An Efficient Channel and Queue Aware Resource Allocation Strategy in Wireless Access Networks

Vaibhav Singh[†], Swades De[†], Hari M. Gupta[†], Navrati Saxena[‡], and Abhishek Roy[*]
[†] Electrical Eng. Department, Indian Institute of Technology Delhi, New Delhi, India
[‡] School of Info. and Commun. Eng., Sungkyunkwan Univ., Suwon, South Korea
[*] WiBro System Lab, Samsung Electronics, Suwon, South Korea

***Abstract*—We propose a new channel rate and queue state adaptive time resource allocation heuristic for the base stations in a multiuser wireless access scenario. Motivation of the scheme is derived through the effective capacity concept. Our proposed strategy accounts for fairness and quality-of-service (QoS) of the system on per frame basis. Our results show that the proposed scheme benefits the users with low service rate substantially without disturbing QoS of the users with high service rate.**

I. Introduction

Resource allocation algorithm in a multiuser access scenario plays a vital role in efficient utilization of limited channel resource and supporting quality-of-service (QoS) of users. Dynamic resource allocation at the physical layer has been extensively studied in recent years. Orthogonal frequency division multiple access (OFDMA), which has been adopted by the WiMAX forum as an air interface standard in mobile broadband access, guarantees robust subchannels to achieve high system QoS. A wide range of work have addressed dynamic subchannel allocation to the users to meet their QoS and fairness criteria. However, dynamic allocation of subcarriers to a large number of users for every OFDM symbol requires a lot of channel estimation and signal processing overhead [1]. If the environment has high fading, this task becomes even more challenging and resource-intensive at the base station (BS). Therefore the subcarrier allocation is performed in every few OFDM symbols, and the channel estimation – which uses the block type pilot arrangement [2] – is spaced over a block of OFDM symbols. Thus, within a frequency resource allocation cycle, the subcarriers assigned to the users remain static. In a highly fading environment this arrangement implies that, some users may encounter poorer subchannels while the others may have much better quality subchannels.

In this paper, we aim at improving the network access performance by adopting a micro-level dynamic resource assignment at the link layer. We propose a heuristic that requires limited information from the physical layer about the channel condition and utilizes the short-term user demand information to optimally adjust the time slice allocated to the users, so as to provide an improved QoS and fairness among the users.

This research was partly supported by the Indo-Korean Joint Programme of Cooperation in Science and Technology under the DST (India) grant no. INT/ROK/PROJ-8-08 and KICOS (South Korea) grant no. 2009-0225-000.

A. Related Work

Wireless resource management can be incorporated into a unified optimization framework across different protocol layers. On the cross-layer resource management, different classes of link layer scheduling optimization schemes were studied in [3]. One class of algorithms aims at maximizing the access performance by exploiting multiuser diversity, where sum capacity maximization of the system and user-level fairness tradeoffs are addressed. Another class approaches to maximize the load balancing in multihop wireless networks by adopting various stabilized scheduling policies. All these optimization schemes are computationally complex, as the users are scheduled in every time slot – which is the smallest unit of time defined in 4G wireless access standards. A feedback control based resource allocation algorithm was proposed in [4], where bandwidth allocation is done as a function of difference between the expected QoS and achieved QoS. Thus, the bandwidth allocated to a user was considered variable, which depends on the total packet loss. The latency and throughput tradeoff was studied in [5] for CDMA systems, where the channel state information was used for proportional user rate allocation. This approach exploits the inherent heterogeneity of service users with varying data rate, but it does not address QoS in terms of bounded delay of the customers.

Another approach of the cross layer resource management involves dynamic subcarrier allocation at the physical layer guaranteeing QoS to different users. An optimal subcarrier allocation for OFDMA was proposed in [6] by maximizing the utility function of average waiting time. An algorithm for slot allocation was proposed in [7] according to different weights of users calculated on the basis of subchannel gains to guarantee fairness and QoS in terms of bounded delay. Both these algorithms are associated with large overheads, as channel estimation and assignment are done in every time slot at the BS for all users. The problem of excess overhead in OFDM subcarrier allocation was noted in [1]. However, as suggested in [2], it can be avoided if the assignment is spaced over every few OFDM symbols.

The concept of effective capacity was introduced in [8] on the similar line of effective bandwidth concept introduced earlier [9]. This concept can be used to study QoS with respect to channel rate and service information, hence it can be adopted for resource allocation at the data link layer.

