

# Efficient Scheduling for Heterogeneous Traffic in Downlink

Rajiv Agarwal, Vinay Majjigi, Rath Vannithamby and John Cioffi

**Abstract**—This paper studies optimal allocation of resources to users in a fading broadcast channel (BC) to support heterogeneous traffic. Resource allocation is done in a cross-layer manner for both deadline-sensitive (DS) and best-effort (BE) data to provide users guaranteed Quality of Service (QoS). Given the user queue-states, the proposed algorithm allocates the minimum resource to ‘just’ meet the deadlines for DS packets a fixed % of time. The remaining resources are then optimally split between the users for their BE data to maximize sum-rate, minimize buffer holding cost or any general utility function. The resource allocation problem is formulated as a linear program (LP) followed by a geometric program (GP) which can be solved very efficiently. Simulation results are presented for Poisson packet arrivals and exponentially distributed packet lengths on a Rayleigh fading channel for both the DS and BE traffic.

## I. INTRODUCTION

Next generation wireless cellular networks will support a variety of quality-of-service (QoS) sensitive applications like streaming multimedia and high-speed data for downlink users. Channel and QoS aware downlink packet scheduling policies are going to play a key role in utilizing system resources in an efficient manner. Schedulers specifically designed to support non-real-time delay-tolerant services perform poorly for real-time delay-sensitive services, and vice-versa. Hence, one needs to design scheduling schemes that can jointly allocate resources to both real-time and non-real-time traffic. Further, this needs to be done in a cross-layer manner by looking at both the users’ channel conditions and their backlogged packets.

Many cross-layer approaches to resource allocation have been recently proposed in [1], [2], [3] and the references therein. These works show that consideration of both channel and queue states allows the entire network capacity region to be achieved in fading broadcast channels (BC) and multiple-access channels (MAC). Recently, in [4], the authors studied cross-layer allocation of resources for heterogeneous traffic using a dynamic programming formulation. The scheduler, which is the base station (BS) in the downlink, has access to instantaneous channel gains for all users and at any time instant chooses the best strategy that minimizes expected cost-to-go. The cost is a function of the penalty for missing a deadline for DS data and buffer-holding cost for BE data.

As noted in [5], [6], if along with instantaneous channel gains, the long-term channel statistics are also available at the BS, system performance can be improved. In [6], two resource allocation schemes were formulated as convex optimization problems - Queue-Proportional Scheduling (QPS) and Maximum-Weight-Matching Scheduling (MWMS) utilizing both the instantaneous and long-term channel statistics. However, [6] does not address the issue of DS traffic and packet-drop rates. Thus the resource-allocation strategies dis-

cussed in [6] like QPS, MWMS, are more suited for BE traffic where strict packet-deadlines are not enforced.

This paper, combines the concepts in [4] and [6] and extends their results to handle both DS and BE traffic by utilizing both instantaneous and long-term channel-state knowledge at the BS. The available resources are first optimally allocated to ‘just’ meet QoS requirements for real-time traffic; this problem is formulated as a Linear Program (LP). The remaining resources are optimally used for non-real-time traffic in a manner same as in [6] with one change; in [6], transmit power is always fixed to be a constant in the Geometric Program (GP) formulation. We extend this to handle both average and peak power constraints and show that the resource allocation can still be formulated as a GP. Since, the entire problem is cast as a convex optimization problem that involves solving an LP and a GP; the allocation of resources can be done in an efficient manner. Performance results show that the proposed scheme can meet delay guarantees and provide good service to delay-tolerant data.

## II. SYSTEM MODEL AND DEFINITIONS

In this paper, a discrete-time fading channel is assumed where the channel-state is constant over a scheduling instant and changes in an independent and identically distributed (i.i.d.) manner. The BS knows both the instantaneous channel-states, which remain constant over a scheduling instant, and the long-term channel-state statistics, which remain constant over a block of scheduling instants, for all users. We consider a flat-fading Gaussian broadcast channel with a single transmitter sending independent messages to  $K$  users over two-sided bandwidth  $2W$ . At  $t^{\text{th}}$  scheduling instant, the received signal of user  $k$  is expressed as

$$Y_k(t) = h_k(t)X(t) + z_k(t), k = 1, \dots, K \quad (1)$$

where the transmitted signal  $X(t)$  is composed of  $K$  independent messages, the complex channel gain of user  $k$  is denoted by  $h_k(t)$ , and  $z_k(t)$ 's are i.i.d. zero-mean Gaussian noise with power  $N_0W$ . As in [7], the channel gain can be combined with the noise component by defining an effective noise  $\tilde{z}_k(t) = z_k(t)/h_k(t)$ . Then, the equivalent received signal is given by

