lambda_i = \lambda_1(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

VII. Conclusion

In Gaussian broadcast channels, Queue Proportional Scheduling (QPS) is shown to have desirable delay and fairness properties. QPS achieves throughput optimality by allocating a maximum data rate vector that is proportional to the queue state vector. In addition, this policy can arbitrarily control the ratio of each user's average queuing delay in order to satisfy each user's different QoS. Numerical results demon-

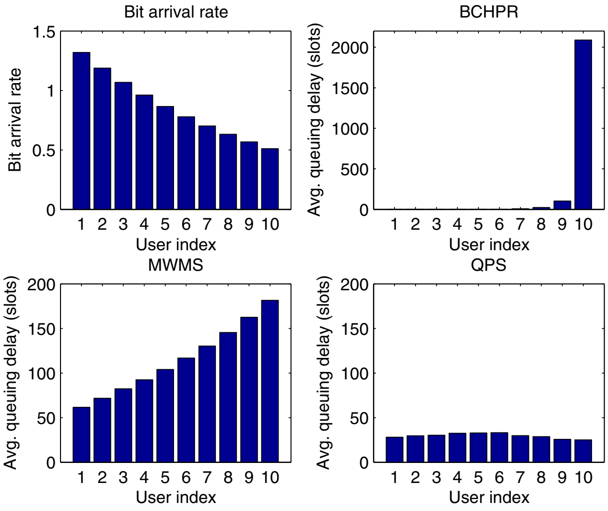


Fig. 5. Each user's bit arrival rate and each user's average queuing delay under BCHPR, MWMS and QPS (10 users, user i's channel gain $h_i = 2 - 0.1(i-1)$ and $\lambda_i = 1.32(0.9)^{i-1}$ for $i = 1, \cdots, 10$).

strate that QPS provides significantly smaller average queuing delay compared to Maximum Weight Matching Scheduling (MWMS) in a Gaussian BC.

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. Networking*, vol. 11, no. 1, pp 138-152, Feb. 2003.

[2] R. Barry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, pp 59 - 68, Sept. 2004.

[3] H. Viswanathan and K. Kumaran, "Rate scheduling in multiple antenna downlink," in *Proc. 39th Annual Allerton Conf. Communumications, Control and Computing*, Allerton, IL, Oct. 2001, pp. 747-756.

[4] C. Swannack, E. Uysal-Biyikoglu, and G. Wornell, "Low complexity multiuser scheduling for maximizing throughput in the MIMO broadcast channel," *Proc. 42nd Annual Allerton Conf. Communumications, Control and Computing*, Allerton, IL, Oct. 2004.

[5] L. Tassiulas and A. Ephremides, "Dynamic server allocation to parallel queues with randomly varying connectivity," *IEEE Trans. Info. Theory*, vol. 39, Mar. 1993, pp. 466-78.

[6] N. McKeown, A. Mekkittikul, V. Anantharam, and J. Walrand, "Achieving 100% throughput in an input-queued switch," in *Proc. IEEE INFOCOM '96*, San Francisco, CA, pp. 296-302.

[7] 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 Communications and Networking Conference (WCNC)*, Atlanta, Georgia, 2004.

[8] E. Yeh and A. Cohen, "Throughput and delay optimal resource allocation in multi-access fading channels," in *Proceedings of the International Symposium on Information Theory*, Yokohama, Japan, p. 245, 2003.

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

[10] A. Eryilmaz, R. Srikant, and J.R. Perkins, "Throughput-optimal scheduling for broadcast channels," in *Proc. of SPIE*, vol. 4531, pp. 70-78, 2001.

[11] R. Leelahakriengkrai and R. Agrawal, "Scheduling in multimedia wireless networks," in *Proc. 17th Int. Teletraffic Congress*, Salvador da Bahia, Brazil, December 2001.

[12] T. Cover and J. Thomas, *Elements of Information Theory*, New York: John Wiley and Sons Inc., 1997.

[13] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2003.

[14] K. Seong, R. Narasimhan, and J.M. Cioffi, "Queue proportional scheduling via geometric programming in fading broadcast channels," *IEEE J. Select. Areas Comm.*, vol. 24, no. 8, Aug. 2006.

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

[16] S. Ross, *Stochastic Processes*, John Wiley and Sons Inc., 1996.

[17] D. Bertsekas and R. Gallager, *Data Networks*. NJ: Prentice Hall, 1992.

[18] E. Yeh and A. Cohen, "Information theory, queueing, and resource allocation in multi-user fading communications," in *Proceedings of the 2004 Conference on Information Sciences and Systems*, Princeton, NJ, 2004, pp. 1396-1401.