The effective capacity was used in [10] for downlink QoS support by dynamically adjusting transmit power and time slots. This approach requires to calculate effective bandwidth and effective capacity functions, which could be complex depending on the statistical characteristics of traffic arrival and service process. Also, fairness to users with different channel quality was not addressed in the work.

In contrast to prior works, in this paper we attempt on optimizing the user QoS and fairness by assigning appropriate weights to the channel rate and queue state information. Unlike in multiuser diversity, the optimization approach can offer QoS guarantee to the users irrespective of the channel quality.

B. Our Contribution

A simple heuristic scheme is devised by assigning an appropriate weight to the channel information and link layer queues to optimally improve the QoS performance of the users having low channel rate without disturbing the QoS of other users. Application domain of our proposed approach is in resource allocation scenarios where the assigned wireless channel to the users cannot be altered.Cellular wireless system is one such example, where the assigned frequency band is fixed at the session initiation level. Another more recent application area would be OFDMA based WLAN (802.11) or WRAN (802.16) access networks, where, although subcarrier reassignment is possible, due to high overhead this operation is spaced out over every few frames.We devise an allocation strategy similar to the idea of [5] which assigns time slices in the inverse ratio of the channel service rates, though our scheme is more flexible terms of the fairness and QoS support.

We first derive the motivation for our proposed heuristic scheme from the effective capacity concept. We then analyze the fairness and system performance provided by the scheme. Our numerical and simulation results show that our scheme provides significant advantage over the conventional scheme where just the queue length or just the channel condition are considered for resource allocation and also over the scheme which considers the queue information and does resource allocation in inverse proportion of the channel rates of the users. Our approach provides a wide range of flexibility in deciding system level fairness and the system performance which can be chosen on the basis of QoS requirements of different users by varying the channel information weight.

II. Problem Formulation

Let us consider a BS in a cellular system or a WiMAX system with N number of active mobile users. We focus on the uplink scheduling, although our approach can be extended to downlink as well. The service rate of a user is considered fixed in a particular frame but varies across different frames due to channel fading. A 2-user scenario is depicted in Fig. 1. For simplicity, we consider all users have a constant and equal packet generation rate, however the analysis can be easily extended to the case of users having different packet generation rate. We assume that the average transmission power of all the users is equal. Each user requests for some bandwidth from the base station to serve its traffic queue. An efficient service mechanism should use channel information as well as queue state to ensure fairness and predefined QoS (e.g., delay or loss bound) of individual users.

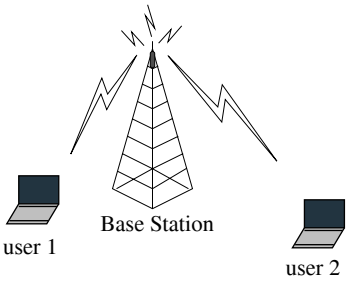

Fig. 1. A two-user scenario served by a base station.

As pointed out in [5] and [6], the signal-to-noise ratio (SNR) at each user is different due to different distances from the BS. Moreover there can be some users facing much higher channel fading environment than the others. By Shanon's theorem, the achievable transmission rate or the channel service rate of different users are related with their corresponding SNRs, and thus each user can encounter different channel rate.

On the other hand, each user has some QoS constraint which should be satisfied by the base station. In real-time applications the probability of packet delay of a user exceeding a maximum delay bound D_{max} should be less than a predefined threshold ϵ, i.e., $\Pr[D > D_{max}] \leq \epsilon$.

In a TDMA based allocation (as in WiMAX), the job of a BS is to optimally allocate the number of slots to the users by accounting their individual channel rate and queue status. Let the i-th user transmits for a time duration $t_i(n)$ in the n-th frame. Considering the transmission time allocation for all users per frame and assuming the traffic demand is at least equal to the resources available, we have:

$$\sum_{i=1}^{i=N} t_i(n) = T_f, \tag{1}$$

where T_f is the frame length of fixed size and $0 \leq t_i(n) \leq T_f$.

Our aim is to devise a fair resource allocation algorithm that assigns each user a time slice (number of slots) per frame based on its queue status and also the available channel rate.

The relation between QoS based on bounded delay D_{max} and the average waiting time W of a packet in a queue, as given in [11], can be approximated by:

$$\Pr[D > D_{max}] \simeq \exp(-\frac{D_{max}}{W}), \tag{2}$$

where D is the packet delay in a queue and $W = E[D]$. From (2) it is clear that if we minimize W we can minimize the delay violation of a queue.

