Cross Layer Designs 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 system with heterogeneous delay requirements. Unlike a lot of existing cross layer designs where the queueing theory and source statistics are decoupled from the information theoretical model, we shall focus on the cross layer design which takes into account of both the 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 individual user's delay constraint and total base station transmit power constraint. The cross layer scheduling algorithm dynamically allocates the radio resource based on users' channel state information (CSIT), source statistics and the delay requirements. 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. Asymptotic multi-user diversity gain is obtained analytically and 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— Orthogonal Frequency Division Multiple Access (OFDMA), power control, subcarrier allocation, heterogeneous applications, delay-sensitive Cross Layer Scheduling

I. INTRODUCTION

OFDM has been proposed as the modulation and multiple access schemes for providing high speed data transmission over next generation networks such as IEEE 802.16 Wireless Metropolitan Area Network because of its robust performance over frequency selective channel. Conventional multiuser OFDM system, e.g. OFDM-FDMA and OFDM-TDMA, only allows a single user to transmit on all of the subcarriers or a fixed subset of subcarriers [1]. However, such a fixed subcarrier allocation scheme fails to exploit the multi-user diversity in the time varying wireless channel. OFDMA with cross layer scheduling exploits this multi-user diversity, by carefully assigning multiple users to transmit simultaneously on the different subcarriers for each

OFDM symbol with optimal power and rate allocations, and as a result, the overall system throughput is increased significantly. There are quite a number of existing works on cross layer scheduling design for OFDMA systems such as [2, 3, 4, 5, 6] 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 rates requirements have been studied in [2] and [3] respectively, while the data rate maximization problem is considered in [4]. The authors in [5] and [6] 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, in 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 higher layer system performance perspective, particularly upon the provision of diverse QoS requirements in terms of delay performance. Hence, these cross layer designs were targeted for delay insensitive applications only. On the other hand, initial attempts on cross layer schedulers designs that incorporated both the source statistics and queue dynamics were reported in [7, 8, 9, 12, 13] where a simple On-Off physical layer model was assumed in [7], and the multiple access channel model with homogeneous users was studied in [8, 9] through combined information theory [10] and queueing theory [11] with the objective to minimize the average system delay. Cross layer heuristic schedulers were proposed in [12] and [13]. The authors of [12] presented a heuristic urgency based allocation policy for Multiuser MISO system with only two classes of users - delay sensitive VoIP users and delay insensitive data users. In [13], a heuristic scheduler design for maximizing the system throughput while providing fairness between users in an OFDMA system. Yet, it is not clear how good these proposed heuristic allocation policy in [12] and [13] performs compared with the optimal performance. Furthermore, all these designs, except heuristic design in [12], were targeted for systems with homogeneous users only. To our best knowledge, the optimal design for cross layer over OFDMA systems with heterogeneous delay constraints still have not yet been addressed.

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 the optimal delay-sensitive subcarrier allocation and power allocation policies to maximize the total system throughput and at the

same time, satisfying the 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 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 number of users and number of subcarriers. An iterative algorithm for finding the "multi water-levels" of heterogeneous users is also proposed. Asymptotic multiuser diversity gain with heterogeneous delay constraints is obtained from the analytical model.

This paper is organized as follows. In Section II, we 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 policy are presented in Section IV. Section V illustrates the asymptotic multiuser diversity gain for the proposed cross layer scheduler. Simulation results are studied in Section VI and a conclusion is given in Section VII.

II. SYSTEM MODEL

This section outlines the downlink OFDMA system model which is the basis of the resource allocation problem formulated in section III. The general cross-layer system model of multiuser wireless systems and the specific system architecture of a multiuser downlink OFDM scheduler are shown in Figure 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 obtains channel state information (CSI) through the uplink dedicated pilots from all mobile users[2] and collects queue state information (QSI) by observing 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

[1] Unlike [7] which simplified the physical layer into a simple ON-OFF model, we consider a more sophisticated information theoretical model to capture the performance of the OFDMA physical layer.

[2] 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 FDD system, explicit feedback of downlink CSIT from mobile users is required.

decision made by the base station transmitter is assumed to be announced to individual 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 and perfect channel coding on each subcarrier can be performed according to channel characteristic.

A. Channel Model

We consider an OFDMA system with 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 denotes the subcarrier index and j denotes the user index. The received symbol Y_{ij} at the j-th mobile user on the i-th subcarrier is given by

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

where X_{ij} is the data symbol from the base station to the j-th mobile user 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 base station to user j through subcarrier i is given by $p_{ij} = E[|X_{ij}|^2]$. 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 base station is constrained by P_{TOT}, i.e. $E[\sum_{j=1}^{K}\sum_{i=1}^{N_F} s_{ij} p_{ij}] \le P_{TOT}$, where P_{TOT} is the average total available transmit power in the base station.

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}[3] (bits/s/Hz) conveying from base station 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

---

[3] c_{ij} is called "instantaneous channel capacity", does not require to be achieved by "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 a finite delay.

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

where $I(X_{ij};Y_{ij} \mid h_{ij})$ 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 base station[4]. We also represent the transmission rate (scheduled at maximum achievable data rate) in matrix form by $R_{N_F \times K} = [r_{ij}]$ with individual matrix element equal to $r_{ij} = c_{ij}$.

C. Source Model

In this paper, we assume packets 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. Furthermore, we consider the scenario with heterogeneous mobile user applications. 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 the 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). We further assume each user has an individual buffer that is sufficiently large enough for storing packets arrived from higher layer, so that there is no buffer overflow.

D. MAC Layer Model

The system dynamics are characterized by system state $\chi = (H_{N_F \times K}, Q_K)$, which composes of channel state $H_{N_F \times K} = [|h_{ij}|^2]$ and buffer state Q_K, where $Q_K = [q_j]$ is a $K \times 1$ vector with the *j*-th component denotes the number of packets remains 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 χ as illustrated in Figure 1. At the beginning of every frame, the base station 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

[4] 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 802.11n WLAN system, packet length is of 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.

on the downlink common channels to all mobile users before the subsequent downlink packets transmissions at scheduled rates.

III. PROBLEM FORMULATION

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

Cross Layer 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:

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

$$subject\ to\ (C1): s_{ij}\in\{0,1\},\quad (C2): \sum_{j=1}^{K} s_{ij}=1,\quad (C3): p_{ij}\ge 0,\quad (C4): E\left[\frac{1}{N_F}\sum_{j=1}^{K}\sum_{i=1}^{N_F} s_{ij}p_{ij}\right]\le P_{TOT},\qquad (3)$$

$$(C5): E[\tilde{W}_j]\le T_j,\qquad \forall \chi, i, j$$

where $\tilde{W}_j$ is the system time (the duration of staying in the system) of user j's packet in system state $\chi=(CSI,QSI)$, P_{TOT} is average total power constraint, and rate allocation r_{ij} from policy $R_{N_F\times K}$ is related to power allocation from policy $P_{N_F\times K}$ by $r_{ij}=c_{ij}=\log_2(1+p_{ij}|h_{ij}|^2)$ as described in Section IIB.

In optimization problem (3)[5], 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 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 $E[\tilde{W}_j]$[6] (including average waiting time and average service time) is required to be smaller than the user j's delay requirement T_j. We assume that the arrival rates of the system are large enough so that there are

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

[6] The system time of user j's packet consists of two components: one is the waiting time, which is the duration that the packet 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 complete the encoding of this packet and start encoding another packet if there is another packet in the queue).

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 the following lemma from queueing analysis.

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

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

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, t_s is the duration of the scheduling slot . Note that S_j and $\overline{S_j}$ are indicator variables for availability and unavailability of subcarrier for user j respectively, i.e.

$\begin{cases}(s_j(m)=1,\overline{s_j}(m)=0) & \text{if there is a subcarrier allocated to user } j \text{ at time slot index } m,\\ (s_j(m)=0,\overline{s_j}(m)=1) & \text{if none of the } \mathrm{N_F} \text{ subcarriers is assigned to user } j \text{ at time slot index } m\end{cases}$. In practical OFDMA system, 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. $E[S_j]=1$ and $E[\overline{S_j}]=0$.

From Lemma 1, the constraint (C5) is ready to be transformed to an equivalent rate constraint that directly relate 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 $E[S_j R_j] \ge F\lambda_j$. This corollary shows that average scheduled data rate $E[S_j R_j]$ 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 *stability* of the queue.

Corollary 2: A necessary condition for the constraint (C5), which is called the equivalent rate constraint, is

$$E[S_j R_j] \ge \rho_j(\lambda_j, T_j, F), \text{ where } \rho_j(\lambda_j, T_j, F) = \frac{(2T_j\lambda_j+2)+\sqrt{(2T_j\lambda_j+2)^2-8T_j\lambda_j}}{4T_j}F \tag{5}$$

Proof: Proof of Lemma 1, Corollary 1 and Corollary 2 are presented in the Appendix A.

IV. SCHEDULING STRATEGIES

The optimization problem (4) is 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 compute the optimal power allocation $\{p_{ij}\}$ for selected user over 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 K^{N_F} which is not feasible for moderate N_F. In this section, 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 complexity of $N_F \times K$ only. Using Corollary 2 and equation (5), optimization problem (3) can be reformulated as follows:

$$\max_{S:\{s_{ij}\in\{0,1\},\sum_{j=1}^{K} s_{ij}=1\},P:\{p_{ij}\geq 0\}} \mathrm{E}\left[\sum_{i=1}^{N_F}\sum_{j=1}^{K} s_{ij}\log_2\left(1+p_{ij}\,|\,h_{ij}\,|^2\right)\right]$$
$$\textit{subject to}\ (C4): E\left[\frac{1}{N_F}\sum_{j=1}^{K}\sum_{i=1}^{N_F} s_{ij}\,p_{ij}\right]\leq P_{TOT},\ (C5): E\left[\sum_{i=1}^{N_F} s_{ij}\log_2\left(1+p_{ij}\,|\,h_{ij}\,|^2\right)\right]\geq \tilde{\rho}_j(\lambda_j,T_j,F),\quad \forall \chi,i,j \tag{6}$$

where $\tilde{\rho}_j(\lambda_j,T_j,F)=\rho_j(\lambda_j,T_j,F)\times(\frac{1}{t_s}/\frac{BW}{N_F})$ and BW is the total Bandwidth of the OFDMA system.