$$\tilde{Y}_k(t) = X(t) + \tilde{z}_k(t), k = 1, \dots, K \quad (2)$$

where the power of  $\tilde{z}_k(t)$  conditioned on the channel gain is defined as  $n_k(t) = N_0W/|h_k(t)|^2$ . Without loss of generality,  $W = 1$  is assumed throughout this paper for simplicity. The effective noise power  $\mathbf{n}(t) = [n_1(t) n_2(t) \dots n_K(t)]^T$  is considered as the denotation of a channel-state (equivalently also called noise-state) and the vector  $\mathbf{n}(t)$  is assumed to take one of  $|\mathcal{N}|$  possible values from the set  $\mathcal{N}$  with some

probabilities. Thus the joint distribution of the noise-states is assumed to be discrete. In general, the BS can have both a block (average) power constraint  $\bar{P}$  i.e.  $\mathbb{E} [\|X(t)\|^2] \stackrel{(a)}{=} \frac{1}{T} \sum_{t=1}^T \|X(t)\|^2 \leq \bar{P}$ , where, each scheduling block consists of  $T$  scheduling instants<sup>1,2</sup>, and an instantaneous (peak) power constraint  $P^{\text{peak}}$  i.e.  $\|X(t)\|^2 \leq P^{\text{peak}}, \forall t$ .

Each user is assumed to have both real-time and non-real-time data to be transmitted. For the DS data of each user  $k$ , the BS maintains a constant transmission rate  $r_k$  irrespective of the channel realization, as long as the channel-gain is above a threshold. When the channel-gain is below the threshold, which happens with some probability  $\text{Pr}_k^{\text{req}}$  depending on the channel-gain distribution, an outage is declared for the user and no transmission takes place for the particular user. Thus for the DS data, rate-tuples are assigned from the outage-capacity or delay-limited capacity region<sup>3</sup> [7] as the coded packet needs to be decoded immediately at the end of transmission. The constant transmission rate  $r_k$  is chosen based on the packet-arrival process, the deadline  $D_k$  associated with the application and the QoS guarantee  $q_k$  the user is assigned. Typical values of  $q_k$  can be 0.9 or 0.95. Since, packet-delay is a function of the packet arrival-process and the constant transmission rate  $r_k$  with outage probability  $\text{Pr}_k^{\text{req}}$ , mathematically, the following relationship must hold for guaranteeing QoS to user  $k$ :

$$\text{Prob}(\text{Packet-delay}(\text{Arrival-Process}_k, r_k, \text{Pr}_k^{\text{req}}) > D_k) < 1 - q_k \quad (3)$$

In general, for each user given its packet-arrival process distribution and application deadline, one can calculate  $\text{Pr}_k^{\text{req}}$  and  $r_k$  for a given  $q_k$  using Eqn (3). Thus for each user,  $\text{Pr}_k^{\text{req}}$  and  $r_k$  are computed and are fixed for the scheduling block. Let  $\mathbf{R}_c = \{r_1, r_2, \dots, r_K\}$  and  $\mathbf{Pr}^{\text{req}} = \{\text{Pr}_1^{\text{req}}, \text{Pr}_2^{\text{req}}, \dots, \text{Pr}_K^{\text{req}}\}$ . In the remaining discussion, we assume that  $\mathbf{R}_c$  and  $\mathbf{Pr}^{\text{req}}$  are given.

For the BE data of each user  $k$ , the BS uses  $B_K$ , the number of bits waiting to be sent in the BE queue of user  $k$ , to compute the total rate assignment  $R_k$  for the user (the rate assigned to BE traffic will be  $R_k - r_k$ ). There are no deadlines for the BE packets. Thus for BE data, ergodic rate-tuples are chosen from the minimum-rate *outage* capacity region [8]. The actual rate assigned to the BE data depends on the BE metric being used by the scheduler (BS) like MWMS or QPS. For each user,  $R_k$  is computed and is fixed for the scheduling block.

### III. PROPOSED SCHEDULING SCHEME

As mentioned in the previous section, two different capacity measures for the fading BC will be useful in the description

<sup>1</sup>In practice, the value of  $T$  is determined from the long-term noise-state statistics and the affordable complexity at the BS because rate allocations for the users need to be updated every scheduling block, as described later.

