

Throughput-Optimized Opportunistic Scheduling for Rate-Guaranteed and Best-Effort Users in Wireless Networks

Wen Huang and Kwan L. Yeung
Department of Electrical and Electronic Engineering
The University of Hong Kong, Hong Kong, PRC.
Tel: (852) 2857-8493, Fax: (852) 2559-8738
Email: {whuang, kyeung}@eee.hku.hk

Abstract—We study opportunistic scheduling algorithms in a wireless network with a central base station communicating with multiple users on a single shared channel using time division multiple access. We assume the coexistence of both rate guaranteed (RG) users and best effort (BE) users in the system. A RG user asks for a specific transmission rate and the system should provision the exact amount being asked. In this paper, we present an optimal opportunistic scheduler that maximizes the system throughput. An analytical model is constructed to evaluate its performance in a homogenous system. Closed form solutions for channel access delay, worst case delay and system throughput are derived. Extensive simulations in both homogeneous and heterogeneous systems are conducted to verify the effectiveness of our scheduler.

Index Terms—Opportunistic scheduling, rate-guaranteed users, static scheduling policy, cross-layer optimization

I. INTRODUCTION

Opportunistic scheduling is a cross-layer approach which exploits the time-varying radio channel to improve the system performance of a wireless network. In a multi-user single-carrier system, different users communicate with the base station (BS) by transmitting/receiving on different time slots. The duration of a slot is given but a user can send data in different slots at different rates depending on its current channel condition. An opportunistic scheduler [1] is designed for selecting the user with the best channel condition to send (at the highest rate). With such a scheduler, the system throughput is maximized.

Other factors may also affect the scheduler decision. For example, if fairness in bandwidth allocation among different users is a concern, the simple scheduler in [1] may not apply. In [2], Liu, Chong and Shroff proposed a general utility maximization opportunistic scheduling framework. They further proposed an algorithm in [3] under this framework to maximize the long-term throughput while providing a fair allocation of the transmission time fraction among all the users. Also aiming at providing a fair allocation of transmission time fraction, Park, Seo, Kwon and Lee [4] proposed a cumulative distribution function based scheduling algorithm. The authors of [5] extended the framework proposed by Liu [2] to maximize the system throughput for the situation that each

user has both a minimal and maximal throughput constraint. Long and Feng proposed in [6] a scheduling scheme to provide the generalized-processor-sharing like service among users.

The authors of [7] extended the framework in [2] to a scenario where two types of users are considered, rate guaranteed (RG) users and best effort (BE) users. Each RG user has a minimal throughput requirement. The objective in the paper is to maximize the total fraction of time slots allocated to all BE users while satisfying the throughput requirements for the RG users. The optimal solution proposed in the paper shows that the scheduler should ensure *exact* rate requirements of all RG users with the *minimum* fraction of time slots. However, while maximizing the fraction of time slots for all BE users provides the flexibility on how to allocate the resources among BE users, it does not necessarily imply that the system throughput is maximized.

In this paper, we also consider opportunistic scheduling in a network with both RG and BE users. We propose an optimization problem to maximize the system throughput (whereas the total fraction of time slots allocated to BE users is maximized in [7]). Our contribution is twofold. First, we find the optimal solution for the optimization problem and a *throughput-optimal* scheduling algorithm is thus designed. Second, we analyze the performance of the scheduling algorithm in a homogeneous system. Closed form results for several important system metrics are derived including user channel access delay, total system throughput, and the maximal rate that can be supported given the number of RG users in the system. Insights are obtained on how to tune and set different system parameters to enhance the system performance. We detail the design of our optimal scheduler in Section II. In Section III, its performance in a homogeneous system is analyzed. Both numerical and simulation results are presented for performance evaluations in Section IV. We conclude the paper in Section V.

II. DESIGN OPTIMAL OPPORTUNISTIC SCHEDULER

A. System model

Without loss of generality, we focus on the time-slotted downlink transmission in a multi-user single-carrier system, where the duration of a slot is fixed but the BS can transmit

This work was supported by General Research Fund HKU 719108E.

at different rates depending on the user perceived channel quality. A saturated traffic condition, as used by [2] [3] [7], is assumed, where each user is always backlogged. Such a model is suitable for elastic traffic that can tolerate variable delays.