This optimization problem (6) is also 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 time sharing factor $s_{ij}=[0,1]$ and reformulate the problem using the variable $\tilde{p}_{ij}=p_{ij}s_{ij}$. The resultant reformulated problem from optimization problem (6) would be a convex maximization problem. Using the Lagrange Multiplier techniques, the following Lagrangian of the reformulated problem is obtained as follows:

$$L=\sum_{j=1}^{K}\sum_{i=1}^{N_F} s_{ij}\log_2\left(1+(\tilde{p}_{ij}/s_{ij})\,|\,h_{ij}\,|^2\right)-\mu\left(\sum_{j=1}^{K}\sum_{i=1}^{N_F}\tilde{p}_{ij}-N_F P_{ToT}\right)+\sum_{j=1}^{K}\left(\gamma_j s_{ij}\log_2\left(1+(\tilde{p}_{ij}/s_{ij})\,|\,h_{ij}\,|^2\right)-\tilde{\rho}_j\right)+\sum_{i=1}^{N_F}\phi_i\left(\sum_{j=1}^{K}s_{ij}-1\right) \tag{7}$$

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:

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

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

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

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 user *j* in problem (6). The search of the Lagrange multipliers requires a numerical procedure which would be presented in Section IV C.

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 higher power water-level $(1+\gamma_j)/\mu$ (where the value of γ_j depend 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 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, the optimal subcarrier allocation policy (8) can be implemented by a greedy algorithm with linear

complexity $N_F \times K$. The proof of Theorem 1 is shown in the Appendix B.

B. Minimal power required for provision of delay requirements guarantee

It should be noted that the user delay requirements may not be always feasible. There is a minimum average transmit power requirement (P_{min}) in order to satisfy of the delay requirements of all users. Given all the K users characteristic tuples $[\lambda_j, T_j]$, under joint subcarrier and power allocation policy presented in (8) and (9), the minimum power required to support delay constraints of all users are given by P_{min} which is calculated by solving the system of equations in (10):

$$\begin{cases} P_{\min} = E\left[\sum_{i=1}^{N_F}\sum_{j=1}^{K} s_{ij}\left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{|h_{ij}|^2}\right)^+\right] \\ E\left[\sum_{i=1}^{N_F} s_{ij}\left(\log_2\left(\frac{(1+\gamma_j)}{\mu}|h_{ij}|^2\right)\right)^+\right] = \tilde{\rho}_j(\lambda_j, T_j, F), \forall j \end{cases} \quad (10)$$

When $P_{TOT} \geq P_{\min}$, the delay constraints (C5) of problem (6) for all delay sensitive users are active; Otherwise, at least one of the delay constraints cannot be satisfied for any power and subcarrier allocation policy. Numerical examples on minimum required power are shown in Section VI.

C. Iterative Lagrange Multiplier Finding Algorithm

Let $\boldsymbol{\gamma} = \{\gamma_1, ..., \gamma_K\}$. The Lagrange multipliers are obtained by solution of the following system of equations:

$$P(\mu, \boldsymbol{\gamma}) = P_{TOT} - E\left[\sum_{i=1}^{N_F}\sum_{j=1}^{K} s_{ij}\left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{|h_{ij}|^2}\right)^+\right] = 0, \text{ and } f_j(\mu, \boldsymbol{\gamma}) = \gamma_j\left[E\left[\sum_{i=1}^{N_F} s_{ij}\left(\log_2\left(\frac{(1+\gamma_j)}{\mu}|h_{ij}|^2\right)\right)^+\right] - \tilde{\rho}_j\right] = 0, \forall j \quad (11)$$

$f_j(\mu, \boldsymbol{\gamma}) < 0$ means delay constraint is violated and $P(\mu, \boldsymbol{\gamma}) > 0$ means power P_{ToT} is not used up. The iterative algorithm to find the Lagrange multipliers μ, $(\gamma_1, ..., \gamma_K)$ is presented in a flow chart format in Figure 2, and is described as follows:

Step 1: Fixed μ, use Bisection Algorithm to find $\boldsymbol{\gamma}^*$

(a) Choose an arbitrary μ. Initialize a feasible search region of $\boldsymbol{\gamma}$, denoted as $[\underline{\gamma_{j,0}}, \overline{\gamma_{j,0}}]$ for all user j, such that $\begin{cases} f_j(\mu, \underline{\boldsymbol{\gamma}}_0) < 0 \\ f_j(\mu, \overline{\boldsymbol{\gamma}}_0) > 0 \end{cases}$ for all $j \in [1, K]$.

(b) For each user $j \in [1, K]$, we shall update $\boldsymbol{\gamma}_n, \bar{\boldsymbol{\gamma}}_n, \underline{\boldsymbol{\gamma}}_n$ based on (12) until $|f_j(\mu, \boldsymbol{\gamma}_n)|^2 < \frac{\delta}{K}$, where δ is a sufficiently small number.

$$\gamma_{j,n} = \frac{\underline{\gamma_{j,n}} + \overline{\gamma_{j,n}}}{2} \quad \text{and} \quad \underline{\gamma_{j,n+1}} = \begin{cases} \underline{\gamma_{j,n}} & \text{if } f_j(\mu, \boldsymbol{\gamma}_n) > 0 \\ \gamma_{j,n} & \text{if } f_j(\mu, \boldsymbol{\gamma}_n) < 0 \end{cases}, \quad \overline{\gamma_{j,n+1}} = \begin{cases} \gamma_{j,n} & \text{if } f_j(\mu, \boldsymbol{\gamma}_n) > 0 \\ \overline{\gamma_{j,n}} & \text{if } f_j(\mu, \boldsymbol{\gamma}_n) < 0 \end{cases} \tag{12}$$

(c) Repeat bisection algorithm in (b)[7] until we find a $\boldsymbol{\gamma}^*$ such that $\left\|\mathbf{f}(\mu, \boldsymbol{\gamma}^*)\right\|^2 < \delta$ where $\mathbf{f}(\mu, \boldsymbol{\gamma}^*) = \left\{ f_1(\mu, \boldsymbol{\gamma}^*), .., f_K(\mu, \boldsymbol{\gamma}^*) \right\}$ and $\|\mathbf{x}\|^2$ means $\sum_{j=1}^{K} |x_j|^2$, given the vector $\mathbf{x} = [x_1, x_2, ..., x_K]$.

Step 2: Redistribution of the remaining power by adjusting μ

Given $\boldsymbol{\gamma}^*(\mu)$ obtained in Step 1, determine the "remaining power" $P(\mu, \boldsymbol{\gamma}^*)$ from (12). As illustrated in Figure 2, if $P(\mu, \boldsymbol{\gamma}^*) < 0$, the problem is infeasible because only insufficient power is provided to meet all the delay requirements. If $P(\mu, \boldsymbol{\gamma}^*) = 0$, then $(\mu, \boldsymbol{\gamma}^*)$ obtained in Step 1 is the solution. If $P(\mu, \boldsymbol{\gamma}^*) > 0$, the solution $(\mu^*, \boldsymbol{\gamma}^*)$ is obtained as follow.

Given $(\mu, \boldsymbol{\gamma}^*)$ obtained in Step 1, we first initialize a feasible search region of μ^*, denoted as $[\underline{\mu_0}, \overline{\mu_0}]$, such that $\begin{cases} P(\overline{\mu_0}, \boldsymbol{\gamma}^*) > 0 \\ P(\underline{\mu_0}, \boldsymbol{\gamma}^*) < 0 \end{cases}$[8]. The search for the correct μ^* is based on the following bisection algorithm:

$$\mu_n = \frac{\underline{\mu_n} + \overline{\mu_n}}{2} \quad \text{and} \quad \underline{\mu_{n+1}} = \begin{cases} \underline{\mu_n} & \text{if } P(\mu_n, \boldsymbol{\gamma}^*) > 0 \\ \mu_n & \text{if } P(\mu_n, \boldsymbol{\gamma}^*) < 0 \end{cases}, \quad \overline{\mu_{n+1}} = \begin{cases} \mu_n & \text{if } P(\mu_n, \boldsymbol{\gamma}^*) > 0 \\ \overline{\mu_n} & \text{if } P(\mu_n, \boldsymbol{\gamma}^*) < 0 \end{cases} \tag{13}$$

For each μ_n obtained from (13), repeat Step 1 and Step 2 to update $\boldsymbol{\gamma}^*(\mu_n)$. The iteration on μ_n in (13) terminates if $\left|P(\mu_n, \boldsymbol{\gamma}^*(\mu_n))\right|^2 < \delta$. The final solution is given by $(\mu_n, \boldsymbol{\gamma}^*(\mu_n))$.

V. Asymptotic Multiuser Diversity Gain

In this section, we study the asymptotic multiuser diversity gain under heterogeneous delay constraints. We consider an OFDMA system with 2 classes of users, where delay sensitive Class 1 contains K_1 users and delay

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

[8] A smaller μ means more power consumption, which in turns means less remaining power would be resulted.

insensitive Class 2 contains K_2 users. The optimal subcarrier allocation policy in (8) is given by $j^* = \arg\max_{j\in[1,K]} \left((1+\gamma_j/\mu)\,|\,h_{ij}\,|^2\right)^{(1+\gamma_j)/\mu}$ for each subcarrier i.

