

Persistent Resource Allocation in OFDMA Networks for Real-Time and Non-Real-Time Traffic

Vinay Majjigi, Rajiv Agarwal, Rath Vannithamby and John Cioffi

Abstract—We consider optimal allocation of resources to users in a downlink OFDMA system to support heterogeneous applications consisting of both deadline-sensitive (DS) and best-effort (BE) users. Given the user rate-requirements and long-term statistics, the proposed persistent scheduling algorithm allocates the minimum resources to ‘just’ meet the deadlines for DS packets. The remaining resources are optimally split to maximize a rate objective for BE users. We resort to a greedy-tone allocation algorithm to solve the optimization problem. By considering persistent scheduling instead of instantaneous scheduling, the proposed scheme’s complexity and required feedback overhead is significantly reduced, hence a practical algorithm for a real-system.

I. INTRODUCTION

The rapid growth of multimedia traffic in the wireless downlink necessitates efficient resource allocation to cope with system constraints. OFDMA is an attractive transmission strategy due to its spectrum efficiency and robustness to inter-symbol interference. In OFDMA networks, the base-station (BS) allocates tones to users in a time- and/or frequency-division manner. In static allocation schemes, the transmitter predetermines and fixes the tone allocation for all time. Such schemes require negligible complexity but are not optimized for varying channel conditions and Quality of Service (QoS) criteria. Hence, a large increase in performance is possible by dynamically allocating resources: tone assignment, rate, and power based on Channel State Information (CSI) and QoS requirements.

The degrees of freedom available to the BS to allocate tones, rate, and power allow for highly optimized allocations and improved system performance, albeit at the expense of high computational complexity. For a globally optimal solution, the discrete nature of tone allocation results in a NP-Hard 0-1 Integer Program [1]. In published literature, sub-optimal algorithms have been proposed with varying complexity, performance, and assumptions to allocate resources, [1], [2],[3],[4],[5],[6],[7],[8]. The benefit of these sub-optimal algorithms is their performance increase over static-allocation schemes, and their low complexity compared to the optimal 0-1 Integer Program.

Nonetheless, most current research has focused on allocating resources based on the instantaneous CSI, which changes within the coherence time of the channel. Hence, there is an implicit need to solve the optimization problem on a similar time scale. This limits their practicality when designing a real-system where the complexity in solving the allocation problem and overhead of informing users of a new allocation strategy

occurs on a time scale of 5 ms¹.

Resource allocation objectives typically fall into two broad categories: (i) minimize total power while satisfying rate requirements; or (ii) maximize rate subject to power constraints. These two objectives generally align with serving deadline-sensitive (DS) and non-deadline-sensitive traffic, respectively. That is, DS users are thought to have a hard rate-requirement to ensure packets meet their expiry deadline. While users with non-deadline-sensitive traffic are treated as best-effort (BE) users, where their rate is maximized subject to resource constraints. In the published literature, most schemes consider only one of the two objectives, whereas practical networks have both DS and BE users, and a single objective algorithm will perform poorly in satisfying all users.

In this paper, three ideas for OFDMA resource allocation are considered: (i) persistent tone-allocation; (ii) heterogeneous traffic consisting of DS and BE users; and (iii) a low complexity algorithm. Persistent tone allocation refers to fixing the tone allocation for users over many channel coherence times. The long-term statistics of the channel are employed to allocate tones to specific users. By reducing the frequency of solving the tone-assignment problem, the computational complexity required at the transmitter and the overhead between users and BS are reduced, as compared with instantaneous scheduling, where the allocation problem is solved on the same order of the channel coherence time. A low complexity greedy-allocation scheme is introduced that jointly solves the optimization problem for both DS and BE users.

Notation: $|\mathcal{S}|$ denotes the cardinality (the number of elements) of a set \mathcal{S} . $\mathcal{S} = \mathcal{S}_1 \cup \mathcal{S}_2$ denotes set-union i.e. \mathcal{S} contains all elements of \mathcal{S}_1 and \mathcal{S}_2 . $\mathcal{S} = \mathcal{S}_1 \setminus \mathcal{S}_2$ denotes set-difference i.e. \mathcal{S} contains all elements of \mathcal{S}_1 which are not present in \mathcal{S}_2 . $x := y$ denotes the value of y is assigned to x , to distinguish from $x = y$, which states a mathematical fact.