Define random variable vector $\vec{r} = \{r_1, \dots, r_N\}$, where r_i is a random variable representing the transmission rate of user i at a generic time slot and N is the total number of users in the system. Follow the notations in [3], we refer r_i as channel rate for user i and \vec{r} as the channel rate vector. We assume that each r_i is an independent stochastic random sequence whose probability distribution depends on its channel condition. At the beginning of a time slot, we assume each user can feedback its channel state information (CSI) to the BS in an error-free manner such that the BS knows the particular realization of the channel rate vector \vec{r} at that time slot. Then the BS can make scheduling decision based on such information. We focus on *static* scheduling policy which is a *mapping* from the channel rate vector space \vec{r} to the user index space $\{1, 2, \dots, N\}$. (i.e., the scheduling decision does not rely on the time index)

B. Problem formulation

Suppose there are m rate guaranteed (RG) users U_1, \dots, U_m and n best effort (BE) users U_{m+1}, \dots, U_{m+n} in the system, and $N = m + n$ is the total number of users. The guaranteed rate is a long term rate for every RG user, that is, the average transmission rate for the user. We define a guaranteed rate vector $\vec{c} = \{c_1, \dots, c_m\}$ for all the RG users, where c_i is the long term (average) rate requirement of RG user i . We further define 1_A as an indicator function of the event A .

The per time-slot throughput for any user i can then be represented as $r_i 1_{\{Q(\vec{r})=i\}}$. A scheduler should ensure that every RG user's average transmission rate satisfies the rate guaranteed constraint, i.e., $E(r_i 1_{\{Q(\vec{r})=i\}}) = c_i$ for $i = 1, \dots, m$.

Then the system throughput can be represented as:

$$E(\vec{r}_{Q(\vec{r})}) = E\left(\sum_{i=1}^N r_i 1_{\{Q(\vec{r})=i\}}\right) = \sum_{i=1}^N E(r_i 1_{\{Q(\vec{r})=i\}})$$

Unlike the scheduler in [7] which maximizes the total fraction of time slots allocated to all BE users, our design goal is to find a feasible static policy Q that maximizes the system throughput while providing just enough bandwidth to all RG users. (Note that under such objective, we are implicitly maximizing the throughput of the BE users.)

Let Θ denote the set of all feasible static policies. Then the scheduling problem is to find a feasible static policy that

$$\begin{aligned} & \operatorname{argmax}_{Q \in \Theta} E(\vec{r}_{Q(\vec{r})}) \\ & \text{s.t.} \quad E(r_i 1_{\{Q(\vec{r})=i\}}) = c_i, \text{ for } i = 1, \dots, m \end{aligned} \quad (1)$$

C. Optimal scheduling policy

If there exists a feasible solution for the problem in (1), i.e., any static scheduling policy Q satisfying the constraint that $E(r_i 1_{\{Q(\vec{r})=i\}}) = c_i$, for $i = 1, \dots, m$, then the optimal scheduling policy exists and it is unique. (The proof of the existence and uniqueness can be found by the convex property

of the channel rate region and is omitted here for brevity.) We define the optimal policy Q^* by:

$$Q^*(\vec{r}) = \operatorname{argmax}_i (\alpha_i^* r_i)$$

where α_i^* is the optimal weighting factor for user i , and its value is set according to the following three rules:

1. $\alpha_i^* > 0, \forall i$
2. $E(r_i 1_{\{Q(\vec{r})=i\}}) = c_i, i = 1, 2, \dots, m$
3. $\alpha_i^* = 1, i = m + 1, \dots, N$

At each decision point (i.e. at the beginning of a time slot), the scheduling policy Q^* calculates the weighted channel rate $\alpha_i^* r$ for each user, and chooses the one with the largest $\alpha_i^* r$ to send. (ties are broken randomly) For RG users, their α_i^* values are jointly determined by their rate requirement vector \vec{c} and the distribution of channel rate vector \vec{r} , as stipulated in rule 2 above. For BE users, their α_i^* values are simply set to 1, meaning that among all the BE users, the optimal scheduling scheme always prefers the BE user that can send at the highest channel rate in the current slot.