Given a fixed finite equivalent rate constraint requirements for class 1 and class 2 users $\tilde{\rho}_{j\in Class_1}, \tilde{\rho}_{j\in Class_2}$ and $P_{TOT} \geq P_{\min}$, with large number of users K (= K_1+K_2), the following lemmas summarize the multiuser diversity gain and minimum power requirement by cross layer scheduler for OFDMA systems with heterogeneous users.

Lemma 2: For large number of users K_1 and K_2, the conditional multiuser diversity gain for Class 1 and Class 2 users are both $E[s_{ij}\,|\,h_{ij}\,|^2\,|s_{ij}=1, j\in Class_1] = E[s_{ij}\,|\,h_{ij}\,|^2\,|s_{ij}=1, j\in Class_2] = \Theta(\ln(K))$ [9].

Lemma 3: For large number of users K_1 and K_2, minimum required power under cross layer scheduler P_{min} and fixed allocation scheduler $P_{min,fixed}$ are respectively given by $\Theta(\frac{2^{(\tilde{\rho}_{j\in Class_1}K_1+\tilde{\rho}_{j\in Class_2}K_2)/N_F}}{\ln(K)}) \leq P_{\min} \leq \Theta(\frac{2^{(\max_j \tilde{\rho}_j)(K)/N_F}}{\ln(K)})$ and $P_{\min,\,fixed} \geq 2^{(\max_j \tilde{\rho}_j)(K)/N_F} - 1$, where $\tilde{\rho}_j(\lambda_j, T_j, F)$ is the equivalent rate constraint mentioned in (6) and (5).

Hence, the relative saving in minimum required power using cross layer scheduler compared to fixed allocation scheduler would be $P_{\min,\,fixed}/P_{\min} \geq \Theta(\ln(K))$. Figure 3 and Figure 4 illustrate the order of growth of P_{min} and conditional multiuser diversity gain for Class 1 and Class 2 users with respect to the number of users K_1 and K_2 respectively.

Proof: Proof of Lemma 2 and Lemma 3 are presented in the Appendix C.

VI. SIMULATION RESULTS

In this section, we present the simulation results using Monte Carlo simulation to illustrate the performance of the proposed cross layer scheduler for OFDM system with heterogeneous applications in terms of average total system throughput and delay performance. We also provide some comparisons of the proposed cross layer scheme with the FDMA-like schemes.

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

[9] $a_K = \Theta(b_K)$ if $\limsup_{K\to\infty} |\,a_K\,|\,/\,|\,b_K\,| < \infty$ and $\limsup_{K\to\infty} |\,b_K\,|\,/\,|\,a_K\,| < \infty$.

fading. The duration of a scheduling slot is assumed to be 2ms. We also assume all users are of the same distance from the base station, and thus they are assumed to be homogeneous in terms of path loss. The channel fading between different users and different carriers 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 $(\boldsymbol{\lambda}, \mathbf{T}) = \{(0.3, 2), (0.4, 4), (0.5, 1000), (0.6, 1000)\}$ (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

Figure 5 depicts the average total system throughput versus SNR under various delay constraints of class 2 user. It is observed that in low SNR regime (below 7.4 dB), the system throughput is lower when delay requirement of the Class 2 users are more stringent. This is because more urgent users with heavy traffic loading (higher arrival rate) will have higher water-level and thus have higher chances of seizing subcarriers. It is also observed that the minimum required power to support all delay constraints of the user would increase as the delay requirements become more stringent. In 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 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.

2) Impact of Delay constraints on the throughput gain from Multiuser Diversity

In Figure 6, 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 number of users *K* increases.

3) Throughput Comparison among various schedulers

Figures 7 and 8 illustrate the 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 the 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 Figure 7 and 8, 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 user 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. Figure 7 and 8 also illustrate that the minimum power required to support the delay constraints of all users for the DS-DSA-APA (4.5 dB for 4 users and 10.3dB for 16 users) is substantially reduced compared to 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 Figure 9, the average delay performance versus different arrival rates of delay insensitive class 4 user is depicted given P_{TOT} = 5.65dB. 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, with the only price to be paid through 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 characteristic of delay performance guarantee is important for serving bursty delay-sensitive real time heterogeneous traffic in next generation wireless networks.

VII. CONCLUSION

In this paper, we have presented a delay-sensitive cross-layer scheduler for OFDMA systems with heterogeneous delay requirements. The cross layer design problem is formulated as an optimization problem with consideration of the source statistics, queue dynamics as well as the CSIT in the OFDMA systems. The optimal power allocation and subcarrier allocation solutions are obtained based on the optimization framework. The proposed cross layer scheduler offers a nice balance of maximizing throughput and providing QoS (delay)

differentiation of the mixed heterogeneous users. We also investigated the minimum power required for satisfying all delay requirements and provided the asymptotic multiuser diversity gain under delay sensitive cross layer framework. From the simulation results, it was also shown that substantial throughput gain is achieved by jointly optimal power and subcarrier allocation policy and all users' delay constraints are 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 user. As a result, modeling the distribution of service rate of this server is highly complex and conventional Pollaczek-Khinchin formula [11] is inconvenient for calculation of average system time $E[\tilde{W}_j]$ for each user *j* in this situation.

Consider a particular user *j*'s buffer, let *m* denotes the time slot index and $\tilde{m}$ be the packet index. The random variables representing the number of packet transmitted, availability of subcarrier, total scheduled data rate (bits/time slot) and the service time[10] for user *j* are denoted as N_j, S_j, R_j and X_j respectively (randomness depending on the evolution of system state across time span), where $n_j(m)$, $s_j(m)$[11], $r_j(m)$ and service time of the $\tilde{m}^{\text{th}}$ packet $x_{\tilde{m},j} \triangleq 1/n_j(m) = F/r_j(m)$ [12] are the corresponding realization in m^{th} time slot.

By computing the ensemble average through the time average, then average service time of user *j* (in terms of number of time slot), denoted as $E[X_j]$, can be calculated by the total service time of all packets average over number of packets that are ever served by user *j*, and is mathematically written as (A.1):

$$E[X_j] = \lim_{M\to\infty} \frac{\sum_{m=1}^{M} s_j(m)}{\sum_{m=1}^{M} s_j(m) n_j(m)} = \lim_{M\to\infty} \frac{\frac{1}{M}\sum_{m=1}^{M} s_j(m)}{\frac{1}{M}\sum_{m=1}^{M} s_j(m)(t_s \times \frac{r_j(m)}{F})} = \lim_{M\to\infty} \frac{\frac{1}{M}\sum_{m=1}^{M} s_j(m)}{\frac{1}{M}\sum_{m=1}^{M} \frac{t_s}{F}\sum_{i=1}^{N_F} s_{ij}(m) r_{ij}(m) \frac{BW}{N_F}} = \frac{E[S_j]F}{E[\sum_{i=1}^{N_F} s_{ij} r_{ij} (\frac{t_s BW}{N_F})]} = \frac{E[S_j]F}{E[S_j R_j]}$$

---

[10] Each realization of X_j, $x_{\tilde{m},j}$ is the service time of the $\tilde{m}^{\text{th}}$ packet of user *j* (in terms of number of time slot), and it is defined to be the time from it is started being served to the time it is completed served.

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

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

where $s_{ij}(m), r_{ij}(m)$ (in bits/s/Hz) are subcarrier and rate allocation result for user *j* on subcarrier *i* in m^{th} time slot.

On the other hand, as shown in Figure 10, the waiting time from the perspective of an arriving packet $\tilde{m}$ is

$$w_{\tilde{m},j}(t) = res_{\tilde{m},j}(t) + \sum_{\tilde{m}'=\tilde{m}-N_Q}^{\tilde{m}-1} x_{\tilde{m}',j}(t) + z^T_{\tilde{m},j}(t) \tag{A.2}$$

where $\begin{cases} res_{\tilde{m},j}(t) \text{ is the total residue time of the server for the currently serving packet perceived by packet } \tilde{m} \\ \sum_{\tilde{m}'=\tilde{m}-N_Q}^{\tilde{m}-1} x_{\tilde{m}',j}(t) \text{ is the total service time of other } N_Q \text{ packets in the queue before packet } \tilde{m} \\ z^T_{\tilde{m},j}(t) \text{ is the total idle time of the server due to no subcarrier allocated to the user } j \text{ perceived by packet } \tilde{m} \end{cases}$ at time *t*, and the corresponding random variable for $res_{\tilde{m},j}(t)$, $x_{\tilde{m},j}(t)$, $z^T_{\tilde{m},j}(t)$ are RES_j, X_j, and Z^T_j respectively. By Poisson Arrival See Time Average (PASTA) property of Poisson arrival process of an M/G/1 queue, we could analyze the average waiting time of user *j* $E[W_j] = E[RES_j] + N_Q E[X_j] + E[Z^T_j]$ through (A.2) [11].

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 subcarrier to user *j* could be observed from the queue,

$$E[S_j] = \lim_{t\to\infty} \frac{\sum_{\tilde{m}'=\tilde{m}-N_Q}^{\tilde{m}-1} x_{\tilde{m},j}(t)}{\sum_{\tilde{m}'=\tilde{m}-N_Q}^{\tilde{m}-1} x_{\tilde{m},j}(t) + z^T_{\tilde{m},j}(t)} = \frac{N_Q E[X_j]}{N_Q E[X_j] + E[Z^T_j]}, \text{ thus } N_Q E[X_j] + E[Z^T_j] = \frac{N_Q E[X_j]}{E[S_j]} \text{ and hence}$$

$$E[W_j] = E[RES_j] + N_Q \frac{E[X_j]}{E[S_j]} = \underbrace{E[RES_j] + \lambda_j E[W_j] \frac{E[X_j]}{E[S_j]}}_{\text{(By Little's result)}} = E[RES_j] + \rho_j \frac{E[W_j]}{E[S_j]} = \frac{E[RES_j]}{1 - \rho_j / E[S_j]} \tag{A.3}$$

where $\rho_j = \lambda_j E[X_j]$ is the utilization factor.

