Cross-Layer Design for OFDMA Wireless Systems With Heterogeneous Delay Requirements

David Shui Wing Hui, Vincent Kin Nang Lau, *Senior Member, IEEE*,
and Wong Hing Lam, *Senior Member, IEEE*

***Abstract*— This paper proposes a cross-layer scheduling scheme for OFDMA wireless systems with heterogeneous delay requirements. We shall focus on the cross-layer design which takes into account both queueing theory and information theory in modeling the system dynamics. We propose a delay-sensitive cross-layer design, which determines the optimal subcarrier allocation and power allocation policies to maximize the total system throughput, subject to the individual user's delay constraint and total base station transmit power constraint. The delay-sensitive power allocation was found to be multilevel water-filling in which urgent users have higher water-filling levels. The delay-sensitive subcarrier allocation strategy has linear complexity with respect to number of users and number of subcarriers. Simulation results show that substantial throughput gain is obtained while satisfying the delay constraints when the delay-sensitive jointly optimal power and subcarrier allocation policy is adopted.**

***Index Terms*— Delay-sensitive cross-layer scheduling, heterogeneous applications, orthogonal frequency division multiple access (OFDMA), power control, subcarrier allocation.**

I. Introduction

OFDM has been proposed as a multiple access scheme for providing high speed data transmission over next generation networks such as the IEEE 802.16 Wireless Metropolitan Area Network because of its robust performance over the frequency selective channel. There are quite a number of existing works on cross-layer scheduling design for OFDMA systems such as [1]-[5] and references therein. The optimal transmit power adaptation and subcarrier allocation and the corresponding computational efficient suboptimal algorithm for the total transmit power minimization problem in an OFDMA system having users with fixed data rate requirements have been studied in [1] and [2] respectively, while the data rate maximization problem is considered in [3]. The authors in [4], [5] provided a general theoretical framework, as well as several practical algorithm implementation schemes, addressing the cross-layer optimization problem of OFDMA systems through using a general utility function based objective. However, these cross-layer designs, while achieving throughput gain by exploiting spectral diversity as well as multiuser diversity, were only based on a decoupled approach where source statistics and queue dynamics were decoupled (and ignored) from the physical layer information theoretical models. The negligence of the effect of the source statistics, queueing delays and application level requirements lead to inappropriate design from a higher layer system performance perspective, particularly upon the provision of diverse QoS requirements in terms of delay performance. On the other hand, initial attempts on a cross-layer scheduler design that incorporated both the source statistics and queue dynamics were reported in [6]-[8], [11], where a simple On-Off physical layer model was assumed in [6], and the multiple access channel model with homogeneous users was studied in [7], [8] through combined information theory [9] and queueing theory [10] with the objective to minimize the average system delay. In [11], a heuristic scheduler design maximizes the system throughput while providing fairness between users in an OFDMA system. Yet, all of these designs, were targeted for systems with homogeneous users only.

In this paper, we focus on delay-sensitive cross-layer scheduling design for OFDMA systems consisting of users with mixed traffics and heterogeneous delay requirements. Specifically, we propose optimal delay-sensitive subcarrier allocation and power allocation policies to maximize the total system throughput while at the same time, satisfying heterogeneous user delay requirements. The proposed optimization framework involves both information theory[1] (to model the multiuser OFDMA physical layer) as well as queueing theory (to model the delay dynamics). By transforming the delay constraints into rate constraints, the delay-sensitive cross-layer scheduling problem is formulated into a mixed convex and combinatorial optimization problem. The optimal delay-sensitive power allocation strategy is given by multi-level water-filling where a user with tighter delay constraint will be assigned a higher "water-level." The optimal delay-sensitive subcarrier allocation strategy is shown to be decoupled between subcarriers (i.e. greedy in nature) with a linear complexity with respect to the number of users and number of subcarriers. An iterative algorithm for finding the "multi water-levels" of heterogeneous users is also proposed.[2]

This paper is organized as follows. In Section II, we

Manuscript received September 9, 2005; revised August 2, 2006; accepted June 4, 2007. The associate editor coordinating the review of this paper and approving it for publication was X. Shen.

D. S. W. Hui and V. K. N. Lau are with the ECE Department, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong (e-mail: {davidhui, eeknlau}@ee.ust.hk).

W. H. Lam is with the EEE Department, University of Hong Kong, Pokfulam Road, Hong Kong (e-mail: whlam@eee.hku.hk).

Digital Object Identifier 10.1109/TWC.2007.05716.

[1] In contrast to a simple ON-OFF model [6], we consider a more sophisticated information theoretical model to capture the performance of the OFDMA physical layer.

[2] We have also worked out the asymptotic multiuser diversity with heterogeneous delay constraints based on our analytical model. However, due to page limitation, this part is removed. Interested readers will please refer to our URL for a longer version of our paper.