<sup>2</sup>Since the fading is assumed to be ergodic time-averages are equivalent to expectation over the distribution, and so (a) follows.

<sup>3</sup>In [6], the rate-pairs are allocated from the ergodic capacity region, which cannot be used for DS traffic because rates in the ergodic capacity region are achieved by using codewords that must span all possible noise state realizations before they can be decoded.

of the proposed scheduling algorithm. The first is the outage capacity region of the BC with certain minimum outage probabilities for the users defined in [7]. The second is minimum-rate *outage* capacity as defined in [8, §VII]. The authors in [8], already allude to the use of minimum-rate *outage* capacity in applications where there is both real-time and non-real-time traffic. However, in [8], [9] and [7], the optimal resource allocation problem is solved using Lagrange multiplier techniques, a technique that is generic to any convex optimization problem. As seen in this section, this paper formulates the same problem as a LP followed by a GP. The advantage of reducing a convex optimization problem in standard form (LP, GP, SOCP, etc.) is that any off-the-shelf convex optimization solver like MATLAB or CVX can be used to solve the problem. More discussion on this can be found in [6] and [10].

Given the  $\text{Pr}_k$  and  $r_k$  pairs specified for each user, we first describe resource allocation for the DS traffic. We use the same notation as in [7] and drop the time parameter  $t$  using the fact that the distribution of noise-states is stationary and ergodic. For any given noise state  $\mathbf{n}$ , there are  $\sum_{k=1}^K \binom{K}{k} = 2^K - 1$  different combinations of the  $K$  users that may be chosen to transmit over the channel. Let the  $(2^K - 1)$  possible combinations of users be represented as vector  $[\psi(i, 1), \psi(i, 2), \dots, \psi(i, K)]$ ,  $i = 1, 2, \dots, (2^K - 1)$ . For each value of  $i$  in the vector  $[\psi(i, 1), \psi(i, 2), \dots, \psi(i, K)]$ , if  $\psi(i, k) = 1$ , then user  $k$  is there in the  $i^{\text{th}}$  combination, else not. For each  $i$ , the set of active users  $\mathcal{U}_i$  is defined as

$$\mathcal{U}_i = \{k : \psi(i, k) = 1, 1 \leq k \leq K\}$$

$\mathcal{U}_0$  denoted the empty set (no active users). For any noise state  $\mathbf{n}$ , suppose that the scheduler only transmits information to the users in the active set  $\mathcal{U}_i$ . Then for a given active user set  $\mathcal{U}_i$ , we can compute the minimum total power  $P_i^{\text{min}}(\mathbf{R}_c, \mathbf{n})$  required to support a subvector of  $\mathbf{R}_c$  composed of the required rates of the users in  $\mathcal{U}_i$  by first arranging the users in  $\mathcal{U}_i$  in the order of their noise powers i.e.  $n_{\pi(1)} \leq n_{\pi(2)} \dots \leq n_{\pi(M_i)}$ , where  $M_i = |\mathcal{U}_i|$ . Then  $P_i^{\text{min}}(\mathbf{R}_c, \mathbf{n}) = \sum_{m=1}^{M_i} \left[ 2^{\sum_{j=m+1}^{M_i} r_{\pi(j)}} (2^{r_{\pi(m)}} - 1) n_{\pi(m)} \right] + (2^{r_{\pi(M_i)}} - 1) n_{\pi(M_i)}$ . For the noise state  $\mathbf{n}$ , let  $w_i(\mathbf{R}_c, \mathbf{n})$  denote the probability that the scheduler transmits information to the users in the set  $\mathcal{U}_i$ . Then

$$\sum_{i=1}^{2^K-1} w_i(\mathbf{R}_c, \mathbf{n}) = 1 - w_0(\mathbf{R}_c, \mathbf{n}) \leq 1.$$

For a given rate vector  $\mathbf{R}_c$  and a noise state  $\mathbf{n}$ , let  $\text{Pr}_k^{\text{on}}(\mathbf{R}_c, \mathbf{n})$  be the probability that information is sent to user  $k$  i.e.

$$\text{Pr}_k^{\text{on}}(\mathbf{R}_c, \mathbf{n}) = \sum_{i=1}^{2^K-1} w_i(\mathbf{R}_c, \mathbf{n}) \psi(i, k) \quad (4)$$

Then the average outage probability  $\text{Pr}_k(\mathbf{R}_c)$  of user  $k$  ( $1 \leq k \leq K$ ) is

$$\begin{aligned} \text{Pr}_k(\mathbf{R}_c) &= 1 - \text{Pr}_k^{\text{on}}(\mathbf{R}_c) \\ &= 1 - \mathbb{E}_{\mathbf{n}} [\text{Pr}_k^{\text{on}}(\mathbf{R}_c, \mathbf{n})] \end{aligned} \quad (5)$$

which has to be less than the specified  $\Pr_k^{\text{req}}$  for guaranteeing QoS to user  $k$ .  $\mathbb{E}_{\mathbf{n}}$  denotes expectation w.r.t. the discrete distribution of the noise states  $\mathbf{n}$ .