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 Figure 10. We calculate the ensemble average of residue time $E[RES_j]$ through its time average as follows (A.4):

$$E[RES_j] = \lim_{t\to\infty} \frac{1}{t} \int_0^t RES(\tau) d\tau = \lim_{t\to\infty} \frac{M(t)}{t} \frac{\sum_{\tilde{m}=1}^{M(t)} \frac{1}{2} x^2_{\tilde{m},j}}{M(t)} + \frac{N(t)}{t} \frac{\sum_{\tilde{n}=1}^{N(t)} \frac{1}{2} z^2_{\tilde{n},j}}{N(t)} = \lambda_j \left(\frac{E[X_j^2]}{2}\right) + \lambda_j \frac{E[X_j]}{t_s} \frac{(E[\overline{S_j}])}{E[S_j]} \frac{1}{2} (t_s)^2 = \frac{\lambda_j E[X_j^2]}{2} + \frac{\lambda_j E[X_j]}{2} \frac{(E[\overline{S_j}])t_s}{E[S_j]}$$

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

(It is noted that $\lim_{t\to\infty}\frac{M(t)}{t}=\lambda_j$, as rate of departure = rate of arrival in steady state, and $\frac{E[S_j]}{E[\overline{S_j}]}=\frac{M(t)E[X_j]}{N(t)(t_s)}=\frac{(M(t)/t)E[X_j]}{(N(t)/t)(t_s)}\Rightarrow\frac{N(t)}{t}=\frac{M(t)}{t}\frac{E[X_j]}{t_s}\frac{E[\overline{S_j}]}{E[S_j]}\Rightarrow\lim_{t\to\infty}\frac{N(t)}{t}=\frac{\lambda_j E[X_j]}{t_s}\frac{E[\overline{S_j}]}{E[S_j]}$)

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), average waiting time would be $E[W_j]=\frac{\lambda_j E[X_j^2]+\lambda_j E[X_j](E[\overline{S_j}]/E[S_j])t_s}{2(1-\rho_j/E[S_j])}$ and hence the delay constraint on system time of each user *j*, given by $E[X_j]+E[W_j]\le T_j$, can be equivalently written as:

$$E[X_j]+\frac{\lambda_j E[X_j^2]+\rho_j(E[\overline{S_j}]/E[S_j])(t_s)}{2\left(1-\rho_j/E[S_j]\right)}\le T_j\ . \tag{A.5}$$

which is the result of Lemma 1. By expressing the second order moment of service time $E[X_j^2]$ in terms of average service time $E[X_j]$ through $E[X_j^2]=Var[X_j]+(E[X_j])^2$, where $Var[X_j]$ is variance of X_j, and using standard quadratic formula, (A.5) can be rewritten as:

$$E[X_j]\le\frac{-b-\sqrt{b^2-4ac}}{2a}\ \text{where}\ a=\frac{2\lambda_j}{E[S_j]}-\lambda_j,\ b=-\left(2+2\frac{\lambda_j T_j}{E[S_j]}+\lambda_j\frac{E[\overline{S_j}]}{E[S_j]}t_s\right),\ c=2T_j-\lambda_j Var[X_j] \tag{A.6}$$

It is noted that when delay requirement of user *j* is $T_j\to\infty$, $(-b-\sqrt{b^2-4ac})/2a\to 1/\lambda_j$ using L'Hospital's Rule. Hence using the result of (A.1) and (A.6), a necessary and sufficient condition for the constraint (C6) when $T_j\to\infty$ would be $E[S_jR_j]\ge F\lambda_j$ (Corollary 1). It illustrates that even 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 $E[X_j^2]=Var[X_j]+(E[X_j])^2\ge(E[X_j])^2$, a necessary condition for the constraint (C5) would be

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

By setting $E[S_j]=1,\ E[\overline{S_j}]=0$, the average scheduled data rate required by user *j*, $E[S_jR_j]$ is lower bounded by: $E[S_jR_j]\ge[(2T_j\lambda_j+2)+\sqrt{(2T_j\lambda_j+2)^2-8T_j\lambda_j}](F/4T_j)$ (Corollary 2).

B. *Proof of Theorem 1*

To avoid complicated combinatorial search on s_{ij}, we relax the constraint (C1) to allow s_{ij} to be real number (between 0,1)[13][2]. Let $\tilde{p}_{ij} = p_{ij}s_{ij}$, the original optimization problem in (6) can be transformed into a convex maximization problem over a convex set and obtain the Lagrangian L [14] presented in (7).

After differentiating L with respect to $\tilde{p}_{ij}$, s_{ij}, respectively, we obtain the optimal solution, $\tilde{p}_{ij}^*$, s_{ij}^*.

Specifically, if $s_{ij}^* \neq 0$ [14], we have $\left.\frac{\delta L}{\delta \tilde{p}_{ij}}\right|_{(\tilde{p}_{ij},s_{ij})=(\tilde{p}_{ij}^*,s_{ij}^*)} = (1+\gamma_j)s_{ij}^* \frac{|h_{ij}|^2 / s_{ij}^*}{1+\tilde{p}_{ij}^*|h_{ij}|^2 / s_{ij}^*} - \mu \begin{cases} <0, \text{ if } \tilde{p}_{ij}^* = 0 \\ =0, \text{ if } \tilde{p}_{ij}^* > 0 \end{cases}$ (B.1)

Thus the optimal power allocation is given by $\tilde{p}_{ij}^* = s_{ij}^* \left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{|h_{ij}|^2} \right)^+$ (B.2)

and similarly, $\left.\frac{\delta L}{\delta s_{ij}}\right|_{(p_{ij},s_{ij})=(p_{ij}^*,s_{ij}^*)} = (1+\gamma_j)\left(\log_2(1+(\tilde{p}_{ij}^*/s_{ij}^*)|h_{ij}|^2) - \frac{(\tilde{p}_{ij}^*/s_{ij}^*)|h_{ij}|^2}{1+(\tilde{p}_{ij}^*/s_{ij}^*)|h_{ij}|^2} \right) - \phi_i \begin{cases} =0, \text{ if } 0<s_{ij}^*<1 \\ >0, \text{ if } s_{ij}^* = 1 \end{cases}$. (B.3)

It follows that $s_{ij}^* = \begin{cases} 1, \text{ if } \phi_i < H_{ij}(\mu,\gamma_j) \\ 0, \text{ if } \phi_i > H_{ij}(\mu,\gamma_j) \end{cases}, \forall i$, where $H_{ij}(\mu,\gamma_j) = (1+\gamma_j)\left(\log_2\left(\frac{(1+\gamma_j)}{\mu}|h_{ij}|^2 \right) \right)^+ - \mu\left(\frac{(1+\gamma_j)}{\mu} - \frac{1}{|h_{ij}|^2} \right)^+$ (B.4)

With the constraint (C2) in (3) and (B.4), for each subcarrier i, we know that if $H_{ij}(\mu,\gamma_j)$ are different for all j, only user j with the largest $H_{ij}(\mu,\gamma_j)$ can use that subcarrier i, i.e. $s_{ij^*} = 1, s_{ij} = 0$ for all $j \neq j^*$, where $j^* = \arg\max_j H_{ij}(\mu,\gamma_j)$. If $H_{ij}(\mu,\gamma_j)$ are maximum for more than 1 user, time sharing among them is needed. However, h_{ij} are i.i.d. for different user j, thus the chance for $H_{ij}(\mu,\gamma_j)$ to be the same for different users happens only with probability 0. Hence, the search for optimal subcarrier allocation is given by (B.4).

C. *Proof of Lemma 2 and Lemma 3*

Define the best user within Class 1 and Class 2 to be $j(1) = \arg\max_{j \in Class_1} \left(c_j |h_{ij}|^2 \right)^{c_j}$ and $j(2) = \arg\max_{j \in Class_2} \left(c_j |h_{ij}|^2 \right)^{c_j}$ respectively, where $c_j = (1+\gamma_j)/\mu$, noted that $c_j = c(1), \forall j \in Class_1, c_{j'} = c(2), \forall j' \in Class_2$ and $c(1) > c(2)$. The pdf of $|h_{ij(1)}|^2$ and $|h_{ij(2)}|^2$ are $p(|h_{ij(1)}|^2 = \gamma) = K_1(1-e^{-\gamma})^{K_1-1}e^{-\gamma}$ and $p(|h_{ij(2)}|^2 = \gamma) = K_2(1-e^{-\gamma})^{K_2-1}e^{-\gamma}$.

---

[13] A fractional value of s_{ij} refers to time sharing of the subcarrier i.

[14] If $s_{ij}^* = 0$, then $\tilde{p}_{ij}^* = 0$, we have $\tilde{p}_{ij}\frac{dL}{d\tilde{p}_{ij}} + s_{ij}\frac{dL}{ds_{ij}} \leq 0$ for all $s_{ij} \in (0,1]$ and $\tilde{p}_{ij} > 0$.