1536-1276/07$25.00 © 2007 IEEE

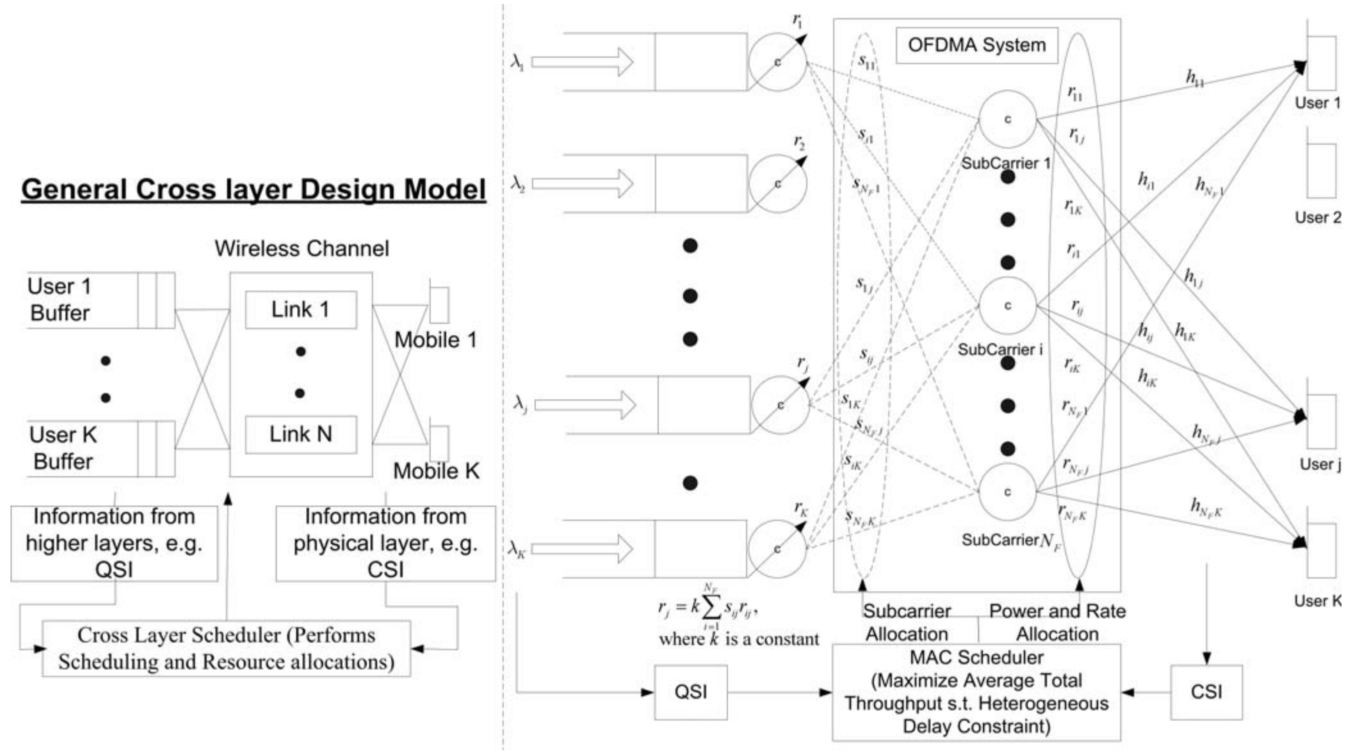


Fig. 1. General Cross-Layer System Model (left) and Cross-Layer Scheduling Model under the Conceptual Channel Model for an OFDMA system with heterogeneous application users (right).

describe the system model, including channel model, physical layer model, source model and MAC layer model. Section III presents the formulated optimization problem, and the corresponding delay-sensitive power and subcarrier allocation policies are presented in Section IV. Simulation results are studied in Section V and a conclusion is given in Section VI.

II. System Model

The general cross-layer system model of a multiuser wireless system and the specific system architecture of a multiuser downlink OFDM scheduler are shown in Fig. 1. Before the scheduling operation is performed, the cross-layer resource scheduler first collects the QoS (delay) requirements of all users. In the beginning of each scheduling interval, the resource scheduler in the base station (BS) obtains channel state information (CSI) through the uplink dedicated pilots from all mobile users[3] and collects queue state information (QSI) by observing the number of backlogged packets in all these users' buffers. The resource scheduler then makes a scheduling decision based on this information and passes the resource allocation scheme to the OFDMA transmitter. The update process of state information of all users and also the scheduling decision process are made once every time slot. The subcarrier allocation and power allocation decision made by the BS transmitter is assumed to be announced to each mobile user through a separate control channel. We further assume perfect channel state information is available at the transmitter (CSIT) and receiver (CSIR), and the transmission rate chosen from a continuous set is realizable according to the channel characteristic and with perfect channel coding on each subcarrier.

A. Channel Model

We consider an OFDMA system with a quasi-static fading channel within a scheduling slot (2ms). This is a reasonable assumption for users with pedestrian mobility where the coherence time of the channel fading is around 20ms or more. Due to OFDMA, the N_F subcarriers are decoupled. Let i denote the subcarrier index and j denote the user index. The received symbol Y_{ij} at mobile user j on subcarrier i is given by

$$Y_{ij} = h_{ij}X_{ij} + Z_{ij} \tag{1}$$

where X_{ij} is the data symbol from the BS to user j on subcarrier i, h_{ij} is the complex channel gain of the i-th subcarrier for the j-th mobile which is zero mean complex Gaussian with unit variance and Z_{ij} is the zero mean complex Gaussian noise with unit variance. The transmit power allocated from the BS to user j through subcarrier i is given by $p_{ij} = E\left[|X_{ij}|^2\right]$. We define a subcarrier allocation strategy $S_{N_F \times K} = [s_{ij}]$, where $s_{ij} = 1$ when user j is selected to occupy subcarrier i, otherwise $s_{ij} = 0$. The average total transmit power from the BS is constrained by P_{TOT}, i.e. $E\left[\frac{1}{N_F}\sum_{j=1}^{K}\sum_{i=1}^{N_F} s_{ij}p_{ij}\right] \le P_{TOT}$, where P_{TOT} is average total available transmit power in BS.

[3] In this paper, we consider OFDMA with TDD systems. Hence downlink CSIT can be obtained from channel reciprocity through CSIT estimation of uplink dedicated pilots. For an FDD system, explicit feedback of downlink CSIT from mobile users is required.

B. Multi-User Physical Layer Model for OFDMA Systems

In order to decouple the problem to be formulated in this paper from specific implementation of coding and modulation schemes, we consider information theoretical Shannon's capacity as the abstraction of the multi-user physical layer model. Given the CSIT h_{ij} and $s_{ij} = 1$, the maximum achievable data rate c_{ij}[4] (bits/s/Hz) conveying from the BS to user j through subcarrier i, during the current fading slot, is given by the maximum mutual information between X_{ij} and Y_{ij}, which can be written as

$$c_{ij} = \max_{p(X_{ij})} I\left(X_{ij}; Y_{ij} | h_{ij}\right) = \log\left(1 + p_{ij}\left|h_{ij}\right|^2\right) \qquad (2)$$

where $I\left(X_{ij}; Y_{ij} | h_{ij}\right)$ denotes the conditional mutual information. As long as the scheduled data rate $r_{ij} \le c_{ij}$, this Shannon's capacity can be achieved by random codebook and Gaussian constellation at the BS.[5] We also represent the transmission rate (scheduled at maximum achievable data rate) in matrix form by $R_{N_F \times K} = [r_{ij}]$ with the individual matrix element equal to $r_{ij} = c_{ij}$.

C. Source Model

Packets from heterogeneous user applications come into each user j's buffer according to a Poisson process with independent rate λ_j packets per time slot with packets of fixed size consisting of F bits without packets overflow. The nature of user j is characterized by a tuple $[\lambda_j, T_j]$, where λ_j is the average packet arrival rate to user j and T_j is the delay constraint requirement by user j. User j with heavier traffic load will have a higher λ_j and more delay-sensitive user j will have stringent delay requirements T_j (smaller T_j value).

D. MAC Layer Model

The system dynamics are characterized by system state $\mathcal{X} = (H_{N_F \times K}, Q_K)$, which is composed of channel state $H_{N_F \times K} = \left[\left|h_{ij}\right|^2\right]$ and buffer state Q_K, where $Q_K = [q_j]$ is a $K \times 1$ vector, with the j-th component denoting the number of packets remaining in user j's buffer. The MAC layer is responsible for the cross-layer scheduling channel resource allocation at every fading block based on the current system state $\mathcal{X}$ as illustrated in Fig. 1. At the beginning of every frame, the BS estimates the CSIT from dedicated uplink pilots. Based on the CSIT and the queue states obtained, the scheduler determines the subcarrier allocation from the policy $S_{N_F \times K}[H, Q]$, the power allocation from the policy $P_{N_F \times K}[H, Q]$ and the corresponding rate allocation from the policy $R_{N_F \times K}[H, Q]$ for the selected users, in each scheduling slot. The scheduling results are then broadcasted on downlink common channels to all users before subsequent downlink packets transmissions at scheduled rates.