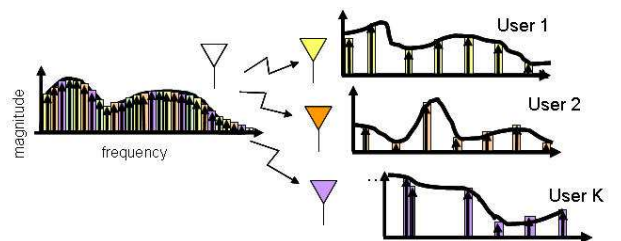


Fig. 1. OFDMA Broadcast Channel [9]

¹A typical value for channel coherence time in the WiMAX standard.

II. SYSTEM MODEL

Consider an OFDMA Broadcast Channel with K users and M tones. Each tone undergoes independent block fading i.e. the channel state is unchanged during a coherence time and then changes in an independent manner to a new realization. If tone m is assigned to user k , the received signal y_{km} is expressed as

$$y_{km} = h_{km}x_{km} + z_{km},$$

$$k = 1 \text{ to } K, m = 1 \text{ to } M \quad (1)$$

where x_{km} is the transmitted symbol, h_{km} is the random complex channel-gain, and z_{km} is the i.i.d. zero-mean complex Gaussian noise with unitary variance. There is a total power constraint on each OFDM symbol P^{total} .

The fading distribution of h_{km} is discrete. Let the channel power-gain γ_{km} be defined as $\gamma_{km} \triangleq |h_{km}|^2$. Γ_{km} denotes the set consisting of the possible values of γ_{km} . γ_{km} takes values γ_{km}^i with probability p_{km}^i , $\sum_{i=1}^{|\Gamma_{km}|} p_{km}^i = 1$.

The BS is assumed to have perfect knowledge of the long-term channel statistics. This assumption is required to allocate tones in a persistent manner. However, unlike many previous works, instantaneous CSI at the BS is not required for *all* users on *all* tones. Instead, for any tone m , instantaneous CSI is only required by the user allocated tone m . Note that the amount of CSI at the BS is significantly less than is required by instantaneous scheduling as discussed in Section III.

A scheduling-period (SP) is the time between the BS re-allocating tones to the users based on their long-term statistics and is of duration T_{SP} seconds. The duration of an SP is many times larger than the coherence time of the channel t_{coh} to ensure that the ergodicity of the fading channel distribution holds. The frame-period (FP) is the time between link-adaptation (re-allocation of power and rate to the users) and is of duration T_{FP} seconds. The frame-period is set equal to the coherence time $T_{\text{FP}} = t_{\text{coh}}$. In persistent scheduling, link-adaptation is done every T_{FP} , but tone allocation is performed every T_{SP} . The reason for persistent tone allocation but instantaneous link-adaptation, is that tone-allocation is optimization over a discrete-set which is fairly complicated, whereas optimal link-adaptation has a closed-form expression for continuous power and rate allocations. A comparison of the different scheduling methods is shown in Table I.

TABLE I
RELATIVE TONE ALLOCATION AND LINK ADAPTATION PERIODS FOR DIFFERENT METHODOLOGIES

Scheduling Method	Tone Assignment Period (T_{SP})	Link Adaptation Period (T_{FP})
Static	∞	T_{FP}
Persistent	$T_{\text{FP}} \ll T_{\text{SP}} < \infty$	T_{FP}
Instantaneous	T_{FP}	T_{FP}

Let \mathcal{M} be the set of all tones. Then, let K^{DS} denote the number of deadline-sensitive users, \mathcal{K}^{DS} denote the subset of DS users, and \mathcal{M}^{DS} denote the subset of tones allocated to

DS users, likewise for K^{BE} , \mathcal{K}^{BE} , and \mathcal{M}^{BE} for BE users, such that $K = K^{\text{DS}} + K^{\text{BE}}$, $\mathcal{K} = \mathcal{K}^{\text{DS}} \cup \mathcal{K}^{\text{BE}}$, and $\mathcal{M} = \mathcal{M}^{\text{DS}} \cup \mathcal{M}^{\text{BE}}$. For a DS user, the associated packet deadline D_k , the packet-arrival process, and the QoS guarantee q_k imply a minimum average required-rate R_k^{req} . 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 transmission rate R_k , mathematically, the following relationship must hold for guaranteeing QoS to user k :

$$\text{Prob}(\text{Packet-delay}(\text{Arrival-Process}_k, R_k) > D_k) < 1 - q_k \quad (2)$$

Each DS user, given the distribution of its packet-arrival process and D_k , calculates R_k for a given q_k using Eqn (2). BE users do not have any deadline for their packets. Hence, their rate allocation is flexible and any general BE objective may be used to assign rates to the BE users [10].