The pdf of $s_{ij(1)} \mid h_{ij(1)} \mid^2$ can be obtained as

$$p(s_{ij(1)} \mid h_{ij(1)} \mid^2 = \gamma) = \begin{cases} p(\mid h_{ij(1)} \mid^2 = \gamma) \Pr(s_{ij(1)}(\mid h_{ij(1)} \mid^2) = 1 \Big\| h_{ij(1)} \mid^2 = \gamma), & \gamma > 0 \\ p(\mid h_{ij(1)} \mid^2 = \gamma) \Pr(s_{ij(1)}(\mid h_{ij(1)} \mid^2) = 1 \Big\| h_{ij(1)} \mid^2 = \gamma) + \Pr(s_{ij(1)}(\mid h_{ij(1)} \mid^2) = 0)\delta(\gamma), & \gamma = 0 \end{cases}$$

$$= P(\mid h_{ij(1)} \mid^2 = \gamma) \Pr((c(1) \mid h_{ij(1)} \mid^2)^{c(1)} > (c(2) \mid h_{ij(2)} \mid^2)^{c(2)} \Big\| h_{ij(1)} \mid^2 = \gamma) + \Pr(s_{ij(1)}(\mid h_{ij(1)} \mid^2) = 0)\delta(\gamma) \quad \text{(C.1)}$$

$$= K_1 e^{-\gamma}(1-e^{-\gamma})^{K_1-1}(1-e^{-(c(1)\gamma)^{c(1)/c(2)}/c(2)})^{K_2} + \Pr(s_{ij(1)}(\mid h_{ij(1)} \mid^2) = 0)\delta(\gamma), \text{ where } \delta(\gamma) \text{ is "delta function"}$$

Moreover, as K_1 and K_2 is large, it can be shown that $c(1)/c(2) \to 1$, thus the conditional diversity gain for Class 1 and Class 2 users (i.e. the average SNR given the specified class is selected) are given by (C.2):

$$E[s_{ij(1)} \mid h_{ij(1)} \mid^2 \mid s_{ij(1)} = 1] = \frac{E[s_{ij(1)} \mid h_{ij(1)} \mid^2]}{\Pr(s_{ij(1)} = 1)} = \frac{\int_0^\infty K_1 \gamma e^{-\gamma}(1-e^{-\gamma})^{K_1-1}(1-e^{-(c(1)\gamma)^{c(1)/c(2)}/c(2)})^{K_2} d\gamma}{\int_0^\infty K_1 e^{-\gamma}(1-e^{-\gamma})^{K_1-1}(1-e^{-(c(1)\gamma)^{c(1)/c(2)}/c(2)})^{K_2} d\gamma} = \Theta(\ln(K_1+K_2)), \text{ as } K_1 \to \infty$$

$$E[s_{ij(2)} \mid h_{ij(2)} \mid^2 \mid s_{ij(2)} = 1] = \frac{E[s_{ij(2)} \mid h_{ij(2)} \mid^2]}{\Pr(s_{ij(2)} = 1)} = \frac{\int_0^\infty K_2 \gamma e^{-\gamma}(1-e^{-\gamma})^{K_2-1}(1-e^{-(c(2)\gamma)^{c(2)/c(1)}/c(1)})^{K_1} d\gamma}{\int_0^\infty K_2 e^{-\gamma}(1-e^{-\gamma})^{K_2-1}(1-e^{-(c(2)\gamma)^{c(2)/c(1)}/c(1)})^{K_1} d\gamma} = \Theta(\ln(K_1+K_2)), \text{ as } K_2 \to \infty$$

The above integral with c(1) = c(2) = 1 could be found in [5] (with the consideration of only one class of user).

In order to satisfy delay constraints of Class 1 and Class 2 users, P_{min} required is calculated based on (C.3):

$$\begin{cases} K_1 \tilde{\rho}_{j(1)} = E[\sum_{i=1}^{N_F} \sum_{j \in Class_1} s_{ij} \log(c(1) \mid h_{ij} \mid^2)] = E[\sum_{i=1}^{N_F} s_{ij(1)} \log(c(1) \mid h_{ij(1)} \mid^2)] \le N_F \Pr(s_{ij(1)} = 1) \log(c(1) E[s_{ij(1)} \mid h_{ij(1)} \mid^2 \mid s_{ij(1)} = 1]) \\ K_2 \tilde{\rho}_{j(2)} = E[\sum_{i=1}^{N_F} \sum_{j \in Class_2} s_{ij} \log(c(2) \mid h_{ij} \mid^2)] = E[\sum_{i=1}^{N_F} s_{ij(2)} \log(c(2) \mid h_{ij(2)} \mid^2)] \le N_F \Pr(s_{ij(2)} = 1) \log(c(2) E[s_{ij(2)} \mid h_{ij(2)} \mid^2 \mid s_{ij(2)} = 1]) \\ P_{\min} = E[\frac{1}{N_F} \sum_{i=1}^{N_F} \sum_{j=1}^{K} s_{ij} p_{ij}] = \Pr[s_{ij(1)} = 1] c(1) + \Pr[s_{ij(2)} = 1] c(2), \text{ since selection process is independently identical for all } i \end{cases}$$

The inequality signs in above equations are due to Jensen's inequality which would asymptotically becomes equality when $K_1 \to \infty$, $K_2 \to \infty$, i.e. $c(1)$ and $c(2)$ are large.

Thus $P_{\min} \ge \Pr[s_{ij(1)} = 1] \frac{2^{K_1 \tilde{\rho}_{j(1)}/(N_F \Pr(s_{ij(1)}=1))}}{E[s_{ij(1)} \mid h_{ij(1)} \mid^2 \mid s_{ij(1)} = 1]} + \Pr[s_{ij(2)} = 1] \frac{2^{K_2 \tilde{\rho}_{j(2)}/(N_F \Pr(s_{ij(2)}=1))}}{E[s_{ij(2)} \mid h_{ij(2)} \mid^2 \mid s_{ij(2)} = 1]}$ and it can be further shown that $\frac{2^{(\tilde{\rho}_{j(1)}K_1 + \tilde{\rho}_{j(2)}K_2)/N_F}}{\Theta(\ln(K_1+K_2))} (\frac{\max\{\tilde{\rho}_{j(1)}K_1, \tilde{\rho}_{j(2)}K_2\}}{\tilde{\rho}_{j(1)}K_1 + \tilde{\rho}_{j(2)}K_2}) \le P_{\min} \le \frac{2^{\tilde{\rho}_{j(1)}(K_1+K_2)/N_F}}{\Theta(\ln(K_1+K_2))} (\frac{K_1}{K_1+K_2}) + \frac{2^{\tilde{\rho}_{j(2)}(K_1+K_2)/N_F}}{\Theta(\ln(K_1+K_2))} (\frac{K_2}{K_1+K_2}))$.

Without cross layer scheduler, $\max_j \tilde{\rho}_j \le (\frac{N_F}{K_1+K_2}) \log(1 + P_{\min, fixed} E[\mid h_{ij} \mid^2])$. Thus $P_{\min, fixed} \ge 2^{(\max_j \tilde{\rho}_j)(K_1+K_2)/N_F} - 1$.

REFERENCES

[1] Rohling, H., Gruneid, R, "Performance comparison of different multiple access schemes for the downlink of an OFDM communication system", in *Proc. IEEE. Vehicular Technology Conf. (VTC.)*, pp. 1365 – 1369, 1997.

[2] C. Y. Wong, R. S. Cheng, K. B. Letaief, and R. D. Murch, "Multiuser OFDM with Adaptive Subcarrier, Bit, and Power Allocation", *IEEE J. Sel. Areas Commun.*, vol. 17, no. 10, Oct. 1999.

[3] 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. On Broadcasting*, vol. 49, no. 4, Dec 2003.

[4] Jiho Jang and Kwang Bok Lee, "Transmit Power Adaptation for Multiuser OFDM Systems," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 2, Feb. 2003.

[5] 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.

[6] 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.

[7] S. Kittipiyakul and T. Javidi, "Resource Allocation in OFDMA: How Load-Balancing Maximizes Throughput When Water-Filling Fails", *UW Technical Report*, UWEETR-2004-0007.

[8] E. M. Yeh and A. S. Cohen, "Information Theory, Queueing, and Resource Allocation in Multi-user Fading Communications," *Proc. of 2004 Conference on Information Sciences and Systems*, Mar. 2004.

[9] E. M. Yeh, *Multiaccess and Fading in Communication Networks*, PhD Thesis, Department of Electrical Engineering and Computer Science, MIT, 2001.

[10] T. Cover and J. Thomas, *Elements of Information Theory*, Wiley, 1991.

[11] D. Bertsekas and R.G. Gallager, *Data Networks*, 2nd Ed., Prentice-Hall, 1992.

[12] V. K. N. Lau, M. L. Jiang, S. Liew and O. C. Yue, "Performance Analysis of Downlink Multi Antenna Scheduling for Voice and Data Applications," in *Proc. 42nd Allerton Conf. Commun., Control and Comp.*, Sept. 2004.

[13] Parimal Parag, Srikrishna Bhashyam and R. Aravind, "A Subcarrier Allocation Algorithm for OFDMA using Buffer and Channel State Information," in *Proc. IEEE VTC.*, pp. 622-625, September, 2005.

[14] S. Boyd and L. Vandenberghe, *Convex optimization*, 2004.