Since in each noise state  $\mathbf{n}$ , the total required minimum power to support  $\mathbf{R}_c$  with non-outage probability  $\Pr_k^{\text{on}}(\mathbf{R}_c, \mathbf{n})$  for user  $k$  is

$$P^{\min}(\mathbf{R}_c, \mathbf{n}) = \sum_{i=1}^{2^K-1} P_i^{\min}(\mathbf{R}_c, \mathbf{n}) w_i(\mathbf{R}_c, \mathbf{n}) \quad (6)$$

where  $w_i(\mathbf{R}_c, \mathbf{n})$  is set to 0, if  $P_i^{\min}(\mathbf{R}_c, \mathbf{n}) > P^{\text{peak}}$ ; the total required minimum average power to support  $\mathbf{R}_c$  will be  $\bar{P}^{\min}(\mathbf{R}_c) = \mathbb{E}_{\mathbf{n}} [P^{\min}(\mathbf{R}_c, \mathbf{n})]$ .

Thus, for a given rate vector  $\mathbf{R}_c$ , we wish to solve the following minimization problem

$$\begin{aligned} & \text{minimize} && s && \text{s. t.} \\ & \mathbb{E}_{\mathbf{n}} [P^{\min}(\mathbf{R}_c, \mathbf{n})] && \leq && s \\ & \sum_{i=1}^{2^K-1} w_i(\mathbf{R}_c, \mathbf{n}) && \leq && 1 \quad \forall \mathbf{n} \\ & 0 \leq w_i(\mathbf{R}_c, \mathbf{n}) && \leq && 1 \quad \forall i, \forall \mathbf{n} \\ & \Pr_k(\mathbf{R}_c) && \leq && \Pr_k^{\text{req}} \quad \forall k \end{aligned} \quad (7)$$

The above problem is an LP in the variables<sup>4</sup>  $\{s, w_i(\mathbf{R}_c, \mathbf{n})\}$ . Knowing the values of  $w_i^*(\mathbf{R}_c, \mathbf{n})$  from the solution of the above problem, we find the optimal  $P_k^{\min*}(\mathbf{R}_c, \mathbf{n})$  and  $r_k^*(\mathbf{n})$ , the minimum power<sup>5</sup> and rate allocated to user  $k$  in each noise state  $\mathbf{n}$ .

Once  $\bar{P}^{\min}(\mathbf{R}_c)$  has been calculated, the remaining power  $\bar{P}^{\text{rem}} = \bar{P} - \bar{P}^{\min}(\mathbf{R}_c)$  is used for the BE traffic. For this we need to evaluate a rate tuple on the boundary of the minimum-rate outage capacity region,  $\mathcal{C}_{\min}(\bar{P}, \mathbf{R}_c, \mathbf{Pr}^{\text{req}})$ , which optimizes some utility criterion. It was shown in [6], that four different utility functions namely (a) QPS, (b) MWMS, (c) Best-Channel Highest-Possible Rate (BCHPR) and (d) Longest-Queue Highest-Possible Rate (LQHPR), can be formulated as a GP for the ergodic capacity region with constant power transmission. In this paper, we show that each of these policies can be implemented for the minimum-rate outage capacity region as a GP with both a peak and average power constraint. Explicit characterization is given only for weighted sum-rate (special cases of which are MWMS and BCHPR) due to space constraints and the formulation for (a) and (d) can be obtained in a similar manner using [6].

Using [8], total power allocated for each user  $k$  in noise state  $\mathbf{n}$  is denoted as  $P_k(\mathbf{n}) = P_k^{\min*}(\mathbf{R}_c, \mathbf{n}) + \hat{P}_k(\mathbf{n})$ , where  $\hat{P}_k(\mathbf{n})$  is the power given to user  $k$  in noise state  $\mathbf{n}$  for BE

<sup>4</sup>The total number of variables is at most  $(|\mathcal{N}|(2^K - 1) + 1)$ , where  $|\mathcal{N}|$  is the number of possible noise states. We have all the variables in the minimization if the peak power constraint  $P^{\text{peak}}$  is inactive or equivalently is always satisfied.