Proposition 1 (Optimal Scheduler). *The policy Q^* is the solution to the optimization problem (1). That is, it maximizes the system throughput while satisfying the RG constraints.*

Proof: Let Q be any feasible policy satisfying the constraint, i.e. $E(r_i 1_{\{Q(\vec{r})=i\}}) = c_i$. By the definition of Q^* , we know that Q^* is always "weighted optimal", i.e.

$$\sum_{i=1}^N \alpha_i^* r_i 1_{\{Q(\vec{r})=i\}} \leq \sum_{i=1}^N \alpha_i^* r_i 1_{\{Q^*(\vec{r})=i\}} \quad (2)$$

From the constraints in (1), we know that the throughput of RG user i equals to c_i . Rewriting (2) by decoupling the throughput contributions from RG users and BE users, we have:

$$\sum_{i=1}^m \alpha_i^* c_i + \sum_{i=m+1}^N r_i 1_{\{Q(\vec{r})=i\}} \leq \sum_{i=1}^m \alpha_i^* c_i + \sum_{i=m+1}^N r_i 1_{\{Q^*(\vec{r})=i\}} \quad (3)$$

From (3), we can manipulate the inequality by subtracting and adding the terms with c from both sides to prove the optimality:

$$\begin{aligned} (3) & \Rightarrow \sum_{i=m+1}^N r_i 1_{\{Q(\vec{r})=i\}} \leq \sum_{i=m+1}^N r_i 1_{\{Q^*(\vec{r})=i\}} \\ & \Rightarrow \sum_{i=1}^m c_i + \sum_{i=m+1}^N E(r_i 1_{\{Q(\vec{r})=i\}}) \leq \\ & \quad \sum_{i=1}^m c_i + \sum_{i=m+1}^N E(r_i 1_{\{Q^*(\vec{r})=i\}}) \\ & \Rightarrow \sum_{i=m+1}^N E(r_i 1_{\{Q(\vec{r})=i\}}) \leq \sum_{i=m+1}^N E(r_i 1_{\{Q^*(\vec{r})=i\}}) \\ & \Rightarrow E(\vec{r}_{Q(\vec{r})}) \leq E(\vec{r}_{Q^*(\vec{r})}) \end{aligned}$$

The proposition is proved. ■

D. Implementation

In the optimal scheduling policy designed above, the weighting factor vector is assumed to be known in advance. Indeed, its value is determined by the channel rate distribution and the guaranteed transmission rate information. In practice the weighting factors can be estimated using stochastic approximation methods as in [2]. In this work, we update the weighting factor vectors follows:

$$\bar{\alpha}^{t+1} = \max\{\bar{\alpha}^t - (\bar{r}^t - \bar{c})/t, 0\}$$

where α^t and r^t are the weighting factor vector and the average transmission rate vector up to time t , respectively.

III. PERFORMANCE ANALYSIS

We construct an analytical model to study the optimal scheduling policy/algorithm designed in the last section. Assume there are m RG and n BE users in the system. We assume that all users experience the i.i.d Rayleigh fading channel with average channel rate of $1/\lambda$, and all the RG users have the same guaranteed transmission rate requirement. Accordingly, due to symmetry all RG users will have the same weighting factor. Thus we drop the subscripts in the following analysis and denote the theoretical optimal weighting factor and transmission rate requirement of each RG user by α and c , respectively. Note that although we only consider Rayleigh fading in this paper, other fading models can be readily accommodated by both the scheduling algorithm in Section II and the analytical framework below.