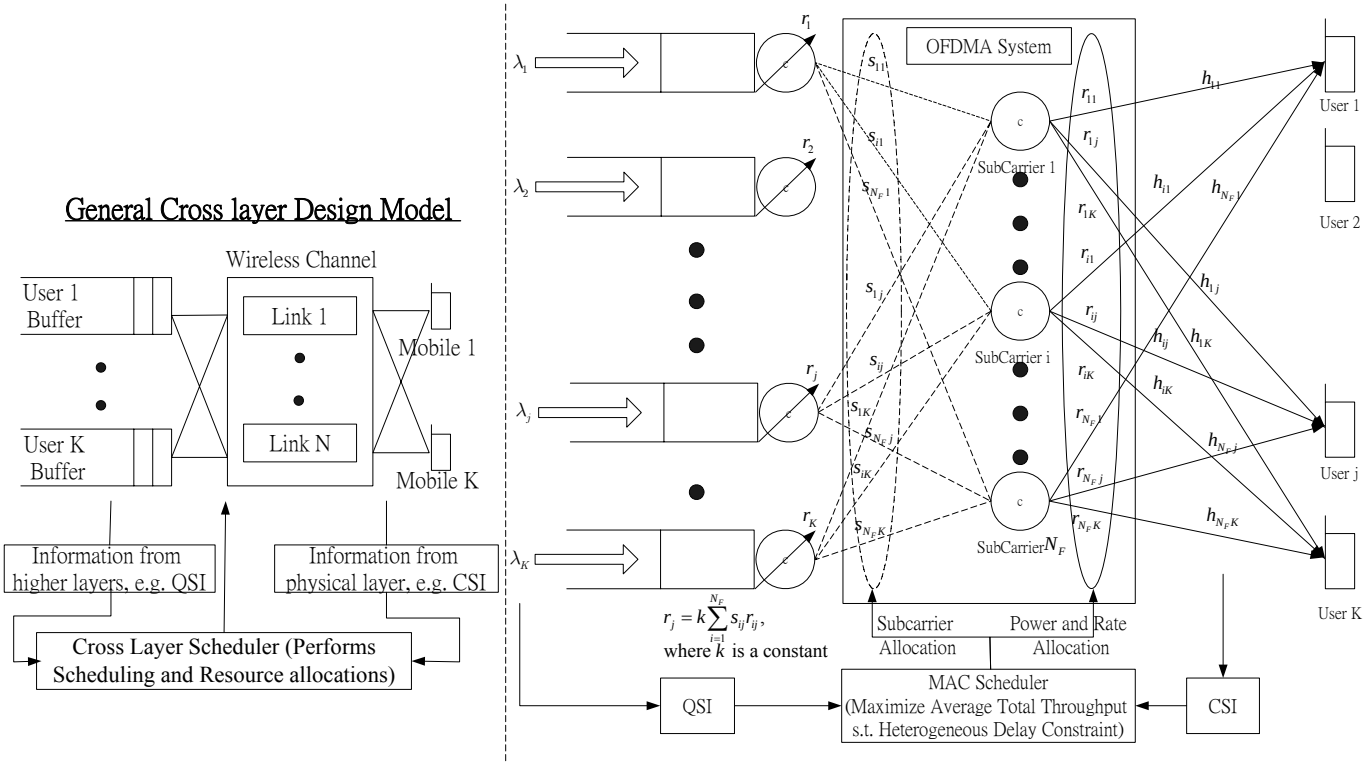


Figure 1. General Cross-layer System model (Left) and Cross Layer Scheduling model under Conceptual Channel Model for OFDMA system with heterogeneous application users (Right)

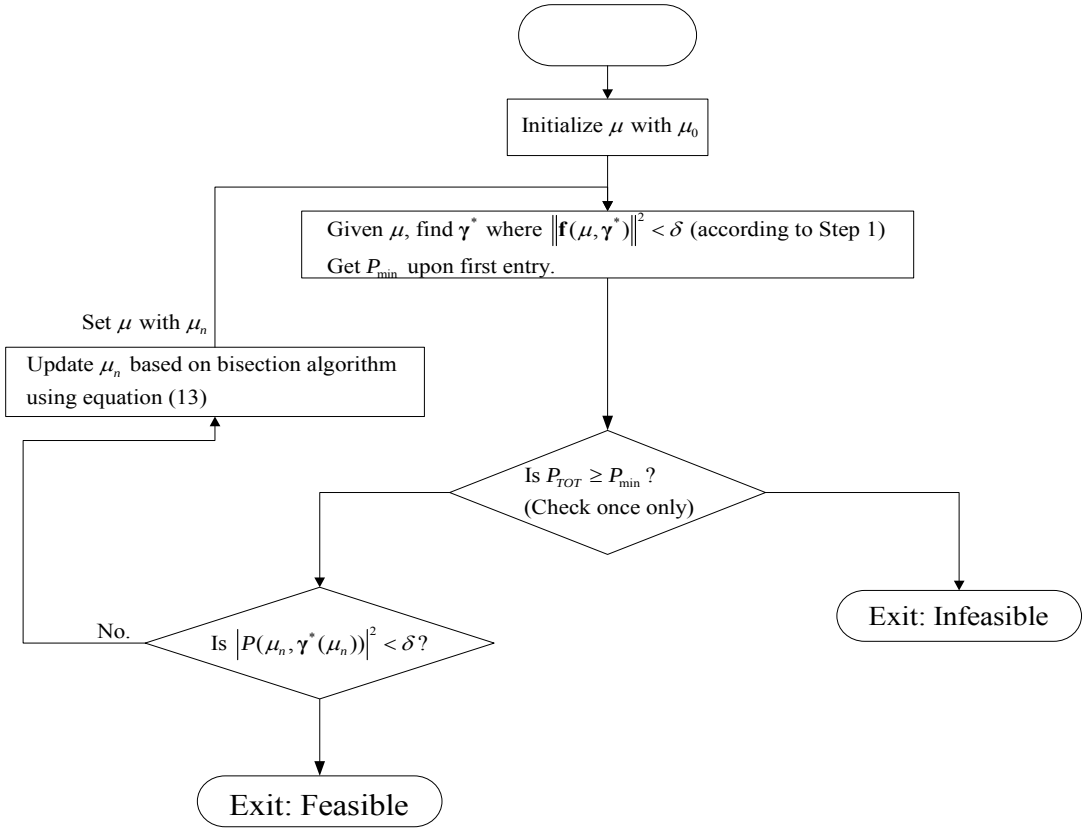


Figure 2. Flow chart of Lagrange Multiplier Finding Algorithm for jointly optimal APA and DSA

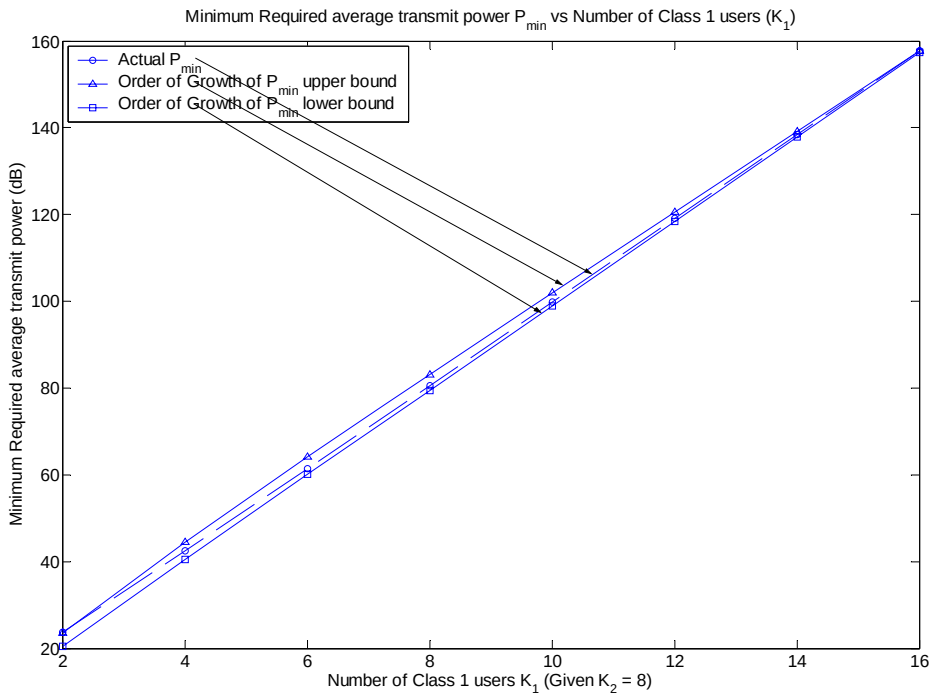


Figure 3. Minimum transmit power vs Number of Class 1 users K_1 (Delay sensitive Class 1 users and delay insensitive Class 2 users have arrival rate and delay requirement of $(\lambda_1, T_1) = (0.8, 2), (\lambda_2, T_2) = (0.1, 1000)$ (packets per time slot, time slots) respectively; given number of subcarrier N_F = 64 and System Bandwidth BW = 20kHz). We can observe that the order of growth of the analytical bound matches closely with the simulation results.

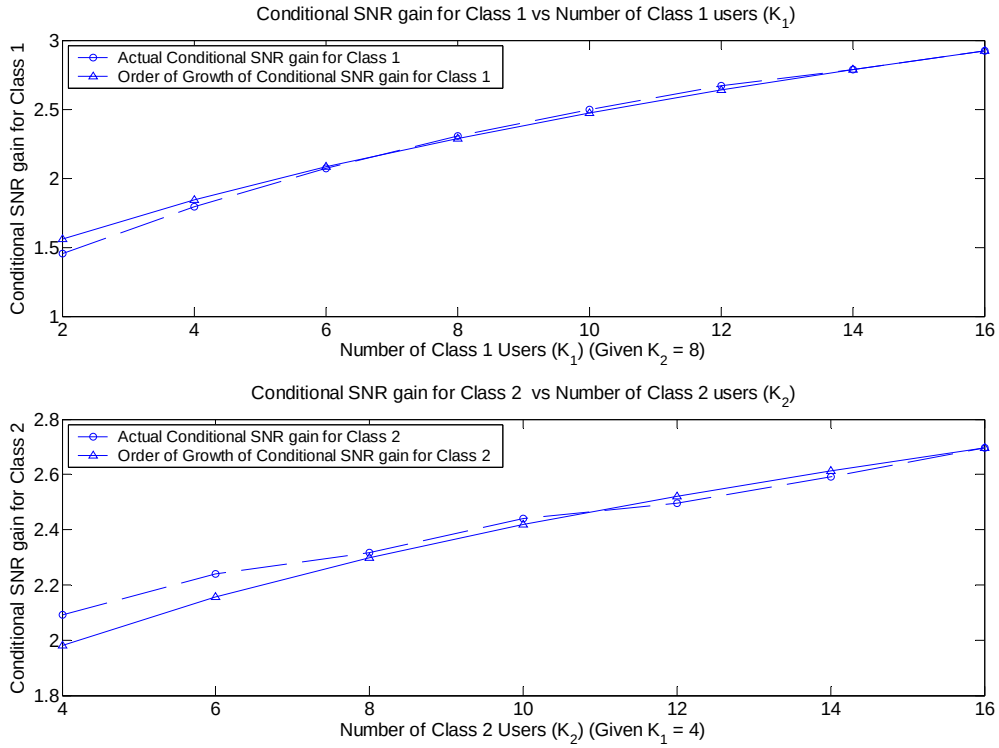


Figure 4. Conditional SNR gain vs Number of users (Arrival rate, delay requirement, number of subcarrier N_F and System Bandwidth BW are the same as Figure 3). It shows that the simulation results match closely with the predicted order of growth of Conditional SNR gain.