III. Problem Formulation

In this section, the OFDMA cross-layer design problem for heterogeneous users is formulated as a constrained optimization problem based on the system model introduced in Section II. The objective is to maximize total system throughput while maintaining OFDMA physical layer constraints on subcarrier selection, transmission power constraint and delay constraints. The optimization problem is formulated as follows:

Cross-Layer Problem Formulation: Find the optimal subcarrier and power allocation policies ($S_{N_F \times K}[H, Q]$, $P_{N_F \times K}[H, Q]$) such that:

$$\max_{S,P} E\left(\sum_{i=1}^{N_F}\sum_{j=1}^{K} s_{ij} r_{ij}\right)$$

subject to:

$$\begin{aligned} &(\mathrm{C1}): s_{ij} \in \{0, 1\},\ (\mathrm{C2}): \sum_{j=1}^{K} s_{ij} = 1,\ (\mathrm{C3}): p_{ij} \ge 0, \\ &(\mathrm{C4}): E\left[\frac{1}{N_F}\sum_{j=1}^{K}\sum_{i=1}^{N_F} s_{ij} p_{ij}\right] \le P_{TOT}, \\ &(\mathrm{C5}): E\left[\widetilde{W}_j\right] \le T_j \quad \forall x, i, j \end{aligned} \qquad (3)$$

where $\widetilde{W}_j$ is the system time (the duration of staying in the system) of user j's packet in system state $\mathcal{X} = (CSI, QSI)$, and the rate allocation r_{ij} from policy $R_{N_F \times K}$ is related to the power allocation from policy $P_{N_F \times K}$ by $r_{ij} = c_{ij} = \log_2\left(1 + p_{ij}\left|h_{ij}\right|^2\right)$ as described in Section II-B.

In optimization problem (3),[6] constraints (C1) and (C2) are used to ensure only one user can occupy a subcarrier i at one time. (C3) is used to ensure transmit power would only take a positive value, (C4) is the average total power constraint, and (C5) is the average delay constraint where the average system time of user j's packet $E\left[\widetilde{W}_j\right]$[7] (including average waiting time and average service time) is required to be smaller than user j's delay requirement T_j. We assume that the arrival rates of the system are large enough so that there are always packets in the user queues to be scheduled.

A. Relationship Between Scheduled Data Rate and Delay Parameters

Before we can solve optimization problem (3), we have to express the delay constraint in terms of physical layer parameters. We shall have Lemma 1 from queueing analysis.

[4] c_{ij} is called "instantaneous channel capacity," and does not require to be achieved by an "infinite delay" random codebook. In slow fading channels, the channel fading remains quasi-static within each scheduling slot. The random coding only spans across one scheduling slot causing only an finite delay.

[5] In practice, the Shannon's Capacity could be approximately achieved by powerful coding such as turbo code and LDPC, provided perfect channel state information is available. For example, in a 802.11n WLAN system, the packet length is 0.5ms, which is much less than the coherent time, and the packet size is 4kBytes = 32kbits, which is more than sufficient for powerful codes (such as turbo code and LDPC code) to have close-to-capacity performance.

[6] In Problem (3), the expectation operator $E[.]$ is taken over all system states $\mathcal{X} = \left(H_{N_F \times K}, Q_K\right)$. It is noted that the subcarrier s_{ij} and power allocation p_{ij} result are functions of CSI $\left|h_{ij}\right|^2$ and QSI q_j. Though s_{ij} and p_{ij} are not random given a state realization, the constraint (C4) refers to the "average power constraint" where "average" (expectation operator in (C4)) refers to average over random realizations of CSIT and QSI. This "average" operator is also applied to the average delay constraint (C5).

[7] The system time of user j's packet consists of two components: one is the waiting time, which is the duration from the time of arrival to the starting time of service (start being encoded); another component is the service time, which is the duration from the starting time of service to the end of service (the time that the system completes the encoding of this packet and starts encoding another packet if the buffer is not empty).

Lemma 1: A necessary and sufficient condition for the constraint (C5) is

$$E\left[X_j\right] + \frac{\lambda_j E\left[X_j^2\right] + \lambda_j E\left[X_j\right]\left(E\left[\overline{S_j}\right]/E\left[S_j\right]\right)(t_s)}{2\left(1-\lambda_j\left(E\left[X_j\right]/E\left[S_j\right]\right)\right)} \le T_j \tag{4}$$

where X_j is the service time of the packet of user j, λ_j is the arrival rate of user j, T_j is the average delay requirement of user j, and t_s is the duration of the scheduling slot. Note that S_j and $\overline{S_j}$ are indicator variables for the availability and unavailability of the subcarrier for user j respectively, i.e. $(s_j(m) = 1, \overline{s_j}(m) = 0)$ if there is a subcarrier allocated to user j at time slot index m, or $(s_j(m) = 0, \overline{s_j}(m) = 1)$ otherwise. In a practical OFDMA system, the number of subcarrier N_F is usually much greater than number of user K, thus there is always a subcarrier available for any particular user j, i.e. $E\left[S_j\right] = 1$ and $E\left[\overline{S_j}\right] = 0$.

From Lemma 1, the constraint (C5) is ready to be transformed to an equivalent rate constraint that directly relates the scheduled data rate R_j of user j to the user characteristic tuple $[\lambda_j, T_j]$, and also the packet size F.

Corollary 1: A necessary and sufficient condition for the constraint (C5) when $T_j \to \infty$ is $E\left[S_j R_j\right] \ge F\lambda_j$.

This corollary shows that average scheduled data rate $E\left[S_j R_j\right]$ of user j should be at least the same as the bits arrival rate to user j's queue (even without any delay requirement) in order to guarantee the *stability* of the queue.

Corollary 2: A necessary condition for the constraint (C5) is called the equivalent rate constraint $E\left[S_j R_j\right] \ge \rho_j$, where

$$\rho_j = \frac{(2T_j\lambda_j + 2) + \sqrt{(2T_j\lambda_j + 2)^2 - 8T_j\lambda_j}}{4T_j} F \tag{5}$$

Proof: The proofs of Lemma 1, Corollary 1 and Corollary 2 are presented in Appendix A. ■

IV. Scheduling Strategies

The optimization problem (3) can be written as a mixed combinatorial (with respect to $\{s_{ij}\}$) and convex optimization problem (with respect to $\{p_{ij}\}$). For each possible subcarrier allocation $\{s_{ij}\}$, we can compute the optimal power allocation $\{p_{ij}\}$ for the selected user over each individual subcarrier and the corresponding user data rates $\{r_{ij}\}$. Based on the computed data rate vector $(r_{11} \ldots, r_{N_F K})$, the total system throughput $\sum_{i=1}^{N_F}\sum_{j=1}^{K} s_{ij} r_{ij}$ can be evaluated. We can evaluate the total system throughput for all different cases by enumerating all possible combinations of $\{s_{ij}\}$ and the one that gives the largest average throughput will be the optimal solution. However, based on the exhaustive search approach for $\{s_{ij}\}$, the total search space is N_F^K which is not feasible for moderate N_F. We shall illustrate that the optimal search for $\{s_{ij}\}$ can be decoupled between the N_F subcarriers and hence the proposed subcarrier allocation is computationally efficient with the complexity of $N_F \times K$ only.