Let $q_i(n)$ and $r_i(n)$ respectively be the queue length (bits) and channel rate of the i-th user in the n-th frame. The queue length in the frame $(n+1)$ can be related as:

$$q_i(n+1) = q_i(n) - r_i(n)t_i(n) + a_i(n), \tag{3}$$

where $a_i(n)$ is the additional arrival during during the n-th frame. The time slice allocation is done by the base station such that for each user

$$q_i(n) \geq r_i(n)t_i(n). \tag{4}$$

Assuming that there is a time window of length T_w, the average queue length over the time window of user i in the n-th frame is given by:

$$\bar{q}_i(n) = (1-\rho_w)\bar{q}_i(n-1) + \rho_w q_i(n), \tag{5}$$

where $\rho_w = \frac{T_f}{T_w}$, which is a small value.

A. Average Waiting Time

By Little's theorem, average waiting time in the n-th frame:

$$W_i(n) = \frac{\bar{q}_i(n)}{\lambda_i}.$$

Thus, from (3) and (5) we obtain:

$$W_i(n+1) = (1-\rho_w)W_i(n) + \rho_w \frac{q_i(n)}{\lambda_i} + \rho_w T_f - \rho_w \frac{r_i(n)t_i(n)}{\lambda_i}. \tag{6}$$

From (2) it is clear that we need to minimize $\sum_{i=1}^{N} W_i$ so as to minimize the average delay violation of the system. To achieve this, As evident from (6), we need to maximize

$$\sum_{i=1}^{i=N} r_i(n)t_i(n),$$

which implies that we should allocate more resource to the users having higher channel gains. However, though this approach provides high resource utilization, it lacks fairness. For some users the QoS performance will be well-guaranteed, but QoS to the other users will suffer quite badly. Therefore, a fairer resource allocation approach is needed.

Taking the expectation over W in (6) and simplifying, we have the average packet delay of user i as:

$$W_i = E\left[\frac{q_i(n)}{\lambda_i}\right] - E\left[\frac{r_i(n)t_i(n)}{\lambda_i}\right] + T_f. \tag{7}$$

Using (4), W_i is minimum when

$$E\left[\frac{q_i(n)}{\lambda_i}\right] - E\left[\frac{r_i(n)t_i(n)}{\lambda_i}\right] = 0. \tag{8}$$

Assuming $r_i(n)$ and $t_i(n)$ to be independent we obtain

$$E[t_i(n)] = \frac{E[q_i(n)]}{E[r_i(n)]}. \tag{9}$$

Hence the average serving time of a user should be directly proportional to the average queue length and inversely proportional to its channel rate. It may be observed that, while (9) gives us a fair resource allocation criteria, we are interested in a frame-by-frame *dynamic* allocation scheme *rather than average estimate* which, besides addressing fairness, should also maximize the overall system performance.

III. Proposed Heuristic Scheme

A. QoS Provisioning

We need to determine the exact relation between serving time alloted to a user as a function of its channel rate and queue information. By the effective capacity concept [8], the delay violation probability of user i is given by,

$$\Pr[D > D_{max}]_i = \gamma_i \exp(-\theta_i D_{max}) = \epsilon_i, \tag{10}$$

where $\theta_i = \frac{\gamma_i \mu_i}{\mu_i \tau_i + q_i}$, γ_i is the probability that the traffic of user i is in service, q_i refers to the number of bits in its queue, τ_i refers to the service time of a certain (fixed) number of bits of user i, and μ_i is its service rate in a frame.

For the n-th frame and i-th user, $\mu_i = r_i(n)t_i(n)$. If the QoS bound ϵ_i is known, from the channel and queue information of the user we can calculate μ_i and hence get the value of $t_i(n)$ that should be alloted to the i-th user out of the frame time T_f to guarantee a particular level of QoS.