<sup>5</sup>Notice that  $P^{\min}(\mathbf{R}_c, \mathbf{n})$  with the subscript  $i$  denotes minimum power needed in the configuration  $i$  (one of  $(2^K - 1)$  possible configurations), whereas  $P^{\min}(\mathbf{R}_c, \mathbf{n})$  with the subscript  $k$  denotes the minimum power assigned to user  $k$  (one of  $K$  possible users).

traffic, a quantity that needs to be determined. A point on the boundary of  $\mathcal{C}_{\min}(\bar{P}, \mathbf{R}_c, \mathbf{Pr}^{\text{req}})$  for a given<sup>6</sup>  $\boldsymbol{\mu}$  is

$$\begin{aligned} & \max_{\mathbf{P}^e(\mathbf{n})} \mathbb{E}_{\mathbf{n}} \left[ \sum_{k=1}^K \mu_k \left[ r_k + \log_2 \left( 1 + \frac{P_{\pi(k)}^e(\mathbf{n})}{n_{\pi(k)}^e + \sum_{j < k} P_{\pi(j)}^e(\mathbf{n})} \right) \right] \right] \\ & \text{s.t.} : \mathbb{E}_{\mathbf{n}} \left[ \sum_{k=1}^K P_k^e(\mathbf{n}) \right] \leq \bar{P}^{\text{rem}} \\ & P_k^e(\mathbf{n}) \geq 0 \quad \forall k, \mathbf{n} \end{aligned} \quad (8)$$

where for every noise state  $\mathbf{n}$ ,  $\pi(\cdot)$  is the permutation s.t.  $n_{\pi(1)} \leq n_{\pi(2)} \leq \dots \leq n_{\pi(K)}$  and  $n_{\pi(k)}^e$  and  $P_{\pi(k)}^e(\mathbf{n})$  are defined as

$$n_{\pi(k)}^e \triangleq \left( n_{\pi(k)} + \sum_{j \leq k} P_{\pi(j)}^{\min*}(\mathbf{R}_c, \mathbf{n}) \right) 2^{\sum_{j > k} r_{\pi(j)}^*(\mathbf{n})} \quad (9)$$

$$\begin{aligned} P_{\pi(k)}^e(\mathbf{n}) & \triangleq \left( \hat{P}_{\pi(k)}(\mathbf{n}) - \left( 2^{r_{\pi(k)}^*(\mathbf{n})} - 1 \right) \sum_{j < k} \hat{P}_{\pi(j)}(\mathbf{n}) \right) \\ & \times 2^{\sum_{j > k} r_{\pi(j)}^*(\mathbf{n})} \end{aligned} \quad (10)$$

Further, it was shown in [8], that the ordering of the effective noises  $n_{\pi(k)}^e$ 's is the same as the ordering of the actual noises  $n_{\pi(k)}$ 's.

Given  $P_k^{\min*}(\mathbf{R}_c, \mathbf{n})$  and  $r_k^*(\mathbf{n})$  from the solution of the LP, we compute  $n_{\pi(k)}^e$  using (9) and then solve the GP based on any of the four utility criterion. We now formulate weighted sum-rate maximization for the BE data as a GP. For readability of equations, symbols  $n_{\pi(k)}^e$ ,  $P^e(\mathbf{n})$  and  $\bar{P}^{\text{rem}}$  in the minimum-rate outage capacity region are replaced by  $n_{\pi(k)}$ ,  $P(\mathbf{n})$  and  $\bar{P}$  respectively in the following GP formulation.