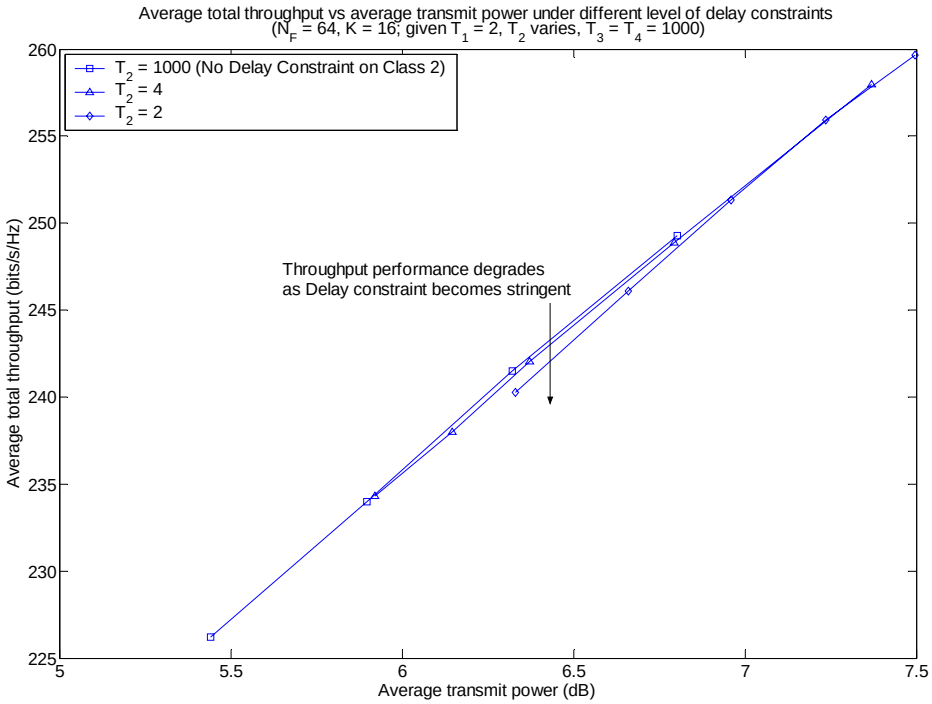


Figure 5. Average total throughput vs average transmit power under different delay constraint T_2 of class 2 users (T_2 = 2, 4, 1000 time slots) (The number of users of each class is (K_1, K_2, K_3, K_4) = (4, 4, 4, 4) respectively)

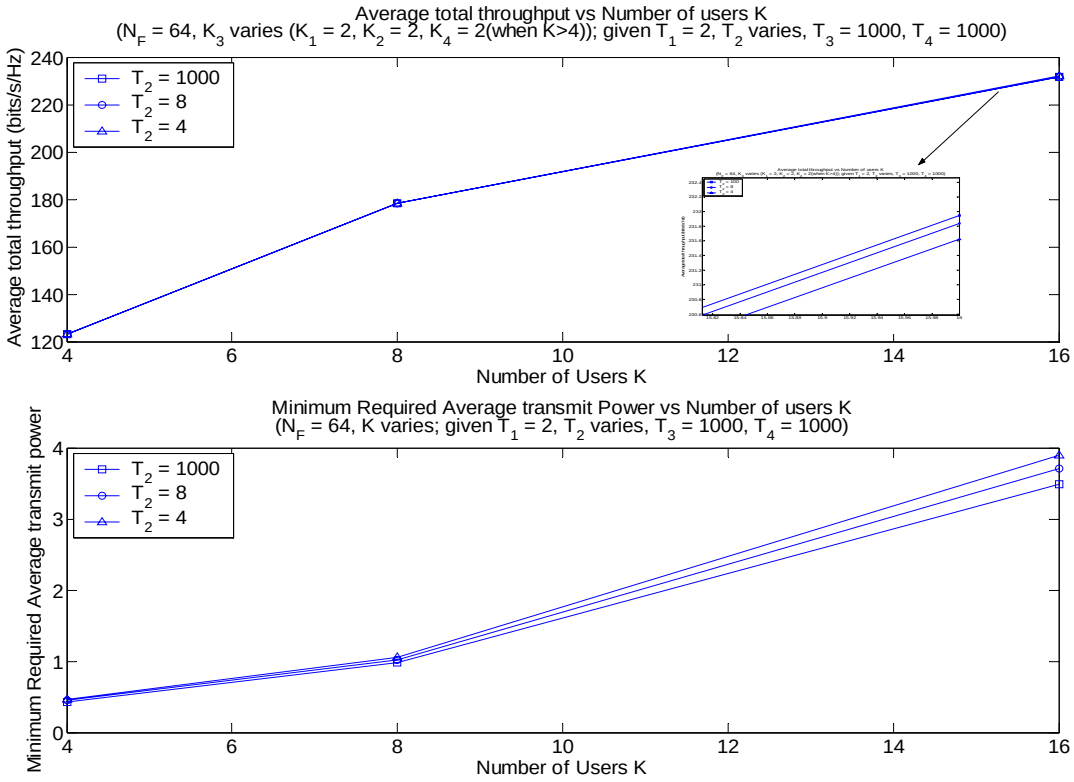


Figure 6 (Upper) Average total system throughput vs different number of users K under different delay constraint T_2 of class 2 users (T_2 = 4, 8, 1000 time slots) (For K = 16, (K_1, K_2, K_3, K_4) = (2, 2, 10, 2); for K = 8, (K_1, K_2, K_3, K_4) = (2, 2, 2, 2); for K = 4, (K_1, K_2, K_3, K_4) = (2, 2, 0, 0)); (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, (K_1, K_2, K_3, K_4) = (4, 4, 4, 4); for K = 8, (K_1, K_2, K_3, K_4) = (2, 2, 2, 2); for K = 4, (K_1, K_2, K_3, K_4) = (1, 1, 1, 1))

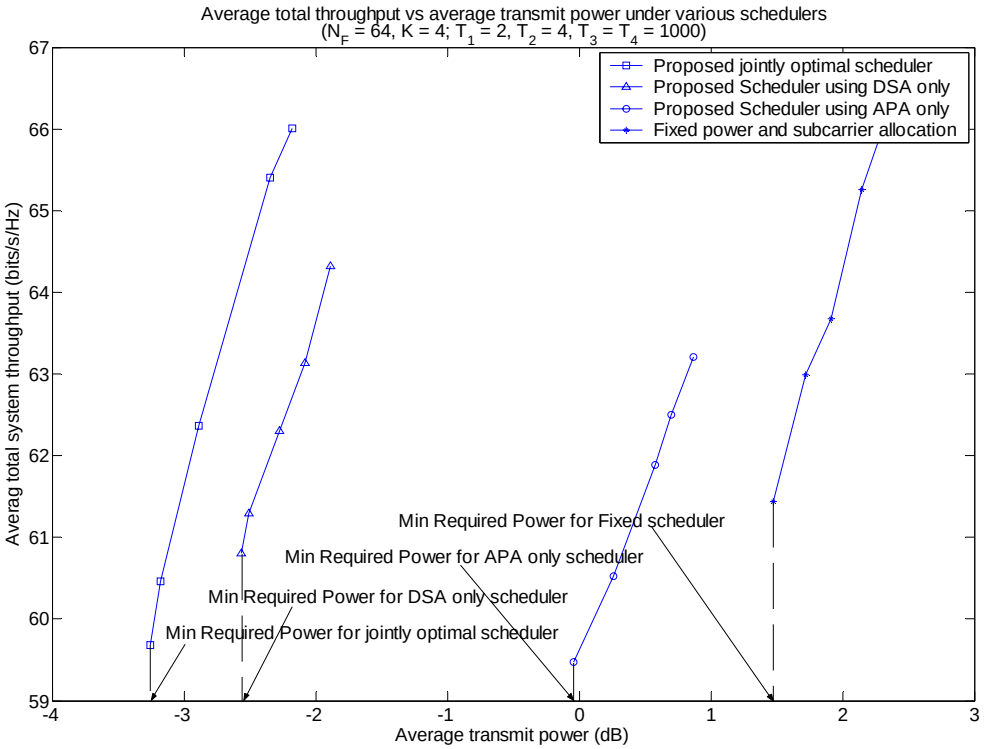


Figure 7. Average total system throughput vs average transmit power under different schedulers when $K = 4$. (The number of users of each class is $(K_1, K_2, K_3, K_4) = (1, 1, 1, 1)$ respectively)