Specifically, using Corollary 2, the optimization problem (3) can be reformulated with constraint (C5) written as:

$$(C5)\text{: } E\left[\sum_{i=1}^{N_F} s_{ij}\log_2\left(1 + p_{ij}\left|h_{ij}\right|^2\right)\right] \ge \widetilde{\rho}_j \tag{6}$$

where $\widetilde{\rho}_j = \rho_j \times \left(\frac{1}{t_s}/\frac{BW}{N_F}\right)$ and BW is the total bandwidth of the OFDMA system. This optimization problem (3) is now a mixed combinatorial and convex optimization problem. In order to make the problem more traceable, we relax the integer constraint on $s_{ij} = \{0, 1\}$ to a time sharing factor between 0 and 1 with problem reformulation using the variable $\widetilde{p}_{ij} = p_{ij}s_{ij}$. The resultant problem would be a convex maximization problem [12]. Using standard techniques, the following Lagrangian of the reformulated problem is obtained:

$$\begin{aligned} L = &\sum_{j=1}^{K}(1+\gamma_j)\sum_{i=1}^{N_F} s_{ij}\log_2\left(1 + \frac{\widetilde{p}_{ij}\left|h_{ij}\right|^2}{s_{ij}}\right) \\ &- \mu\left(\sum_{j=1}^{K}\sum_{i=1}^{N_F}\widetilde{p}_{ij} - N_F P_{TOT}\right) \\ &- \sum_{j=1}^{K}\gamma_j\widetilde{\rho}_j + \sum_{i=1}^{N_F}\phi_i\left(\sum_{j=1}^{K}s_{ij} - 1\right) \end{aligned} \tag{7}$$

After finding the KKT condition through this Lagrangian, we get the following jointly optimal power and subcarrier allocation stated in Theorem 1.

A. Delay-Sensitive Jointly Optimal Power and Subcarrier Allocation

Theorem 1: Given the CSIT realization h_{ij}, the optimal subcarrier allocation policy $S_{opt}[H] = [s_{ij}]$ can be decoupled between N_F subcarriers and is given by:[8] *For each i within $1 : N_F$*

$$\begin{aligned} j^* = \arg\max_{j\in[1,K]}\, &(1+\gamma_j)\left(\log_2\left(\frac{(1+\gamma_j)}{\mu}\left|h_{ij}\right|^2\right)\right)^+ \\ &-\mu\left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{\left|h_{ij}\right|^2}\right)^+, s_{ij} = \begin{cases} 1, & j = j^* \\ 0, & \textit{otherwise}. \end{cases} \end{aligned} \tag{8}$$

The corresponding optimal power allocation policy $P_{opt}[H] = [p_{ij}]$ is given by:

$$p_{ij} = \begin{cases} \left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{\left|h_{ij}\right|^2}\right)^+, & \forall s_{ij} = 1 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $(x)^+$ means $\max(0, x)$, and μ, γ_j are the Lagrange multipliers satisfying the power constraint (C4) and delay constraint (C5) for all users j.

The search for the Lagrange multipliers requires a numerical procedure as follows. Denote $\{\gamma_1, \ldots, \gamma_K\}$ as γ. The Lagrange multipliers are obtained by solving a system of equations on $P(\mu, \gamma) = 0$, and $f_j(\mu, \gamma) = 0, \forall j$, where

$$P(\mu, \gamma) = P_{TOT} - E\frac{1}{N_F}\sum_{i=1}^{N_F}\sum_{j=1}^{K} s_{ij}\left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{\left|h_{ij}\right|^2}\right)^+,$$

$$f_j(\mu, \gamma) = \gamma_j\left(E\sum_{i=1}^{N_F} s_{ij}\left(\log_2\frac{(1+\gamma_j)\left|h_{ij}\right|^2}{\mu}\right)^+ - \widetilde{\rho}_j\right).$$

Notice that $f_j(\mu, \gamma) < 0$ means the delay constraint is violated and $P(\mu, \gamma) > 0$ means power P_{TOT} is not used up. There are

[8] Due to page limits, we have skipped the proof here. Interested readers will please refer to the longer version of the paper in our URL for the proof.

several ways to find the Lagrange multipliers through solving this system of equations iteratively. One way is to update the Lagrange multiplier variables γ and μ alternatively using the bisection method.[9] In each iteration, the upper portion of the interval for bisection on γ_j will be retained if $f_j(\mu, \gamma) < 0$, and the lower half of the interval for μ will be retained if $P(\mu, \gamma) > 0$. Another way to update the Lagrange multipliers is based on the gradient method [12].[10]

In Theorem 1, the optimal power allocation $P_{opt}[H] = [p_{ij}]$ expressed in (9) can be interpreted as a multi-level water-filling strategy. It means that those delay-sensitive users j with more stringent average delay requirements (having more urgent packets to be transmitted) have to be transmitted at a higher power water-level $(1 + \gamma_j)/\mu$ (where the value of γ_j depends on the urgency of the delay requirements). On the other hand, those delay-insensitive users j (i.e. those users with inactive delay constraint (C5)) are allocated with the same power water-level $1/\mu$. Furthermore, the optimal subcarrier allocation strategy (8) can be interpreted as a policy that user j with higher urgency level γ_j has a higher chance of being allocated subcarriers, while users with the same γ_j have the same chance and subcarriers are allocated to the user with the best CSIT among this user group. Besides, it can also be implemented by a greedy algorithm with linear complexity in terms of $N_F \times K$.

B. Minimal Power Required for Provision of Delay Requirements Guarantee

It should be noted that delay requirements may not always be feasible. There is a minimum average transmit power requirement ($P_{\min}$) in order to satisfy the delay requirements of all users. Given all the K users characteristic tuples $[\lambda_j, T_j]$, under the joint subcarrier and power allocation policy presented in (8) and (9), $P_{\min}$ is calculated by solving the system of equations:

$$\begin{cases} P_{\min} = E\left[\frac{1}{N_F}\sum\limits_{i=1}^{N_F}\sum\limits_{j=1}^{K} s_{ij}\left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{|h_{ij}|^2}\right)^+\right] \\ E\left[\sum\limits_{i=1}^{N_F} s_{ij}\left(\log_2\left(\frac{(1+\gamma_j)}{\mu}|h_{ij}|^2\right)\right)^+\right] = \widetilde{\rho}_j, \forall j \end{cases} \tag{10}$$

When $P_{TOT} \geq P_{\min}$, the delay constraints for all delay-sensitive users are active; otherwise, at least one of the delay constraints cannot be satisfied by any power and subcarrier allocation policy. Numerical examples on minimum required power are shown in Section V.

V. Simulation Results

In this section, we present the simulation results to illustrate the performance of the proposed cross-layer scheduler. We also provide some comparisons of the proposed cross-layer scheme with the FDMA-like schemes.

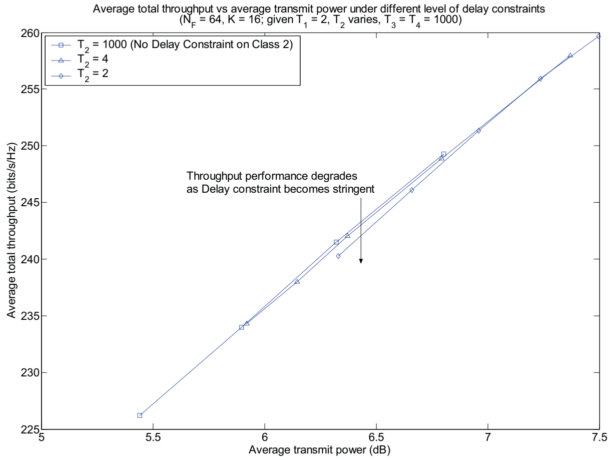


Fig. 2. Average total throughput vs average transmit power under different delay constraint T_2 of Class 2 users (T_2 = 2, 4, 1000 time slots). The number of users of each class is (K_1, K_2, K_3, K_4) = (4, 4, 4, 4), respectively.