$$\begin{aligned} (O) & \text{ minimize } \log e^{-\sum_{k=1}^K g_k \mathbb{E}_{\mathbf{n}} [R_k(\mathbf{n})]} \\ (1) & \log e^{-R_k(\mathbf{n})} \leq 0, \quad \forall k = 1, \dots, K, \quad \forall \mathbf{n} \in \mathcal{N} \\ (2) & \log \sum_{k=1}^K e^{\log t_k(\mathbf{n})} e^{-v(\mathbf{n})} e^{(\ln 2) \sum_{j=k}^K R_{\pi(j)}(\mathbf{n})} \leq 0, \quad \forall \mathbf{n} \\ (W) & t_k(\mathbf{n}) \triangleq n_{\pi(k)} - n_{\pi(k-1)}; \quad v(\mathbf{n}) \triangleq \log (P(\mathbf{n}) + n_{\pi(K)}) \\ (3) & P(\mathbf{n}) \geq 0 \Rightarrow \log n_{\pi(K)} e^{-v(\mathbf{n})} \leq 0, \quad \forall \mathbf{n} \\ (4) & P(\mathbf{n}) \leq P^{\text{peak}} \Rightarrow \log (n_{\pi(K)} + P^{\text{peak}}) e^{v(\mathbf{n})} \leq 0, \quad \forall \mathbf{n} \\ (5) & \mathbb{E}_{\mathbf{n}} [P(\mathbf{n})] \leq \bar{P} \Rightarrow \log \frac{1}{\bar{P} + \mathbb{E}_{\mathbf{n}} [n_{\pi(K)}]} \mathbb{E}_{\mathbf{n}} [e^{v(\mathbf{n})}] \leq 0 \end{aligned} \quad (11)$$

where (O) is the objective function, (1, 2, 3, 4, 5) are the constraints and (W) is definition of the dummy variables. Our optimization variables are  $\{s, \{R_k(\mathbf{n})\}, \{v(\mathbf{n})\}\}$  and there is a one-to-one relationship between the  $P(\mathbf{n})$  (the total power used in a noise state i.e.  $P(\mathbf{n}) = \sum_{k=1}^K P_k(\mathbf{n})$ ) and the  $v(\mathbf{n})$ .

<sup>6</sup>The  $\boldsymbol{\mu}$  vector s.t.  $0 \leq \mu_k \leq 1$ ,  $\sum_{k=1}^K \mu_k = 1$  is a standard technique for representing points on the boundary of any BC capacity region as the region is convex [7], e.g. for two users  $\boldsymbol{\mu} = [0.5 \ 0.5]$  denotes the vector to maximize sum-rate.

The total number of variables is  $(1 + |\mathcal{N}|K + |\mathcal{N}|)$ . The value  $g_k$  specifies the importance of user  $k$  or its weight and is given.

The solution of the GP gives  $\mathbf{P}^e(\mathbf{n}) = [P_1^e(\mathbf{n}), P_2^e(\mathbf{n}), \dots, P_K^e(\mathbf{n})]$ , which is used to solve the system of simultaneous linear equations in (10) to compute  $\hat{\mathbf{P}}(\mathbf{n}) = [\hat{P}_1(\mathbf{n}), \hat{P}_2(\mathbf{n}), \dots, \hat{P}_K(\mathbf{n})]$  and thus find the actual rates  $(R_k - r_k)$  for BE traffic of each user, where

$$R_{\pi(k)} = \mathbb{E}_{\mathbf{n}} \left[ \log_2 \left( 1 + \frac{P_{\pi(k)}^{\min*}(\mathbf{R}_c, \mathbf{n}) + \hat{P}_{\pi(k)}(\mathbf{n})}{n_{\pi(k)} + \sum_{j < k} P_{\pi(j)}(\mathbf{n})} \right) \right] \quad (12)$$

#### IV. NUMERICAL RESULTS

This section presents simulation results for 2 users with Poisson packet arrivals (mean  $a_k$ ,  $k = 1, 2$ ) and exponentially distributed packet lengths (mean  $\nu_k$ ,  $k = 1, 2$ ).  $\bar{P}$  and  $P^{\text{peak}}$  are set to 20dB and 40dB respectively. The channel fading distribution for each user is i.i.d. Rayleigh fading with mean  $\bar{\gamma}$ , quantized to 4 levels where quantization is done by dividing the channel gain pdf into 4 regions of equal probability. We look at quantized channel state feedback because in practice receivers can feedback their channel gain values only with finite number of bits, in this case 2. The non-outage probability is set to  $\mathbf{Pr}^{\text{req}} = [1 \ 1]$ . The deadline for DS packets is set to  $D_k = 20$ ,  $k = 1, 2$  ms for all DS packets as also assumed in [4]. Also, QoS requirements for the two users are set to  $q_1 = 0.97$ ,  $q_2 = 0.95$ . The coherence time of the channel is set to 5 ms, which is also the symbol (transmission time-slot) duration.