A. The relationship between α and c

Under Rayleigh fading, a user's channel rate can be modeled by an exponential random variable with pdf $f(r) = \lambda e^{-\lambda r}$ and cdf $F(r) = 1 - e^{-\lambda r}$, respectively. Consider a tagged RG user U with channel rate r , given a weighting factor value of α , the channel access probability p_{RG} of U , i.e., the probability that U is chosen for transmission, can be obtained:

$$\begin{aligned} p_{RG} &= \int_0^\infty f(r) \prod_{j \in \text{all other users}} P(\alpha_j r_j < \alpha r) dr \\ &= \int_0^\infty f(r) F^{m-1}(r) F^n(\alpha r) dr \end{aligned} \quad (4)$$

where α_j and r_j are the weighting factor and channel rate of other users. Eqn. (5) follows the facts that RG users share the same weighting factor α and all the users experience i.i.d. channel conditions. Thus the achievable average transmission rate of a RG user can be obtained as follows:

$$c = \int_0^\infty \lambda r e^{-\lambda r} (1 - e^{-\lambda r})^{m-1} (1 - e^{-\lambda \alpha r})^n dr \quad (5)$$

By expanding the above formula using the binomial expansion and integration, we get

$$c = \frac{1}{\lambda} \sum_{i=0}^{m-1} \sum_{j=0}^n \binom{m-1}{i} \binom{n}{j} \frac{(-1)^{i+j}}{(1+i+\alpha j)^2} \quad (6)$$

From (5) we observe that c is monotonically increasing with α , therefore there is a one-to-one relationship between α and

c . Once c is given, the value of α can be uniquely determined from (6). Since $1/\lambda$ is the average channel rate of a RG user, the summation part on the r.h.s of (6) represents the gain in transmission rate brought by the weighting factor α .

B. Maximal rate can be supported for the RG users

We next derive the maximal guaranteed rate that can be supported given that the number of RG users is m .

Proposition 2 (Maximal throughput). *The maximal average transmission rate that can be guaranteed for the RG users can be represented as*

$$c_{\max} = \frac{1}{\lambda} \mathbf{O} \left(\frac{\ln(m)}{m} \right)$$

Proof: The maximal guaranteed rate is achieved if and only if all the time slots are devoted to the RG users. That means the weighing factors of RG users α should be set to infinity. Therefore we have from (5):

$$\begin{aligned} c_{\max} &= \lim_{\alpha \rightarrow \infty} \int_0^\infty \lambda r e^{-\lambda r} (1 - e^{-\lambda r})^{m-1} (1 - e^{-\lambda \alpha r})^n dr \\ &= \int_0^\infty \lambda r e^{-\lambda r} (1 - e^{-\lambda r})^{m-1} dr \end{aligned} \quad (7)$$

Again, using binomial expansion and integration again on (8) and after some mathematical manipulations, we can get

$$c_{\max} = (1/\lambda) \cdot (H_m/m) \quad (8)$$

where H_m is the m th harmonic number and can be approximated as

$$H_m = \sum_{i=1}^m i^{-1} \approx \ln(m) + \gamma \quad (9)$$

where $\gamma \approx 0.5882$ is known as Euler's constant. Combine the result obtained from (7) to (9), we reach the conclusion made in Proposition 2. ■

Proposition 2 tells us that c_{\max} is monotonically decreasing with m and is not affected by n . From the proposition, it is also easy to find the maximal number of RG users that can be supported in the system once the channel condition/rate distribution is given.

C. Channel access delay

We define the time gap between two consecutive transmissions by the same user as its *channel access delay*. Let d_{RG} denote the channel access delay for RG user T . To find d_{RG} , we first obtain channel access probability p_{RG} from (4):

$$p_{RG} = \sum_{i=0}^{m-1} \sum_{j=0}^n \binom{m-1}{i} \binom{n}{j} \frac{(-1)^{i+j}}{1+i+\alpha j}$$