For the proof of concept, we consider a two-user system where the channel rate of one user is $r_1(n)$ and that of the other is $r_2(n) = R \cdot r_1(n)$ in the n-th time frame. The ratio of channel service rates of the two users R is assumed constant in a frame. The alloted time by the BS to the respective users is calculated using (9) and (10) as follows.

$$t_1(n) = \frac{q_1(n)\ln(\gamma_1(n))}{(1-\tau_1(n))r_1(n)\gamma_1(n)}, \text{ and} \tag{11}$$

$$t_2(n) = \frac{q_2(n)\ln(\gamma_1(n)/f(R))f(R)}{(R-\tau_1(n))r_1(n)\gamma_1(n)}, \tag{12}$$

where $\frac{\gamma_1(n)}{\gamma_2(n)} \triangleq f(R)$ is intuitively an increasing function of R, and $\frac{\tau_1(n)}{\tau_2(n)} = R$. Dividing the (11) by (12) we obtain

$$\frac{t_1(n)}{t_2(n)} = \frac{f_1(R)q_1(n)}{q_2(n)},$$

where $f_1(R)$ is an increasing function of R. Because of the complexity of the function f_1 we devise a heuristic time allocation scheme by the base station in which

$$\frac{t_1(n)}{t_2(n)} = \frac{R^\alpha q_1(n)}{q_2(n)} = \frac{r_2(n)^\alpha q_1(n)}{r_1(n)^\alpha q_2(n)}. \tag{13}$$

The tuning parameter α $(0 \leq \alpha \leq 1)$ depends on the channel statistics of the users and can be tuned according to the variation in channel.

B. Resource Allocation Fairness

In (1) we have the magnitude of time slice assigned to each user in the n-th frame. If we denote $\phi_i(n)$ to be the fraction of frame time allocated to the i-th user $(i = 1,\ 2)$,

$$\phi_i(n) = \frac{q_i(n) \times r_i(n)^{-\alpha}}{q_1(n) \times r_1(n)^{-\alpha} + q_2(n) \times r_2(n)^{-\alpha}}. \tag{14}$$

The traffic served for user i in the n-th frame is proportional to $\phi_i(n)r_i(n) \triangleq \phi_i'(n)$. To allocate the time resource fairly to

the users we use Jain's fairness index [12] as:

$$F.I. = \frac{\left\{\phi_1'(n) + (\phi_2'(n)\right\}^2}{2\left\{\left(\phi_1'(n)\right)^2 + \left(\phi_2'(n)\right)^2\right\}}. \quad (15)$$

Now, from (7) we observe that if the total packets served in a frame is maximized, the total average delay and hence $\Pr[D > D_{max}]$ in (2) would be minimized. Total packets served in the n-th frame is $\{t_1(n)r_1(n) + t_2(n)r_2(n)\}$. Substituting (14) in the expression for total packets served it can be noted that, an optimum value of α has to be chosen so as to provide the best achievable QoS and satisfy the maximum possible fairness constraint.

IV. Simulation and Results

We conducted MATLAB simulation to verify out analytic observations. Each user generates Poisson distributed traffic at a fixed rate. We tested the proposed scheme in AWGN channel as well as Rayleigh fading channel, with a perfect knowledge of the channel assumed to be known at the transmitter. The channel rates were assumed constant in a frame duration but different for different users, which could be caused by different path losses at different distances from the BS.

A. Simulation Settings

AWGN channel with different rates were simulated by assigning different average SNRs at different users. With a perfect knowledge at the transmitter, the channel rate is equal to instantaneous capacity. An AWGN channel having a capacity r_0 is related to the average SNR as:

$$r_0 = B\log_2(1 + SNR),$$

where B is the channel bandwidth. Instantaneous service rate of the Rayleigh fading channel is given by

$$r_n = B\log_2(1 + x_n^2),$$

where x_n^2 is the power gain at the n-th frame. The channel rate of a Rayleigh faded AWGN channel can be calculated as:

$$r_n = r_0\frac{\log_2(1 + x_n^2)}{\log_2(1 + SNR_{avg})}. \quad (16)$$

Rayleigh fading channels were simulated by taking the distribution of x^2 to be exponential and calculating the channel rate from 16. A fading margin was provided for each user, so as to guarantee a minimum QoS. Each frame is of $T_f = 20$ ms size and the maximum tolerable delay D_{max} of any packet is 50 ms. The channel service rate is considered to be 1 Mbps at a power gain of 32. Only uplink scheduling was studied, although the proposed scheme is applicable in uplink as well as downlink scheduling. We simulated a fixed subcarrier allocation to the users. Therefore we considered the sum of channel utilization of the users to be limited, which can be increased if this scheme is used in conjugation with subcarrier allocation as mentioned earlier.

B. Performance of the proposed scheme

We start with a two-user system to provide an insight of the performance optimality, and then extend it to multiple-users.