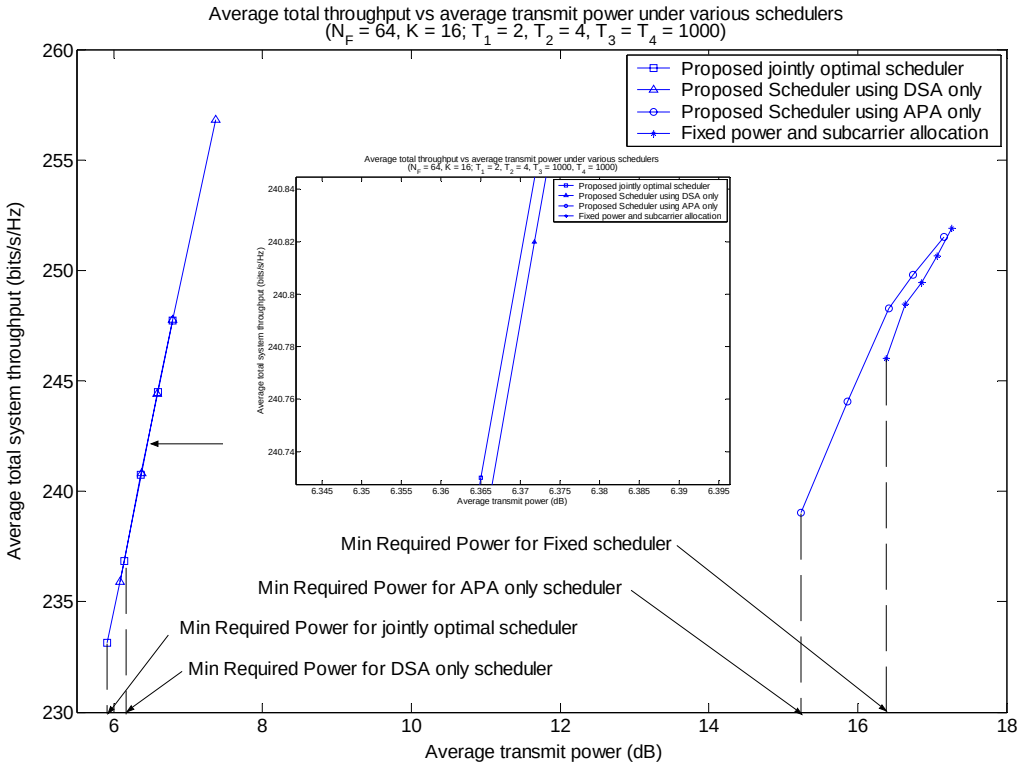


Figure 8. Average total system throughput vs average transmit power under different schedulers when $K = 16$. (The number of users of each class is $(K_1, K_2, K_3, K_4) = (4, 4, 4, 4)$ respectively)

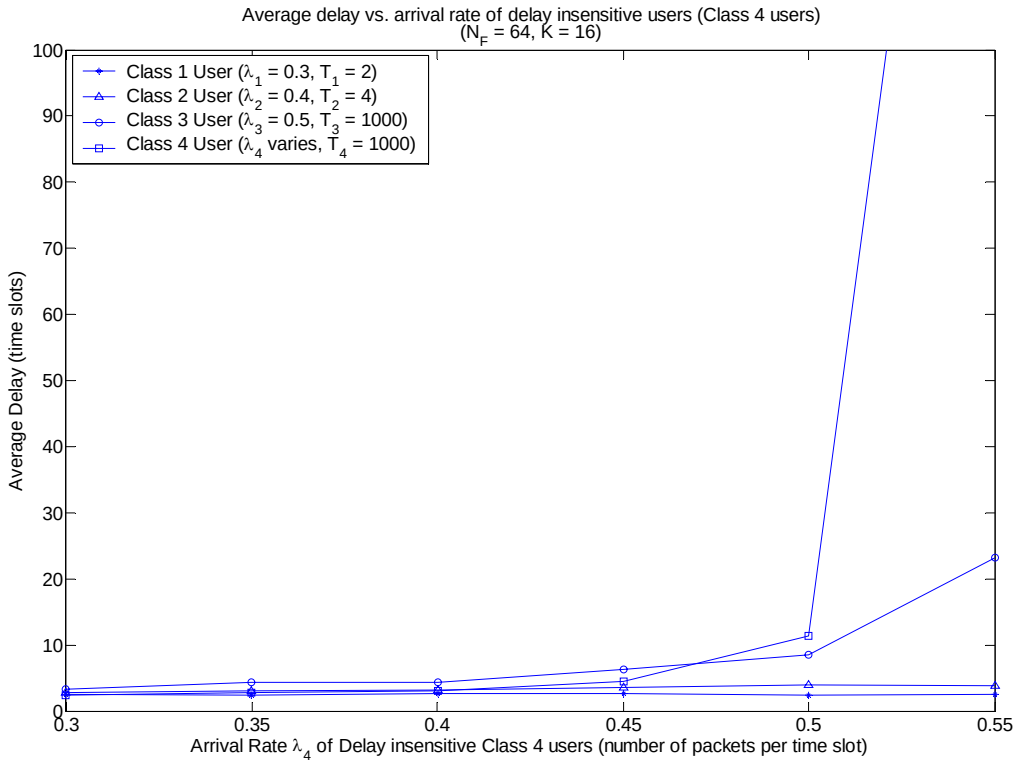
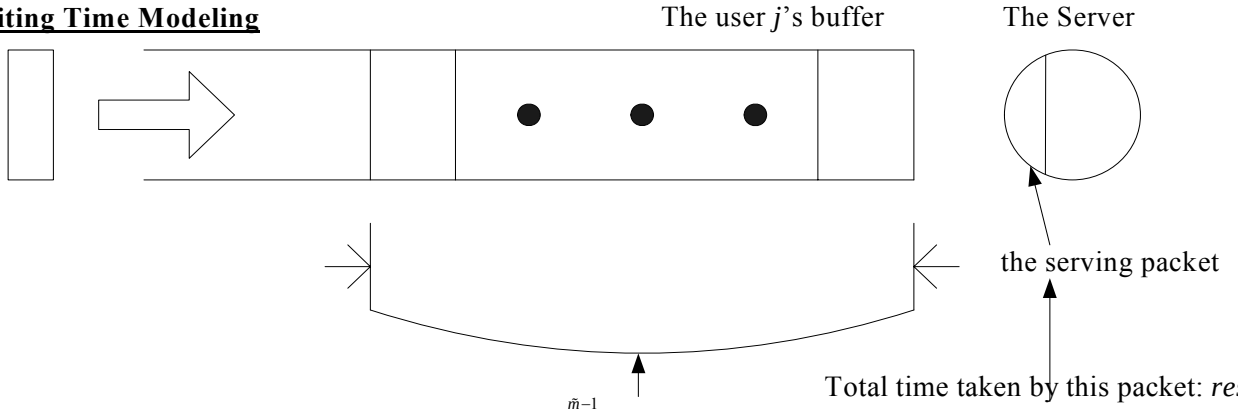


Figure 9. Average delay vs arrival rate of delay insensitive user (Class 4 users) ((K_1, K_2, K_3, K_4) = (4, 4, 4, 4))

Waiting Time Modeling



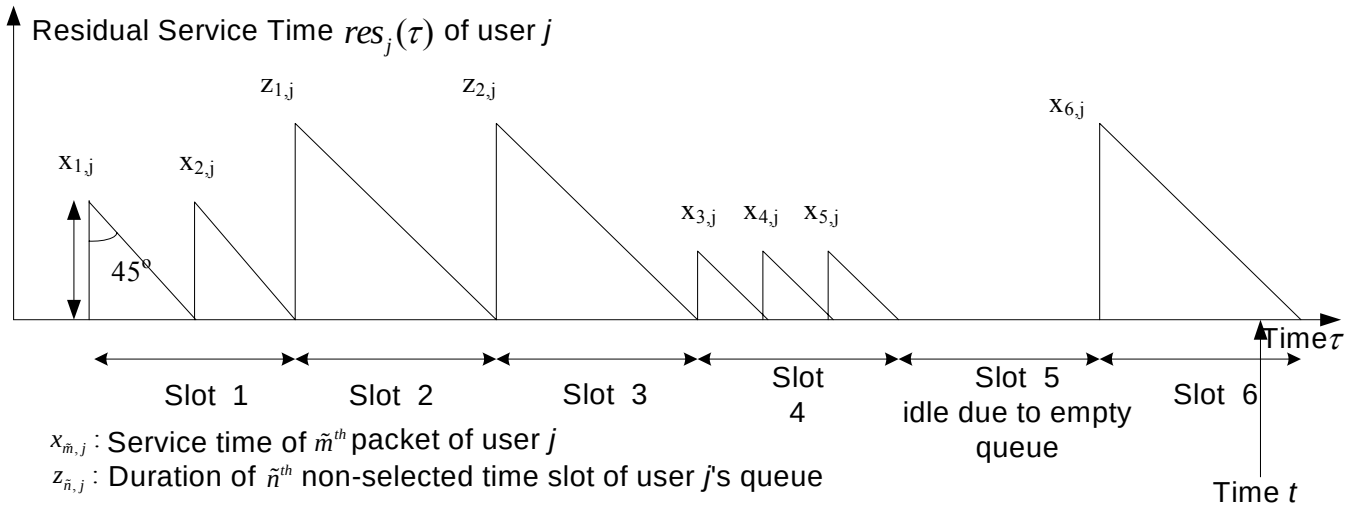
Total time taken by these N_Q packets: $\sum_{\tilde{m}'=\tilde{m}-N_Q}^{\tilde{m}-1} x_{\tilde{m}'}(t) + z^T(t)$

Total time taken by this packet: $res_j(t)$ ≜ 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



$x_{\tilde{m},j}$: 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

Figure 10. Conceptual diagram for waiting time modeling and residual service time modeling