Next, the probability that there are k time slots between two consecutive channel accesses by a RG user is:

$$P(d_{RG} = i) = (1 - p_{RG})^i \cdot p_{RG}, \quad i = 0, 1, 2, \dots$$

We can see that the channel access delay of a RG user follows a geometric distribution with a success probability of p_{RG} . Then the average delay can be easily found as follows:

$$E[d_{RG}] = (1 - p_{RG})/p_{RG}$$

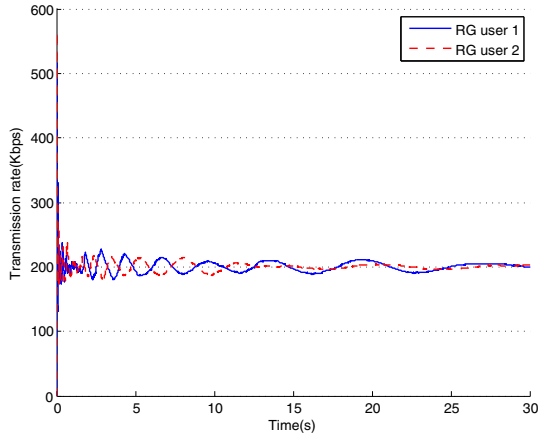


Fig. 1. Transmission rate of two randomly chosen RG users versus time.

D. Worst case delay

The worst case delay of the system is defined as the minimum number of channel uses that guarantees all N users successfully access the channel M times. Such definition is a more stringent notion of delay than average delay per user and is the worst case delay among all the users. Once the channel access probability is obtained, the worst case delay can directly calculated as follows [8]:

$$E(D_{M,N}) = N \int_0^\infty \left[1 - \prod_{i=1}^m \left(1 - S_M(p_{RG} N t) e^{-p_{RG} N t} \right) \times \prod_{i=1}^n \left(1 - S_M(p_{BE} N t) e^{-p_{BE} N t} \right) \right] dt$$

where $S_M(t) = \sum_{k=0}^{M-1} \frac{t^k}{k!}$.

E. System throughput

Since all RG users get exactly what they ask for, we only need to derive the throughput achieved by the BE users. Let Y denote the average transmission rate achieved by a BE user, following the derivation of c in (6), Y can be obtained as:

$$Y = \int_0^\infty \lambda r e^{-\lambda r} (1 - e^{-\lambda r/\alpha})^m (1 - e^{-\lambda r})^{n-1} dr$$

$$= \frac{1}{\lambda} \sum_{i=0}^m \sum_{j=0}^{n-1} \binom{m}{i} \binom{n-1}{j} \frac{(-1)^{i+j}}{(1 + i/\alpha + j)^2}$$

Again, here we can see that the summation part on the r.h.s of above equation represents the gain in transmission rate of a BE user brought by the weighting factor α . Then, the system throughput can be calculated by taking all users into account:

$$\text{Throughput} = mc + nY$$

IV. PERFORMANCE EVALUATION

Four opportunistic scheduling algorithms are studied and compared in this section. They are:

- OS-T: our proposed optimal scheduling algorithm with pre-calculated theoretical parameters (derived by assuming that the channel rate distribution is known a priori);

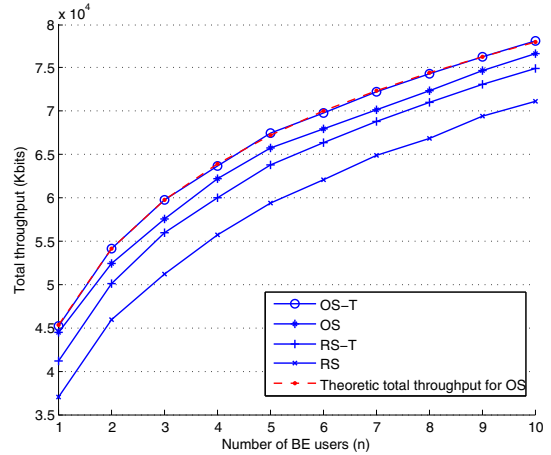


Fig. 2. System throughput under a homogeneous system.

- OS: our proposed optimal scheduling algorithm with stochastic approximation method (given in Section II.D);
- RS-T: the (reference) scheduling algorithm in [7] with pre-calculated theoretical parameters, and
- RS: the (reference) scheduling algorithm in [7] with stochastic approximation method.