In Figures 1-2, we set  $\bar{\gamma}_1 = 6\text{dB}$ ,  $\bar{\gamma}_2 = 5.5\text{dB}$ ,  $a_1^{\text{be}} = a_1^{\text{ds}} = 0.75$  packets per time-slot,  $a_2^{\text{be}} = a_2^{\text{ds}} = 0.375$  packets per time-slot,  $\nu_1^{\text{ds}} = \nu_2^{\text{ds}} = 2$  bits per packet and vary  $\nu^{\text{be}} = \nu_1^{\text{be}} = \nu_2^{\text{be}}$  from 10 to 50 bits per packet. For the given parameter values, the required rates  $r_k$  for DS traffic can be computed using Eqn (3), [11] to be  $r_1 = 1.8507$ ,  $r_2 = 1.0496$  bits per time-slot. Using the weighted sum-rate formulation given in this paper, we compare two schemes, one that maximizes - (a) sum-rate for the BE users and (b) weighted sum-rate for the BE users, where the weights are given by their  $a_k^{\text{be}}$ 's. Hence, user 1 that has double the expected packet-arrival rate is weighted more. The former scheme is the same as BCHPR. The latter scheme is similar to MWMS because queue-states at any instant depend on the expected packet-arrival rate.

Fig. 1 plots the average queuing delay (in multiples of time-slot), denoted by  $\bar{\tau}_k$ ,  $k = 1, 2$ , as a function of the expected packet-arrival rate for BE traffic  $\nu^{\text{be}}$ . As seen from the plot, weighted sum-rate similar to MWMS results in much lower average delay for the better user at the expense of a higher average delay for the worse user. A similar observation was made in [6] in the context of MWMS and only an instantaneous power constraint. Hence Scheme (b) does not ensure fairness to the users. In comparison, Scheme (a) has worse average delay for user 1, however better fairness. Fig. 2 plots the average queue length (in number of packets), denoted by  $\bar{Q}_k$ ,  $k = 1, 2$ , as a function of  $\nu^{\text{be}}$ . As seen from the plot, weighted sum-rate has desirable fairness properties in terms of average queue length, also observed in [6].

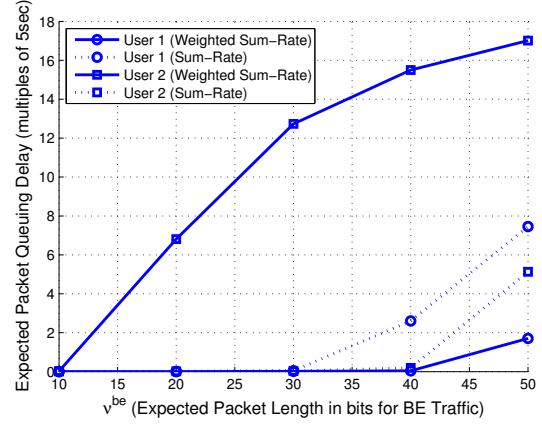


Fig. 1.  $\bar{\tau}_k$  v/s  $\nu^{\text{be}}$  comparison for BE traffic.

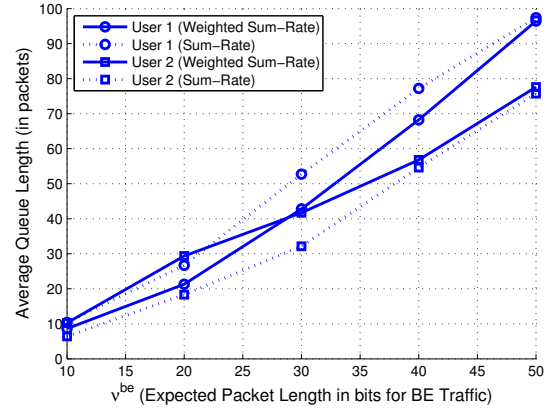


Fig. 2.  $\bar{Q}_k$  v/s  $\nu^{\text{be}}$  comparison for BE traffic.

If Fig. 3, we compare the trade-off between packet-drop-rate for DS traffic and average-queuing-delay for BE traffic for two schemes - the scheme in [4] and the proposed scheme for a single user. The simulation parameters used are  $a^{\text{be}} = a^{\text{ds}} = 0.75$ ,  $\nu^{\text{be}} = \nu^{\text{ds}} = 7$ , packet-deadline = 20 ms, time-slot duration = 5ms and continuous Rayleigh fading with  $\bar{\gamma} = 0$  dB.