A two-user system with AWGN channel is considered where both users' traffic arrival rate is $\lambda = 1200$ packets/s with each packet of size 70 Bytes. Channel rate of user-1 is 1 Mbps and that of user-2 is 2.4 Mbps. It can be noted that the channel utilization of user-1 is quite higher than that of user-2 due to a better channel condition for the second user.

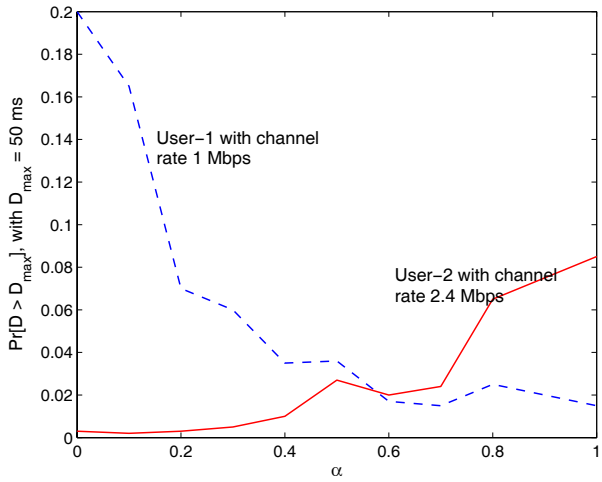

Fig. 2. QoS and fairness performance in AWGN channel.

Performances of the two users are shown in Fig. 2. It can be seen that, for user-1 the probability of violating the delay bound reduces drastically while that for user-2 increases gradually with an increase in α. A nearly equal performance of the two users is achieved at an α value between 0.5 and 0.6. Thus, by optimally choosing an α, compared to the QoS to user-1 at $\alpha = 0$ an about 10 fold advantage (from about 20% to 2% probability of delay bound violation) is achieved with a marginal decrease of QoS to user-2. Note that, though user-1 has a bad channel, the system is still able to satisfy the QoS constraint as it benefits from the low channel utilization of the other user. Moreover, if user-2 has a less-stringent QoS bound (for example if it has data traffic) as compared to user-1, then even a higher value of α can be chosen to optimally guarantee the performances of the two users.

In general, a suitable value of α can be flexibly decided in $[0, 1]$ based on different QoS constraints of the users. $\alpha = 0$ corresponds to the allocation of transmission times purely based on queue information, without any importance to the channel rate information. $\alpha = 1$, on the other hand, corresponds to allocating the transmission times based on the respective queue information and inverse proportion to the channel rates. As evident from Fig. 2, an optimal value of α provides a better and more fairer system performance compared to the conventional schemes with $\alpha = 0$ or $\alpha = 1$.

To show the applicability of the proposed heuristic to dif-

ferentiated service classes, we consider user-1 is delay tolerant and user-2 is delay sensitive with a packet loss constraint of 6%. Their performances are studied in a Rayleigh fading environment. The channel rate of user-1 is taken 0.8 Mbps (lower SNR) and that of user-2 is 2.65 Mbps. Fig. 3 shows

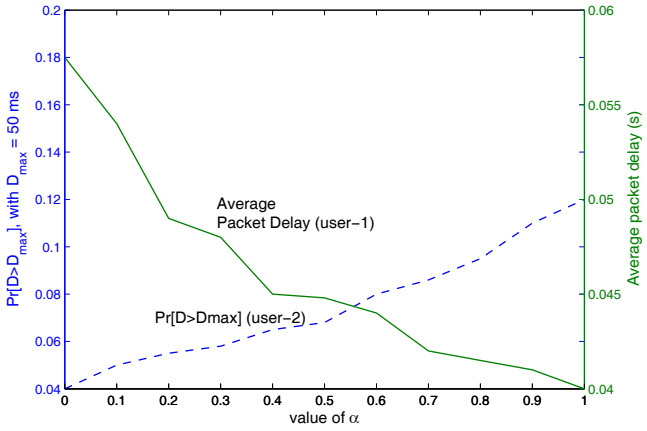

Fig. 3. Delay and loss performance in Rayleigh fading channel.

that, as in single-class users, an optimum α can be decided depending on the loss and delay bounds.If we choose the value of $\alpha = 0.3$ we satisfy the loss constraint of the second user and also minimize the delay of user-1 from 57 ms to 47 ms.