Note that the scheduler in [7] maximizes the total fraction of time slots for BE users, but the criteria for choosing among the BE users is not specified. To maximize the system throughput, our implementation of RS-T and RS always selects the user with the best channel quality among all BE users.

In our simulations, we assume all users experience the same i.i.d Rayleigh fading. Each simulation run consists of 60000 slots (i.e. 60 seconds), where the length of each slot is 1ms. All four scheduling algorithms are implemented and compared.

Figs.1 to 3 show the simulation results in a homogeneous system with the following parameters: mean channel rate $1/\lambda = 800Kbps$, guaranteed rate for each RG user $c = 200Kbps$, number of RG users $m = 4$, and number of BE users n varying from 1 to 10. For the OS scheduler, we randomly choose 2 RG users from the simulation on $n = 8$ and plot their received average transmission rates against time in Fig.1. We can clearly see that the guaranteed rate of $c = 200Kbps$ is achieved very quickly.

We then compare the system throughput performance between our optimal schedulers (OS-T & OS) and the reference schedulers (RS-T & RS) in Fig.2. We can see that due to the imperfect stochastic approximation on channel rate distribution, the system throughputs obtained by OS and RS are consistently lower than their respective counterparts with perfect channel knowledge (i.e. using pre-calculated theoretical parameters). Nevertheless, our OS still outperforms RS-T, the reference scheduler with perfect parameters. On average our OS algorithm achieves more than 10% gain in throughput than the RS scheduler.

Next we plot the channel access delay of a random chosen RG user versus the number of BE users in the system in Fig. 3. In general, we can see that channel access delays (obtained using different schedulers) remain quite stable as the number of BE users increases. This is because all four opportunistic

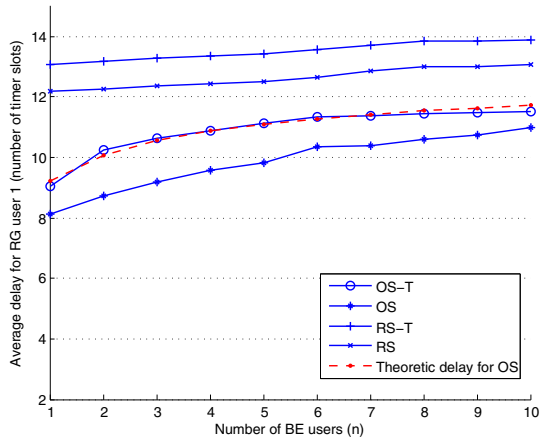


Fig. 3. Average delay of a randomly chosen RG user under a homogeneous system.

schedulers are designed to ensure that RG users get their required bandwidth and thus they have priority over BE users.

From Fig. 3, we also notice that with stochastic approximation, OS and RS algorithms yield shorter access delay than their ideal counterparts, OS-T and RS-T. This is because in the implementation of OS and RS (as described in Section II. D), the vector of weighting factors $\bar{\alpha}^t$ changes with time and is sensitive to the already-achieved transmission rate at a particular instant. According to the updating rule in (4), when a user is below its guaranteed rate, it has a larger chance to be served, and thus its access delay is reduced. While for schedulers with perfect parameters (OS-T and RS-T), this is not the case since the values of weighting factor are pre-determined (based on known channel rate distribution) and do not change with time.

Fig. 4 compares the system throughput and channel access delay performance of the four schedulers under a heterogeneous setting where mean channel rate $1/\lambda = 1000Kbps$, the number of RG users $m = 5$, and 2 RG users requesting for guaranteed rate $c_1 = 300Kbps$ and 3 RG users requesting for $c_2 = 200Kbps$, and number of BE users n varying from 1 to 10. (Note that the RG user rate convergence trace is similar to that in Fig. 1, and thus skipped.) Here we only simulate the schedulers with stochastic approximation implementation (i.e. OS and RS) under such case. The reason is that the heterogeneous case is more likely the case in a practical network and the BS may not have full knowledge of the channel condition of the users. Therefore, it is more appropriate for the BS to use aforementioned adaptive stochastic approximation methods. From the figure, again we see that our OS consistently outperforms RS in [7].