A. Simulation Model

In our simulation, we consider an OFDMA system with total system bandwidth of 80 kHz consisting of 64 subcarriers. Thus each subcarrier has bandwidth of 1.25kHz and each subcarrier channel experiences flat fading. The duration of a scheduling slot is assumed to be 2ms. The channel fading between different users and different subcarriers is modeled as i.i.d. complex Gaussian with unit variance. We consider four classes of users in the system with arrival rates and delay requirements of each class being specified by $(\lambda, \mathbf{T}) = \{(0.3, 2), (0.4, 4), (0.5, 1000), (0.6, 1000)\}$ (packets per time slot, time slots). Class 1 and Class 2 users represent delay-sensitive traffic with heterogeneous delay requirements while Class 3 and Class 4 users represent delay insensitive applications with heterogeneous traffic loading. Each packet consists of 80 bits and each point in the figures is simulated from 10000 independent trials.

B. Simulation Results

1) Throughput performance of the proposed scheduler

Fig. 2 depicts the average total system throughput versus SNR under various delay constraints of a Class 2 user. It is observed that in a low SNR regime (below 7.4 dB), the system throughput is lower when the delay requirement of Class 2 users is more stringent. This is because more urgent users with heavy traffic loading will have higher water-levels and thus have higher chances of seizing subcarriers, causing losses in degree of freedom in exploiting throughput maximization by other users with better CSIT. Besides, the minimum required power to support all delay constraints of the user would increase as the delay requirements become more stringent. In a high SNR regime (above 7.4 dB), the throughput performance is the same regardless of the value of the imposed delay constraint for Class 2. This is because in a high SNR regime, the water-levels are the same for all users and thus the optimal subcarrier allocation reduces to the conventional delay-insensitive scheduling policy.

[9] We could also implement the Lagrange Multiplier Finding Algorithm using other root finding algorithms, e.g. Newton Raphson's Algorithm, for faster convergence.

[10] In the n-th iteration, μ and γ are updated through $\mu(n+1) = [u(n) - \alpha P(\mu(n), \gamma(n))]^+$, $\gamma_j(n+1) = [\gamma_j(n) - \alpha f_j(\mu(n), \gamma(n))]^+$ where α is the chosen stepsize.

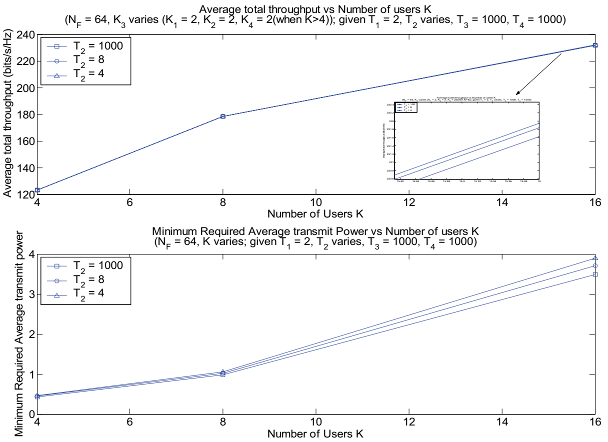


Fig. 3. (Upper) Average total system throughput vs different number of users K under different delay constraint T_2 of Class 2 users (T_2 = 4, 8, 1000 time slots). For K = 16, 8, 4 (K_1, K_2, K_3, K_4) = (2, 2, 10, 2), (2, 2, 2, 2), (2, 2, 0, 0), respectively. (Lower) Minimum required average transmit power vs different number of users K under different delay constraint T_2 of Class 2 users (T_2 = 4, 8, 1000 time slots). For K = 16, 8, 4 (K_1, K_2, K_3, K_4) = (4, 4, 4, 4), (2, 2, 2, 2), (1, 1, 1, 1), respectively.

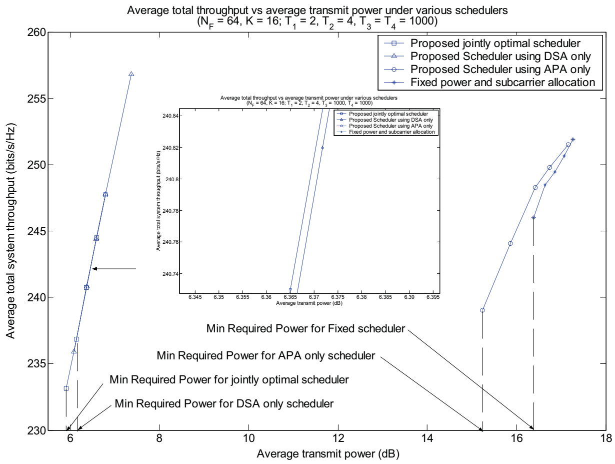


Fig. 5. Average total system throughput vs average transmit power under different schedulers when K = 16. (The number of users of each class is (K_1, K_2, K_3, K_4) = (4, 4, 4, 4), respectively.)

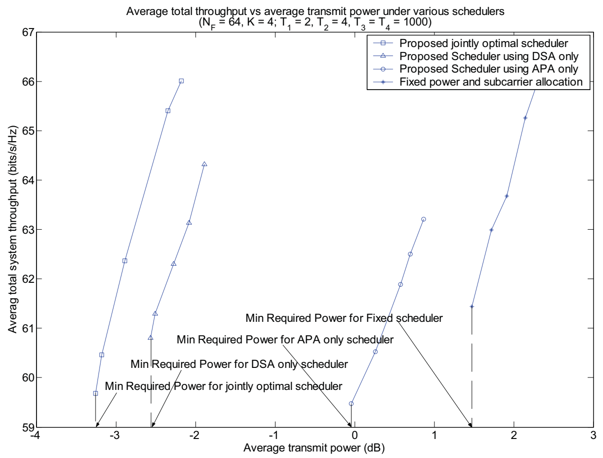


Fig. 4. Average total system throughput vs average transmit power under different schedulers when K = 4. (The number of users of each class is (K_1, K_2, K_3, K_4) = (1, 1, 1, 1), respectively.)

2) Impact of delay constraints on the throughput gain from multiuser diversity

In Fig. 3, the total system throughput versus number of users K is depicted for the case of SNR = 5.64 dB. It shows that the delay-sensitive cross-layer design can exploit multiuser diversity gain as well. However, the multiuser diversity gain decreases for systems with more stringent delay constraints. The minimum power required to support delay constraints of Class 1 and Class 2 users also increases as K increases.