III. PROPOSED SCHEME

This section gives an explicit formulation of the problem, describes the proposed greedy-tone allocation algorithm, and studies two benchmark schemes.

A. Problem Formulation

The objective is to maximize the ergodic weighted sum-rate for BE users while satisfying rate requirements of DS users and total power constraints as seen in Eqn (3)-Eqn (7).

maximize:

$$\sum_{k \in \{K^{\text{BE}}\}} w_k \sum_m b_{km} \sum_i P_{km}^i r(P_{km}^i) \quad (3)$$

subject to:

$$\sum_m b_{km} r_{km}^{\text{DS}} \geq R_k^{\text{req}}, \forall k \in \{K^{\text{DS}}\} \quad (4)$$

$$\sum_m \sum_{k \in \{K\}} b_{km} P_{km}^i \leq P^{\text{total}}, \forall i \quad (5)$$

$$b_{km} \in \{0, 1\} \quad (6)$$

$$\sum_{k \in \{K\}} b_{km} \leq 1, \forall m \quad (7)$$

The discrete optimization variable b_{km} denotes tone allocation, i.e. $b_{km} = 1$ if tone m is allocated to user k , and $b_{km} = 0$ otherwise. The continuous optimization variables are P_{km}^i , the power allocated in noise state i , and r_{km}^i , the bits per tone in noise state i . In this formulation, a continuous r_{km}^i is employed rather than a practical discrete constellation to simplify analysis. The parameters w_k are the user weights for weighted sum-rate maximization, and, in practice, can be altered to consider fairness, QoS requirements, etc.

For BE users, the power allocated P_{km}^i follows a water-filling strategy where more power is allocated in good channel states, and less power in poor channel states. For DS users, a constant rate must be satisfied in all fading states when not in outage, as captured in Eqn (4). In this case, the BS allocates power based on truncated channel inversion, that is

more power is allocated in poor channel states, and less power in good channel states. If the channel state drops below a cutoff value γ_o , outage results [11]. The constant rate maintained for DS users is,

$$r_{km}^{\text{DS}} = \log(1 + \sigma_{km}) \text{Prob}(\gamma_{km} > \gamma_o), \quad (8)$$

where σ_{km} is the constant channel state for DS users that the BS maintains by dynamically adjusting P_{km}^i , given as,

$$\sigma_{km} = P_{km}^i \gamma_{km}^i, \forall i \quad (9)$$

The constraints given by Eqn (5)-Eqn (7) specify that the power allocated in a noise state cannot exceed the total power, tones cannot be fractionally allocated, and tones can be allocated to at most one user, respectively.

B. Greedy-Tone Scheme

As the allocation of tones remains a hard optimization problem over a discrete space, we propose a heuristic scheme to perform tone allocation based on the long-term channel-statistics. The proposed scheme is a greedy-tone allocation scheme that tries to ‘just’ meet the DS users’ rate requirements with minimal resources. The intuition of the scheme arises because of the interdependency between allocating tones and allocating power to the DS users. For example, allocating few tones to the DS users will require high power to meet their rate requirements, leaving little power for the BE users. Likewise, allocating too many tones to the DS users will reduce the required power but will leave only a few tones to the BE users. Therefore, the proposed scheme decides both the number of tones and power to the DS users in a greedy-manner to maximize the BE objective.