For the scheme in [4], a packet transmission is attempted every 5 ms, hence each packet has 4 possible attempts before it is dropped. At any instant, transmission is done with average power  $\bar{P} = 20$  dB, hence the two schemes have the same average power consumption, so  $SNR = 20$  dB. The probability of success for a packet is computed by finding if the channel went in outage. Specifically, if a packet has  $b$  bits, the probability of success for this packet is given as  $e^{-\frac{1}{SNR} \frac{2^b - 1}{\bar{\gamma}}}$ . The trade-off curve is traced by keeping buffer holding cost  $c$  fixed at 1 and changing the packet-drop-penalty  $\lambda$  from 1 through 100.

For our proposed scheme,  $\bar{P}$  and  $P^{\text{peak}}$  are set to 20dB and 40dB respectively. The channel fading distribution is i.i.d. Rayleigh fading with same mean  $\bar{\gamma} = 0$  dB, quantized to 4

levels where quantization is done by dividing the channel gain pdf into 4 regions of equal probability. The trade-off curve is traced by increasing the rate requirement  $r$  for DS traffic from 0.2 to 5 bits per time-slot.

As seen from Fig. 3, the proposed scheme has two advantages over the scheme in [4]. It achieves a better trade-off and the variations in average-packet-delay with changing average packet-drop-rate are minimal. As discussed before, this is achieved by utilizing the long-term channel statistics, which the proposed scheme uses and the scheme in [4] does not.

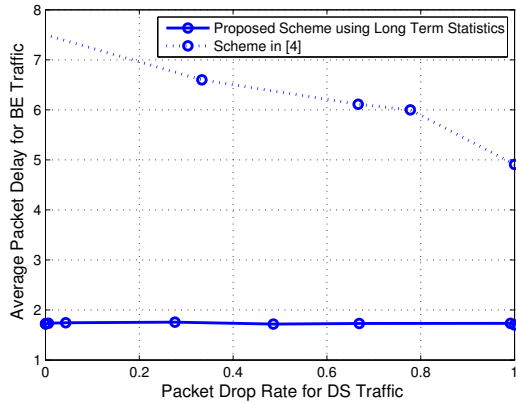


Fig. 3. Performance trade-off between BE and DS traffic.

## V. CONCLUSION

We proposed a scheduling algorithm for fading broadcast channels which always guarantees QoS for deadline-sensitive traffic when not in outage at the same time allocates remaining resources optimally for the BE traffic by utilizing both instantaneous and long-term channel statistics. The minimum power required to meet deadlines was formulated as a Linear Program and the optimal allocation of remaining power to the best-effort traffic was formulated as a Geometric Program. Hence, the resource allocation problem was cast in a convex optimization framework that can be solved efficiently. For a BS with fixed complexity, this means that resource allocation problem can be solved more often to handle fast channel variations; thus always meeting QoS guarantees. It was shown numerically that the proposed scheme achieves a better trade-off between DS and BE traffic as compared to the scheme in [4].

## REFERENCES

- [1] R. Barry and E. Yeh, "Cross-layer wireless resource allocation," *IEEE Signal Processing Magazine*, pp. 59–68, Sept 2004.
- [2] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, and P. Whiting, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, pp. 150–154, Feb 2001.
- [3] C. Swannack, E. Uysal-Biyikoglu, and G. Wornell, "Low complexity multi-user scheduling for maximizing throughput in the mimo broadcast channel," *Allerton*, 2004.
- [4] A. Dua and N. Bambos, "Downlink scheduling of heterogeneous traffic," in *IEEE ICC*, Jun 2006.

- [5] K. Seong, R. Narasimhan, and J. Cioffi, "Cross-layer resource allocation via geometric programming in fading broadcast channels," *IEEE VTC Spring*, May 2006.
- [6] K. Seong, R. Narasimhan, and J. Cioffi, "Queue proportional scheduling via geometric programming in fading broadcast channels," *IEEE Jour. Sel. Areas in Comm.*, vol. 24, pp. 1593–1602, Aug 2006.
- [7] L. Li and A. Goldsmith, "Capacity and optimal resource allocation for fading broadcast channels: Part II outage capacity," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1103–1127, Mar 2001.
- [8] N. Jindal and A. Goldsmith, "Capacity and optimal power allocation for fading broadcast channels with minimum rates," in *Proc. IEEE Globecom*, pp. 1292–1296, Nov 2001.
- [9] A. G. and L. Li, "Capacity and optimal resource allocation for fading broadcast channels: Part I ergodic capacity," *IEEE Trans. Info. Th.*, vol. 47, pp. 1083–1102, Mar 2001.
- [10] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, March 2004.
- [11] D. P. Bertsekas and R. G. Gallager, *Data Networks*, 2nd Edition. Prentice Hall, 1991.