3) Throughput comparison among various schedulers

Figs. 4 and 5 illustrate throughput performance versus SNR for various schedulers by considering a system with (K_1, K_2, K_3, K_4) = (1, 1, 1, 1) and (4, 4, 4, 4) respectively. In addition to the proposed delay-sensitive cross-layer scheduler (delay-sensitive joint dynamic subcarrier allocation and adaptive power allocation) [DS-DSA-APA], we consider two variants of the proposed delay-sensitive cross-layer schedulers, namely delay-sensitive adaptive power allocation (DS-APA) and delay-sensitive dynamic subcarrier allocation (DS-DSA). The DS-APA performs adaptive power allocation only based on (9) [using fixed subcarrier allocation] while the DS-DSA performs adaptive subcarrier allocation (8) only [using fixed power allocation]. From both Figs. 4 and 5, it can be seen that the DS-DSA-APA achieves the best system throughput. When (K_1, K_2, K_3, K_4) =(4,4,4,4), the DS-DSA is close to optimal. This is because when the number of users is large, the multiuser diversity gain ensures that the SNR per subcarrier is high and hence, power adaptation only provides marginal gains. On the other hand, when the number of users is smaller, the power adaptation becomes more important. In both cases, there is significant throughput gain of the proposed schemes relative to the conventional delay-insensitive FDMA-like scheduler. Figs. 4 and 5 also illustrate that the minimum power required to support the delay constraints of all users for the DS-DSA-APA is substantially reduced (4.5 dB for 4 users and 10.3dB for 16 users) compared to a conventional FDMA-like scheme (fixed allocation).

4) Impact of changes in traffic loading on delay performance of delay-sensitive users of the proposed scheduler

In Fig. 6, the average delay performance versus different arrival rates of delay insensitive Class 4 users is depicted given $P_{TOT} = 5.65$dB. It is observed that by using the proposed scheduler, with the increases in traffic loading of Class 4 users, the delay requirements of delay-sensitive users from Class 1 and Class 2 are still satisfied, while the only price to be paid is an increased average delay for those delay insensitive users from Class 3 and Class 4. Similarly, the average delay performance of delay-sensitive users from Class 1 and Class 2 can also be shown to be guaranteed when the arrival rates of other classes of users are increased, whenever the minimum power requirement is satisfied. Such a characteristic of delay performance guarantee is important for serving bursty delay-sensitive real time heterogeneous traffic in next generation wireless networks.

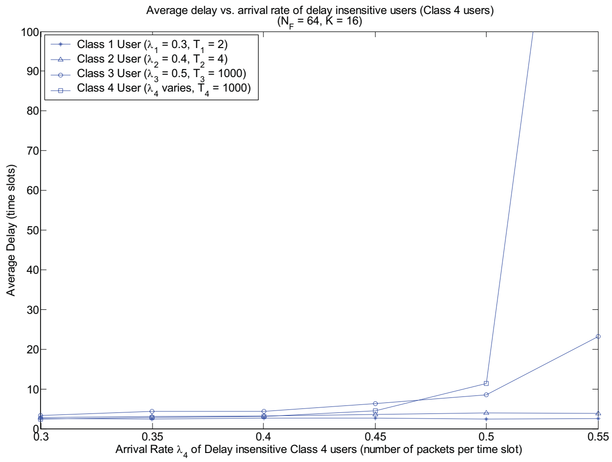


Fig. 6. Average delay vs arrival rate of delay insensitive user (Class 4 users). ($(K_1,\ K_2,\ K_3,\ K_4)$ = (4, 4, 4, 4))

VI. Conclusion

In this paper, we have presented a delay-sensitive cross-layer scheduler for OFDMA systems with heterogeneous delay requirements. Through a proper transformation of the delay constraints to the equivalent rate constraints, the cross-layer design problem is formulated as a convex optimization problem with consideration of the source statistics and queue dynamics as well as CSIT in the OFDMA systems. The proposed cross-layer scheduler offers a nice balance of maximizing throughput and providing QoS (delay) differentiation of the mixed heterogeneous users. From the simulation results, it was also shown that substantial throughput gain and minimum required power saving are achieved by a jointly optimal power and subcarrier allocation policy with all users' delay constraints being satisfied.

Appendix A
Proof of Lemma 1, Corollary 1, and Corollary 2

Proof: For an OFDMA system with Poisson arrival to each user's queue, suppose the service provided by all subcarriers for each user to be considered as a server that changes its service rate according to the system state, then the buffer status dynamic for each user can be modeled as an M/G/1 queue. However due to the subcarrier allocation process, the server may be idle due to no subcarrier being allocated to the user. As a result, modeling the distribution of the service rate of this server is highly complex, and the conventional Pollaczek-Khinchin formula [10] is inconvenient for the calculation of average system time $E\left[\widetilde{W}_j\right]$ for each user j in this situation.

Consider a particular user j's buffer. Let m denote the time slot index and $\widetilde{m}$ denote the packet index. The random variables representing the number of packets transmitted, availability of subcarrier, total scheduled data rate (bits/time slot) and the service time[11] for user j are denoted as N_j, S_j, R_j, and X_j respectively (randomness depends on the evolution of the system state across the time span), where $n_j(m)$, $s_j(m)$,[12] $r_j(m)$, and the service time of the $\widetilde{m}^{\text{th}}$ packet $x_{\widetilde{m},j} \triangleq 1/n_j(m) = F/r_j(m)$ [13] is the corresponding realization in the m^{th} time slot.

The average service time of user j (in terms of the number of time slots), denoted as $E[X_j]$, is written as (A.1):

$$E[X_j] = \lim_{M\to\infty} \frac{\frac{1}{M}\sum\limits_{m=1}^{M} s_j(m)}{\frac{1}{M}\sum\limits_{m=1}^{M} s_j(m)\, n_j(m)} = \frac{E[S_j]\,F}{E\left[\sum\limits_{i=1}^{N_F} s_{ij} r_{ij}\left(\frac{t_s BW}{N_F}\right)\right]} = \frac{E[S_j]\,F}{E[S_j R_j]}. \tag{A.1}$$

On the other hand, as shown in Fig. 7, the waiting time from the perspective of an arriving packet $\widetilde{m}$ is

$$w_{\widetilde{m},j}(t) = res_{\widetilde{m},j}(t) + \sum_{\widetilde{m}'=\widetilde{m}-N_Q(\tilde{m})} x_{\widetilde{m}',j}(t) + z^T_{\widetilde{m},j}(t) \tag{A.2}$$

where $res_{\widetilde{m},j}(t)$ is the total residue time of the server for the currently serving packet perceived by packet $\widetilde{m}$, $\sum\limits_{\widetilde{m}'=\widetilde{m}-N_Q(\tilde{m})} x_{\widetilde{m}',j}(t)$ is the total service time of the other $N_Q(\tilde{m})$ packets in the queue before packet $\widetilde{m}$, $z^T_{\widetilde{m},j}(t)$ is the total idle time of the server due to the fact that no subcarrier is allocated to user j perceived by packet $\widetilde{m}$ at time t, and the corresponding random variables for $res_{\widetilde{m},j}(t)$, $x_{\widetilde{m},j}$, and $z^T_{\widetilde{m},j}$ are RES_j, X_j, and Z^T_j respectively.

By the Poisson Arrival See Time Average (PASTA) property of the Poisson arrival process of an M/G/1 queue, (A.2) allows us to analyze the average waiting time of user j as $E[W_j] = E[RES_j] + N_Q E[X_j] + E\left[Z^T_j\right]$ [10], where N_Q is the average queue size.