The algorithm is split into an initialization stage and the greedy-tone iterative stage. In the initialization, the BS ranks tones based on their potential contribution to the BE objective. For each tone and BE user, the possible contribution is approximated by the weighted geometric mean². Therefore the value a tone contributes to the BE objective is given by μ_m ,

$$\mu_m = \max_{k \in \mathcal{K}^{\text{BE}}} \prod_i \gamma_i^{w_k p(\gamma_i)}, m = 1, 2, \dots, M \quad (10)$$

Then the tones are ranked among other tones as,

$$\begin{aligned} m_{\pi(1)} &= \arg \max \mu_m, m \in \mathcal{M} \\ m_{\pi(2)} &= \arg \max \mu_m, m \in \mathcal{M} \setminus m_{\pi(1)} \\ &\vdots \\ m_{\pi(M)} &= \arg \max \mu_m, m \in \mathcal{M} \setminus \{m_{\pi(1)}, \dots, m_{\pi(M-1)}\} \end{aligned} \quad (11)$$

Where $m_{\pi(1)}$ and $m_{\pi(M)}$ correspond to the best and worst tones for BE user objective, respectively. Then, each DS user

²The geometric mean is the natural metric because for water-filling rate r is given by $\sum_i \log(\gamma^i) p^i = \log(\prod_i \gamma_i^{p^i})$ [11].

is allocated one tone from among the worst tones for the BE objective. Given their rate requirements, the BS computes the initial power required for meeting DS users’ rate requirements. The remaining tones and power are then used to calculate the BE user objective function.

After the initialization, tones are incrementally taken from the BE users and given to the DS users. The BS recomputes the power required by DS users, and with the remaining power, maximizes the BE objective. The scheme is explicitly described below:

Initialization:

Rank BE tones using Eqn (10)-Eqn (11);

Let K^{DS} worst BE tones be assigned to DS users

so, $|\mathcal{M}_{(0)}^{\text{DS}}| = K^{\text{DS}}$ and $|\mathcal{M}_{(0)}^{\text{BE}}| = M - K^{\text{DS}}$;

From $\mathcal{M}_{(0)}^{\text{DS}}$, allocate one tone to each DS user such that the user with the highest rate requirement is given its best tone and so on;

Determine power required for DS users $P_{(0)}^{\text{DS}}$;

Set iteration number $j := 1$;

Greedy-Tone Iteration:

While ($\mathcal{M}^{\text{BE}} \neq \{\phi\}$);

Remove worst tone $m_{(j)}^* \in \mathcal{M}_{(j-1)}^{\text{BE}}$

so, $|\mathcal{M}_{(j)}^{\text{BE}}| = |\mathcal{M}_{(j-1)}^{\text{BE}}| - 1$;

Allocate $m_{(j)}^*$ to the DS user³ that minimizes $P_{(j)}^{\text{DS}}$

so $|\mathcal{M}_{(j)}^{\text{DS}}| = |\mathcal{M}_{(j-1)}^{\text{DS}}| + 1$;

Compute BE Objective $Obj_{(j)}$,

with $\mathcal{M}_{(j)}^{\text{BE}}$ and $P_{(j)}^{\text{BE}} = P^{\text{total}} - P_{(j)}^{\text{DS}}$;

$j := j + 1$;

End;

Finish:

Greedy-Tone Iteration j corresponding to maximum $Obj_{(j)}$

is the final allocation strategy;

In any iteration j , $P_{(j)}^{\text{BE}}$ increases but $|\mathcal{M}_{(j)}^{\text{BE}}|$ decreases, hence $Obj_{(j)}$ may either increase or decrease. Notice that $P_{(j)}^{\text{BE}}$ increases since $P_{(j)}^{\text{DS}}$ always decreases as DS users are allocated more tones. This reduction in required power is clear because the rate requirements are fixed, and additional bandwidth and degrees of freedom are available to the BS to allocate rate and power, hence $P_{(j)}^{\text{DS}} \geq P_{(k)}^{\text{DS}}, j < k$.

C. Benchmark Schemes

The proposed persistent scheme is compared with an instantaneous scheduling and a static allocation scheme to highlight the performance advantages and trade-offs of each scheme. The instantaneous scheme utilizes the same greedy-tone algorithm described in Section III-B with a modification to Eqn (10) given as,

$$\tilde{\mu}_m^i = \max_{k \in \mathcal{K}^{\text{BE}}} \gamma_i^{w_k}, m = 1, 2, \dots, M \quad (12)$$

The static scheme allocates an equal amount of tones to each user and only performs link-adaptation, i.e. $\lfloor \frac{M}{K} \rfloor$ tones are allocated to each user. If $M - \lfloor \frac{M}{K} \rfloor K \neq 0$, DS users

³If the k^{th} DS user has L tones, the rate on each tone is $\frac{R_k^{\text{req}}}{L}$

are randomly selected and given one tone until all tones are allocated.