V. CONCLUSION

We studied opportunistic scheduling algorithms in a wireless network with both rate guaranteed (RG) users and best effort (BE) users. We proposed an optimal opportunistic scheduler that maximizes the throughput for all BE users. An analytical model was also constructed to evaluate its performance in a homogenous system. Closed form solutions for channel

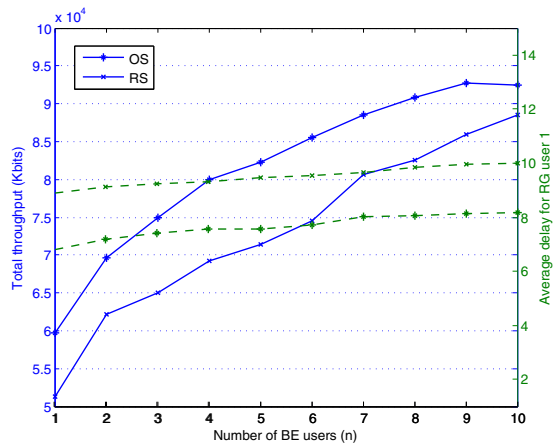


Fig. 4. System throughput and average channel access delay of a randomly chosen RG user under a heterogeneous system.

access delay and system throughput were derived. Extensive simulations in both homogeneous and heterogeneous systems were conducted to verify the effectiveness of our scheduler. A possible future work is to extend the opportunistic scheduling algorithms to support multihop communications [9].

REFERENCES

- [1] R. Knopp and P. A. Humblet, "Information Capacity and Power Control in Single Cell Multiuser Communications," *Proc. of the IEEE Int. Conf. on Communications*, pp. 331-335, June 1995.
- [2] X. Liu, E. K. P. Chong, and N. B. Shroff, "Opportunistic transmission scheduling with resource-sharing constraints in wireless networks," *IEEE J. Select. Areas Commun.*, vol. 19(10), pp. 2053-2064, Oct. 2001.
- [3] X. Liu, E. K. P. Chong, and N. B. Shroff, "A framework for opportunistic scheduling in wireless networks," *Computer Networks*, vol. 41, no. 4, pp. 451-474, Mar. 2003.
- [4] Daeyoung Park, H. Seo, Hojoong Kwon and Byeong Gi Lee, "Wireless packet scheduling based on the cumulative distribution function of user transmission rates," *IEEE Transactions on Communications*, vol.53, no.11, pp. 1919-1929, Nov. 2005.
- [5] Syed Hussain Ali, Ki-Dong Lee, and Victor C.M. Leung, "Optimum Throughput Constrained Opportunistic Scheduling in Cellular Data Networks," *IEEE International Symposium on Signal Processing and Information Technology*, pp.708-713, Aug. 2006.
- [6] Fei Long and Gang Feng, "Efficient Rate-Guaranteed Opportunistic Scheduling for Wireless Networks," *IEEE Transactions on Vehicular Technology*, vol.57, no.2, pp.1199-1208, March 2008.
- [7] M. Zafer and E. Modiano, "Joint Scheduling of Rate-Guaranteed and Best-Effort Users over a Wireless Fading Channel," *IEEE Transactions on Wireless Communications*, vol.6, no.10, pp.3825-3835, October 2007.
- [8] M. Sharif and B. Hassibi, "A delay analysis for opportunistic transmission in fading broadcast channels," *Proc. IEEE INFOCOM*, 2005.
- [9] H. S. Chiu, Kwan L. Yeung and K. S. Lui, "J-CAR: An Efficient Joint Channel Assignment and Routing Protocol for IEEE 802.11-based Multi-Channel Multi-Interface Mobile Ad Hoc Networks," *IEEE Transactions on Wireless Communications*, Vo. 8, No. 4, pp.1706-1715, April 2009.