1) Express average waiting time $E[W_j]$ in terms of average residue time $E[RES_j]$ and $E[S_j]$

Since in steady state, the availability of a subcarrier to user j could be observed from the queue, as a result

$$E[S_j] = \frac{N_Q E[X_j]}{N_Q E[X_j] + E\left[Z^T_j\right]}, \text{ and so}$$

$$E[W_j] = E[RES_j] + \lambda_j E[W_j]\frac{E[X_j]}{E[S_j]} = \frac{E[RES_j]}{1-\lambda_j E[X_j]/E[S_j]}. \tag{A.3}$$

2) Express average residue time $E[RES_j]$ in terms of moments of X_j and $E[S_j]$

The residual service time is also graphically depicted in Fig. 7. We calculate the ensemble average of residue time

[11] Each realization of X_j, $x_{\widetilde{m},j}$ is the service time of the $\widetilde{m}^{\text{th}}$ packet of user j (in terms of the number of the time slot), and is defined as the time from when it started being served to the time it is completely served.

[12] If there is at least one subcarrier allocated to user j at the m^{th} time slot, then $s_j(m) = 1$, otherwise $s_j(m) = 0$.

[13] It is supposed that the $\widetilde{m}^{\text{th}}$ packet is transmitted in the m^{th} time slot.

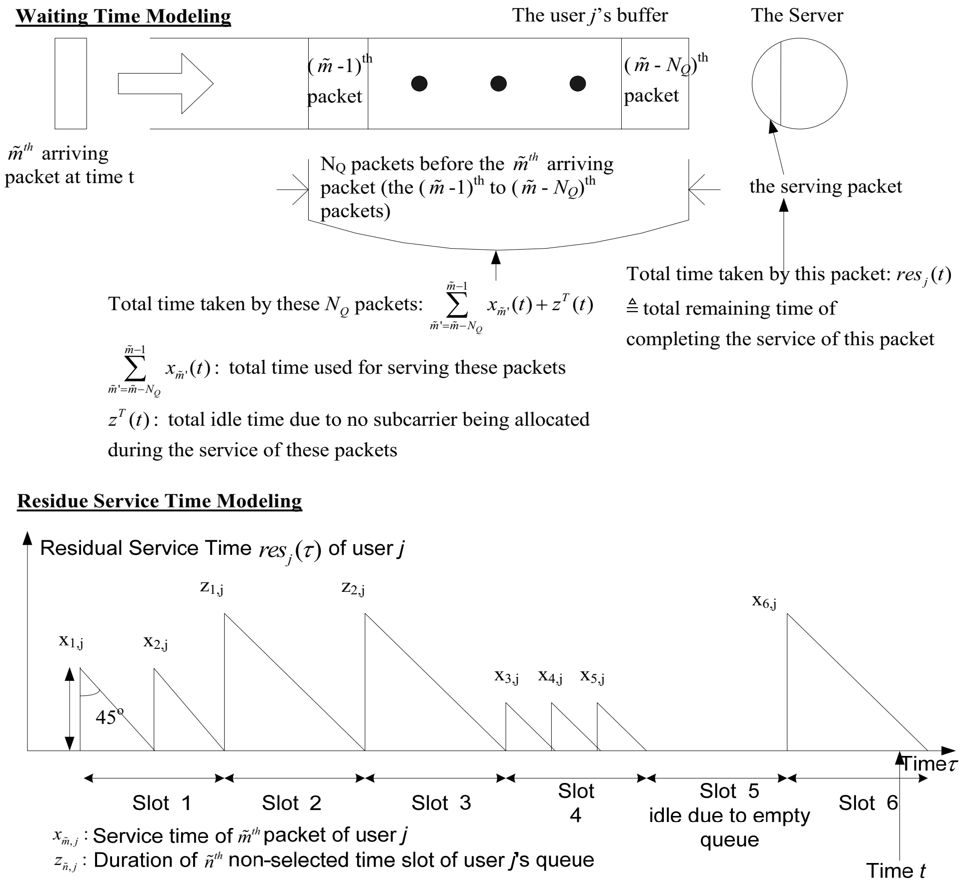

Waiting Time Modeling
The user j's buffer
The Server
$(\tilde m$ -1)th packet
$(\tilde m$ - $N_Q)^{th}$ packet
$\tilde m^{th}$ arriving packet at time t
N$_Q$ packets before the $\tilde m^{th}$ arriving packet (the $(\tilde m$ -1)th to $(\tilde m$ - $N_Q)^{th}$ packets)
the serving packet
Total time taken by these N_Q packets: $\sum_{\tilde m'=\tilde m-N_Q}^{\tilde m-1} x_{\tilde m'}(t)+z^T(t)$
Total time taken by this packet: $res_j(t)$ $\triangleq$ total remaining time of completing the service of this packet
$\sum_{\tilde m'=\tilde m-N_Q}^{\tilde m-1} x_{\tilde m'}(t)$: total time used for serving these packets
$z^T(t)$: total idle time due to no subcarrier being allocated during the service of these packets
Residue Service Time Modeling
Residual Service Time $res_j(\tau)$ of user j
x$_{1,j}$
x$_{2,j}$
z$_{1,j}$
z$_{2,j}$
x$_{3,j}$
x$_{4,j}$
x$_{5,j}$
x$_{6,j}$
45°
Timeτ
Slot 1
Slot 2
Slot 3
Slot 4
Slot 5 idle due to empty queue
Slot 6
Time t
$x_{\tilde m,j}$: Service time of $\tilde m^{th}$ packet of user j
$z_{\tilde n,j}$: Duration of $\tilde n^{th}$ non-selected time slot of user j's queue

Fig. 7. Conceptual diagram for waiting time modeling and residual service time modeling.

$E[RES_j]$ through its time average as follows:

$$\begin{aligned} E[RES_j] &= \lim_{t\to\infty} \frac{M(t)}{t}\frac{\sum_{\tilde m=1}^{M(t)}\frac{1}{2}x_{\tilde m,j}^2}{M(t)} + \frac{N(t)}{t}\frac{\sum_{\tilde n=1}^{N(t)}\frac{1}{2}z_{\tilde n,j}^2}{N(t)} \\ &= \frac{\lambda_j E\left[X_j^2\right]}{2} + \frac{\lambda_j E[X_j]}{2}\frac{\left(E\left[\boxed{S_j}\right]\right) t_s}{E[S_j]} \qquad \text{(A.4)} \end{aligned}$$

where $z_{\tilde n,j}$ is the duration of the $\widetilde{n}^{th}$ non-selected time slot, $M(t)$ is number of the packet departure up to time t, and $N(t)$ is number of the non-selected time slot up to time t.