The feedback requirement for the persistent scheme is considerably less than the instantaneous schemes, and is close to the required feedback of the static tone allocation scheme as numerically shown in Figure 3. In the static scheme, users must feedback their channel values for link-adaptation resulting in an overhead of MB_{real} bits, where quantization level of a real number is B_{real} bits. In the persistent scheduling scenario, in addition to the MB_{real} bits overhead for link-adaptation, every T_{SP} seconds ($T_{\text{SP}} \gg T_{\text{FP}}$) all users feedback the long-term statistics, e.g. the geometric mean of each tone. Due to the infrequency of this additional term, the feedback requirement does not grow as quickly as the instantaneous scheduling, where MKB_{real} bits must be fed back every T_{FP} seconds. The inherent assumption of the persistent scheme is the validity of the long-term statistics over T_{SP} seconds.

The complexity of each scheme is compared in Table II⁴. The static scheme has constant complexity $O(1)$ due to the closed form expressions for link-adaptation. The instantaneous scheduling scheme solves a greedy-allocation algorithm every T_{FP} seconds resulting in a complexity $O(MK)$. The proposed scheduling scheme solves the greedy-allocation algorithm only every T_{SP} seconds, so the complexity is reduced to $O(\frac{T_{\text{FP}}}{T_{\text{SP}}}MK)$.

TABLE II
COMPLEXITY FOR THE DIFFERENT SCHEMES

Scheduling Method	Complexity
Static	$O(1)$
Persistent	$O(\frac{T_{\text{FP}}}{T_{\text{SP}}}MK)$
Instantaneous	$O(MK)$

IV. NUMERICAL RESULTS

In this section, the proposed persistent scheduling scheme is compared with instantaneous scheduling and a static tone allocation scheme for heterogeneous traffic. The numerical results suggest a large decrease in feedback overhead and complexity are possible with some performance degradation when allocating tones in a persistent versus instantaneous manner.

The results are simulated considering a Rayleigh fading distribution discretized to 4 levels where quantization is done by dividing the channel gain pdf into 4 regions of equal probability. Quantized channel state values are considered because of the finite number of bits that users feedback, in this case $B_{\text{real}} = 2$ bits. The number of DS users is varied to evaluate the performance as the system is stressed with additional rate requirements. The BS maximizes the the ergodic weighted sum-rate for BE users, where the weights are equal. We study performance in terms of two quantities

⁴ $f = O(\phi)$ means that $|f| \leq A\phi$ for some constant A and for all n with a sequence f_n .

of interest: (i) throughput of the BE users, and (ii) required feedback in bits for a particular scheme.

For the plots, $R_k^{\text{req}} = 5$ for all DS users, $P^{\text{total}} = 300$, the distribution of $\tilde{\gamma}_{km} \sim U[20, 27]$. Figure 2 studies the effect of increasing the number of DS users from 1 to 15 on the BE users' throughput. The number of tones in the OFDM symbol is $M = 64$, the number of BE users is $K^{\text{BE}} = 2$, and the BE users' weights are equal $w_1 = w_2 = 0.5$. As can be seen in Figure 2, the instantaneous scheme performs better than the persistent scheme but requires a significantly larger amount of feedback as seen in Figure 3. Also, the persistent scheme achieves better performance than the static scheme, but only requires an incremental increase in feedback. As is expected, all three schemes suffer a decrease in BE throughput as the number of DS users increases.

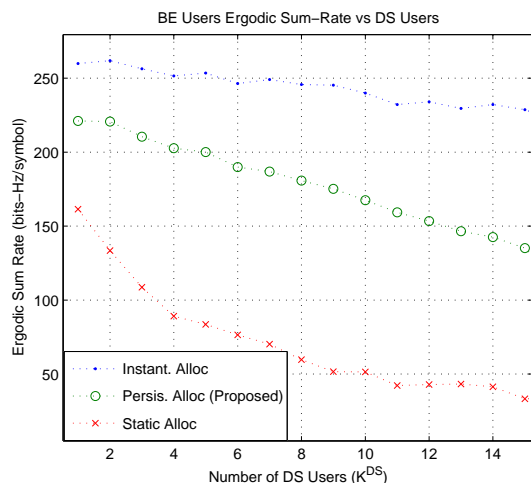


Fig. 2. Various schemes performance considering BE throughput versus the number of DS users