Fig. 4 shows the effect of channel gain ratio R of two users on the system performance gain in terms of the average packet delay in our scheme over the conventional one when channel state information is not used. It can be seen that, with widely

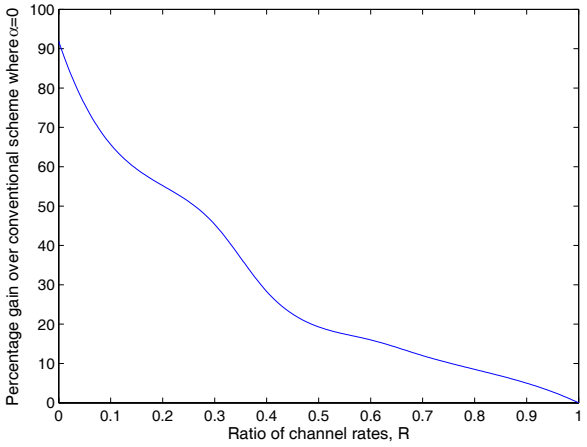

Fig. 4. Percentage QoS gain with varing service rate ratios.

different channel rates the performance gain of the proposed approach is as high as 90%. However, as the channel rates tend to be identical, the gain of the proposed scheme reduces sharply, reducing to zero when the channel gains are identical. This is because, the proposed approach exploits from the users with low channel utilization (higher channel rates). As the channels tend to become identical, this room shrinks and the corresponding gain of the approach reduces.

V. Conclusion

We have proposed a heuristic channel-and-queue aware dynamic time resource allocation scheme for link layer scheduling in a multiuser wireless access environment that exploits the user channels having low utilization to offer a significantly improved overall system performance and at the same time ensuring fairness to users. It allows to flexibly assign weight on the channel quality in achieving a desired system performance optimality. The proposed approach is quite simple to implement with less overhead.

In the current work, an analytic proof-of-concept of performance optimality is given without explicitly addressing the service class differentiation. The proposed approach is however applicable to differentiated service classes, which we will extensively look into in our future work.

References

[1] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. Murch, "Multiuser OFDM with adaptive subcarrier, bit, and power allocation," *IEEE J. Sel. Areas in Commun.*, vol. 17, no. 10, pp. 1747–1758, Oct. 1999.

[2] S. Coleri, M. Ergen, A. Puri, and A. Bahai, "Channel estimation techniques based on pilot arrangement in OFDM systems," *IEEE Trans. Broadcasting*, vol. 48, no. 3, pp. 223–229, Sept. 2002.

[3] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in wireless networks," *IEEE J. Sel. Areas in Commun.*, vol. 24, no. 8, pp. 1452–1463, Aug. 2006.

[4] M. Rabby and K. Ravindran, "Dynamics of end-to-end bandwidth allocations in qos-adaptive data connections," in *Proc. IEEE Wksp. Local and Metropolitan Area Networks*, Princeton, NJ, USA, June 2007.

[5] P. Bender, P. Black, M. Grob, R. Padovani, N. Sindhushayana, and A. Viterbi, "CDMA/HDR: A bandwidth-efficient high-speed wireless data service for nomadic users," *IEEE Commun. Mag.*, July 2000.

[6] G. Song, Y. Li, L. J. Cimini, Jr., and H.-T. Zheng, "Joint channel-aware and queue-aware data scheduling in multiple shared wireless channels," in *Proc. IEEE Wireless Communications and Networking Conference*, Atlanta, GA, USA, Mar. 2004.

[7] T. Ali-Yahiya, A.-L. Beylot, and G. Pujolle, "An adaptive cross-layer design for multiservice scheduling in OFDMA based mobile WiMAX systems," *Elsevier Computer Commun.*, vol. 32, pp. 531–539, Feb. 2009.

[8] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. Wireless Commun.*, vol. 2, no. 4, pp. 630–643, July 2003.

[9] C.-S. Chang and J. A. Thomas, "Effective bandwidth in high speed digital networks," *IEEE J. Sel. Areas in Commun.*, vol. 13, no. 6, pp. 1091–1100, Aug. 1995.

[10] J. Tang and X. Zhang, "Cross-layer-model based adaptive resource allocation for statistical qos guarantees in mobile wireless networks," in *Proc. ACM Intl. Conf. Quality of Service in Heterogeneous Wired/Wireless Networks*, Waterloo, Ontario, Canada, Aug. 2006.

[11] S. Asmussen, *Applied Probability and Queues*. Springer, 2nd ed., 2003.

[12] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination for resource allocation in shared systems," in *Research Report, DEC-TR-301*, Digital Equipment Corporation, Hudson, MA, USA, Sept. 1984.