In (A.4), we noted that $\lim_{t\to\infty}\frac{M(t)}{t}=\lambda_j$ (the rate of departure is the same as the rate of arrival in steady state) and

$$\begin{aligned} \frac{E[S_j]}{E\left[\boxed{S_j}\right]} &= \frac{M(t)E[X_j]}{N(t)(t_s)} = \frac{(M(t)/t)E[X_j]}{(N(t)/t)(t_s)} \\ \Longrightarrow \lim_{t\to\infty}\frac{N(t)}{t} &= \frac{\lambda_j E[X_j]}{t_s}\frac{E\left[\boxed{S_j}\right]}{E[S_j]} \end{aligned}$$

3) Resultant model of average waiting time $E[W_j]$ in terms of moments of X_j and $E[S_j]$

By (A.3) and (A.4), the average waiting time would be

$$E[W_j] = \frac{\lambda_j E\left[X_j^2\right] + \lambda_j E[X_j]\left(E\left[\boxed{S_j}\right]/E[S_j]\right)t_s}{2\left(1-\lambda_j E[X_j]\ /E[S_j]\right)}$$

and hence the delay constraint on the system time of each user j can be equivalently written as:

$$E[X_j] + \frac{\lambda_j E\left[X_j^2\right] + \lambda_j E[X_j]\left(E\left[\boxed{S_j}\right]/E[S_j]\right)(t_s)}{2\left(1-\lambda_j E[X_j]/E[S_j]\right)} \leq T_j. \qquad \text{(A.5)}$$

which is the result of Lemma 1. By expressing the second order moment of service time $E\left[X_j^2\right]$ in terms of average service time $E[X_j]$ through $E\left[X_j^2\right]=Var[X_j]+(E[X_j])^2$, where $Var[X_j]$ is the variance of X_j, and using the standard quadratic formula, (A.5) can be rewritten as $E[X_j]\leq\frac{-b-\sqrt{b^2-4ac}}{2a}$, where

$$\begin{aligned} a &= \frac{2\lambda_j}{E[S_j]}-\lambda_j, \\ b &= -\left(2+2\frac{\lambda_j T_j}{E[S_j]}+\lambda_j\frac{E\left[\boxed{S_j}\right]}{E[S_j]}t_s\right), \\ c &= 2T_j-\lambda_j Var[X_j] \qquad \text{(A.6)} \end{aligned}$$

It is noted that when the delay requirement of user j is $T_j\to\infty$, $\left(-b-\sqrt{b^2-4ac}\right)/2a\to 1/\lambda_j$ using L'Hospital's Rule. Hence using the result of (A.1) and (A.6), a necessary and sufficient condition for the constraint (C5) would be $E[S_jR_j]\geq F\lambda_j$ when $T_j\to\infty$ (Corollary 1). It illustrates that even though user j does not have any delay requirement, the system should provide an average scheduled data rate of at least the same as the bits arrival rate to user j's buffer to guarantee the stability of the queue. Besides, since $E\left[X_j^2\right] = Var\left[X_j\right] + \left(E\left[X_j\right]\right)^2 \geq \left(E\left[X_j\right]\right)^2$, a necessary condition for constraint (C5) would be

$$\frac{E\left[S_j\right]F}{E\left[S_jR_j\right]} + \frac{\lambda_j\left(\frac{E[S_j]F}{E[S_jR_j]}\right)^2 + \lambda_j\left(\frac{E\left[\overline{S}_j\right]F}{E[S_jR_j]}\right)(t_s)}{2\left(1-\frac{\lambda_j F}{E[S_jR_j]}\right)} \leq T_j \tag{A.7}$$

And thus, by setting $E\left[S_j\right] = 1$ and $E\left[\overline{S}_j\right] = 0$, we obtain a lower bound of the average scheduled data rate required by user j, $E\left[S_jR_j\right]$, as shown in Corollary 2. ■

References

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

[2] M. Ergen, S. Coleri, and P. Varaiya, "QoS aware adaptive resource allocation techniques for fair scheduling in OFDMA based broadband wireless access systems," *IEEE Trans. Broadcasting*, vol. 49, no. 4, pp. 362-370, Dec. 2003.

[3] J. Jang and K. B. Lee, "Transmit power adaptation for multiuser OFDM systems," *IEEE J. Select. Areas Commun.*, vol. 21, no. 2, pp. 171-178, Feb. 2003.

[4] G. Song and Y. (G.) Li, "Cross-layer optimization for OFDM wireless network–Part I: Theoretical framework," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 614-624, Mar. 2005.

[5] G. Song and Y. (G.) Li, "Cross-layer optimization for OFDM wireless network–Part II: Algorithm development," *IEEE Trans. Wireless Commun.*, vol. 4, no. 2, pp. 625-634, Mar. 2005.

[6] S. Kittipiyakul and T. Javidi, "Resource allocation in OFDMA: How load-balancing maximizes throughput when water-filling fails," UW Technical Report, UWEETR-2004-0007, 2004.

[7] E. M. Yeh and A. S. Cohen, "Information theory, queueing, and resource allocation in multi-user fading communications," in *Proc. Information Sciences Syst. Conf.*, Mar. 2004, pp. 1396-1401.

[8] E. M. Yeh, "Multiaccess and Fading in Communication Networks," Ph.D. Thesis, Department of Electrical Engineering and Computer Science, MIT, 2001.

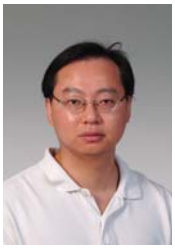
[9] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[10] D. Bertsekas and R. G. Gallager, *Data Networks*, 2nd ed. Upper Saddle River, NJ: Prentice Hall, 1992.

[11] P. Parag, S. Bhashyam, and R. Aravind, "A subcarrier allocation algorithm for OFDMA using buffer and channel state information," in *Proc. IEEE VTC-Fall*, Sep. 2005, pp. 622-625.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

David Shui Wing Hui (S'06) obtained the B. Eng. (with First Class Honor) in Information Engineering in 2004 and the M. Phil. in 2007, both from the University of Hong Kong. He is currently working toward the Ph.D. degree at the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology (HKUST). His current research interest is on the cross-layer optimization for MIMO/OFDM wireless systems with delay requirements, and applications of information theory and queueing theory to wireless communications.



Vincent K. N. Lau (M'98–SM'01) obtained the B.Eng (Distinction 1st Hons) from the University of Hong Kong (1989-1992) and a Ph.D. from Cambridge University (1995-1997). He was with HK Telecom (PCCW) as a system engineer from 1992-1995, and Bell Labs - Lucent Technologies as a member of technical staff from 1997-2003. He joined the Department of ECE, Hong Kong University of Science and Technology (HKUST), as an Associate Professor. At the same time, he is a technology advisor for HK-ASTRI, leading the Advanced Technology Team on Wireless Access Systems. His current research focus is on robust cross-layer scheduling for MIMO/OFDM wireless systems with imperfect channel state information, and communication theory with limited feedback as well as cross-layer scheduling for users with heterogeneous delay requirements.

Wong Hing Lam (S'86–M'87–SM'96) was born in Hong Kong on March 7, 1960. He received the B.Sc. degree in computer and communication engineering from the University of Essex, U.K., in 1983, the M.Sc. degree in telecommunication engineering from the Imperial College, University of London, U.K., in 1984, and the Ph.D. degree from the University of Southampton, U.K. In 1991, he joined the Department of Electrical and Electronic Engineering, University of Hong Kong. He has published more than 63 technical publications in the field of digital cellular mobile radio communications. His current research interests are pico-cellular/indoor personal communications networks (PCN), UMTS, wireless LAN and WAN, Global Positioning System (GPS) and Intelligent Transport Systems (ITS).