Figure 3 highlights the necessary feedback between static, persistent, and instantaneous scheduling as the system size increases in terms of tones. The long-term statistics are taken to be valid over 100 coherence times, i.e. $T_{\text{SP}} = 100T_{\text{FP}}$. The number of users is taken as $K = 20$ where the specific number of DS and BE users is irrelevant for this feedback comparison. The total number of tones M varies from 16 to 128. From the discussion in Section III-B, it is clear that the necessary feedback for instantaneous eclipses the other two schemes especially when K and T_{SP} are large.

V. CONCLUSION

This paper proposed a persistent scheduling algorithm for the OFDMA broadcast channel that provides QoS for deadline-sensitive service and allocates the remaining resources for the best effort service. The scheme utilizes a greedy-tone allocation scheme that allocates tones to users in a persistent manner, but still performs link-adaptation every channel coherence time. Through simulations, it is shown that the persistent allocation scheme greatly reduces the necessary feedback compared to the instantaneous scheduling

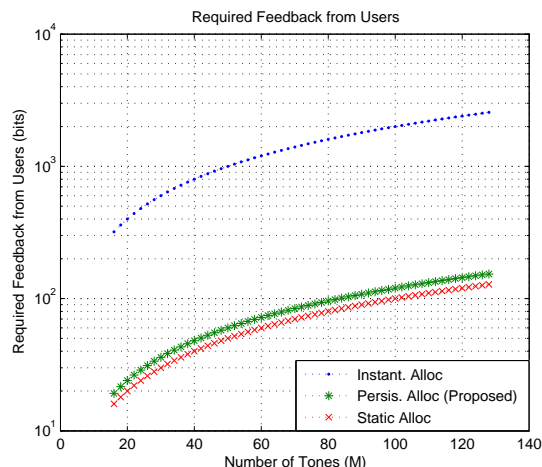


Fig. 3. Required feedback for the various schemes

schemes, but suffers only a slight throughput degradation. For practical implementations where limited feedback is a necessity, and thus instantaneous scheduling is not feasible, persistent scheduling offers a low-complexity flexible solution to accommodate heterogeneous traffic in an optimized manner.

REFERENCES

- [1] C. Wong, R. Cheng, K. Letaief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE Journal on Selected Areas in Communications*, vol. 17, pp. 1747–1758, Oct. 1999.
- [2] J. Chen, R. A. Berry, and M. L. Honig, "Performance of limited feedback schemes for downlink ofdma with finite coherence time," in *IEEE ISIT*, Jun 2007.
- [3] A. Dua and N. Bambos, "Downlink scheduling of heterogeneous traffic," in *IEEE ICC*, Jun 2006.
- [4] J. Gross, H. Karl, F. Fitzek, A. Wolisz, "Comparison of heuristic and optimal subcarrier assignment algorithms," *Proc. of Intl. Conf. on Wireless Networks*, June 2003.
- [5] D-H. Kim, B-H. Ryu, and C-G. Kang, "Comparison of heuristic and optimal subcarrier assignment algorithms," *Proc. of Intl. Conf. on Wireless Networks*, June 2003.
- [6] K. Seong, M. Mohseni, and J. M. Cioffi, "Optimal resource allocation for ofdma downlink systems," in *IEEE ISIT*, July 2006.
- [7] Zhu Han, Zhu Ji, K.J. Ray Liu, "Fair multiuser channel allocation for OFDMA networks using nash bargaining solutions and coalitions," *IEEE Trans. on Communications*, vol. 53, Aug. 2005.
- [8] Ahmed K. F. Khattab and Khaled M. F. Elsayed, "Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks," *IEEE Computer Society*, 2006.
- [9] I. C. Wong and B. L. Evans, "Optimal OFDMA resource allocation with linear complexity to maximize ergodic rates," *IEEE Trans. on Wireless Communications*, Feb 2007.
- [10] A. Ibing and H. Boche, "Fairness vs. efficiency: Comparison of game theoretic criteria for ofdma scheduling," in *41st Asilomar Conf. on Sig, Sys, and Comp*, November 2007.
- [11] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.