Abstract—In this paper we propose a two step solution to the joint flow control and physical resource allocation problem for multi-service OFDMA system. We develop a flow control algorithm for the data generated at the remote server stations. This data corresponds to the service demanded by a particular user in a wireless cell. Our flow control algorithm derives an output data rate for each user while considering the variable nature of the channel capacity and pass these rates to the Base Station (BS). At the BS, physical layer resource allocation algorithm allocates the subcarrier and power to the users according to the proposed data rates. Simulation results show good performance and increased system throughput.

I. INTRODUCTION

The ever increasing demand for higher data rates and user satisfaction are pushing the efforts toward cross-layer design in wireless networks because interaction and exchange of parameters among different layers can result in tremendous performance gains. In this paper, we consider the problem of cross layer design in Orthogonal Frequency Division Multiple Access (OFDMA) system. OFDMA inherits the immunity to Inter Symbol Interference (ISI) in frequency selective fading channels. It also makes possible the dynamic allocation of subcarriers because of its remarkable ability to multiplex several users on different subcarriers without interference [1]. Moreover, efficient power allocation becomes possible by waterfilling over the inverse of the channel gains [2]. Dynamic resource allocation efficiently utilizes multiuser diversity and frequency variations in the wireless channel [3]-[4].

In this paper we consider the problem of joint flow control at the transport layer and resource allocation at the physical layer for multi-service OFDMA system. This leads to cross layer design at the transport and physical layers. In most of the previous works on cross layer design, scheduling at the MAC layer and resource allocation at the physical layer are considered. In such a problem setting each user is assumed to maintain a separate queue at the BS which receives data from higher layers. Several schedulers have been proposed in the literature such as [5]-[7]. However, relatively few authors have considered the cross layer design at the transport and physical layers [8]-[9]. In [8], the author considers wireless multihop networks and regulate the source rates at the Transport layer and power control at the physical layer. Although distributed algorithms are developed for these two layers, they cannot be implemented in OFDMA system because of the constraints used in their problem development and multi-carrier nature of the OFDMA systems. In [9], the authors consider different time scales for transport and physical layers algorithms. For each iteration of the packet level algorithm, their physical layer algorithm takes several iterations to converge. However, this assumption is not always accurate because the packets size at the transport layer is not very high e.g. in Transmission Control Protocol (TCP), packet size of a TCP segment can have a maximum value of 1500 bytes which can be treated in one time slot at the physical layer using physical layer technologies like OFDM. Moreover, flow control algorithm should adapt according to the channel variations at the physical layer. Therefore, time scale for flow control at the Transport layer should be the same as the time scale for physical layer resource allocation.

In this paper we propose a two step distributed cross layer design for downlink multi-service OFDMA system. The BS receives service demands from wireless users in the cell and it sends the users requests to the remote servers via a fiber optics network. We assume that each server contains the data demanded by a particular user in the wireless cell. The user demand can either be a Real Time (RT) service such as VoIP or video / audio streaming or a Non-Real Time (NRT) service such as web browsing or file transfer. Traditionally rate control algorithms like TCP were designed for use in wired networks where the capacity of the links is often a fixed quantity. TCP was mainly designed for congestion control and it increases or decreases the data rates depending on the Acknowledgment (ACK) signals [10]-[12]. In wireless system, the capacity of the channel is variable and rate control algorithms designed for wired systems cannot be directly used as they fail to take into account the variable channel capacity. Therefore, in order to develop a good cross layer design we develop an intelligent flow control algorithm at the remote servers which takes into account the variable channel capacity and derives the data rates accordingly. These data rates are then transmitted to the BS. At the BS the subcarrier and power is allocated to different users according to the algorithm developed in [13]. Simulation results show that our proposed cross layer design
ensures traffic priority for multiple flows as well as higher system throughput.

The rest of the paper is organized as follows. In section II system model is described. The problem is formulated in section III. Flow control algorithm is developed in section IV. Utility functions are discussed in section V. Physical layer resource allocation algorithm is described in section VI. Simulation results are presented in section VII while the paper is concluded in section VIII.

II. SYSTEM MODEL

We consider a time slotted downlink OFDMA system with \( K \) users and \( F \) subcarriers. The total transmit power from the base station is constrained to \( P_{\text{max}} \). The data transmission is on time frame basis and each frame consists of \( D \) OFDM symbols. User channels remain unchanged for the duration of a frame but may change from one frame to another. We assume that perfect Channel State Information (CSI) is available at the BS. The channel gain to noise ratio that perfect CSI is available at each user, but may change from one frame to another. We assume that the base station is constrained to assign only one user to each available subcarrier. The channel gain to noise ratio that perfect CSI is available at each user is denoted by \( g(k,f) \), where, \( |h(k,f)| \) denotes the channel coefficient of user \( k \) on subcarrier \( f \) after Fast Fourier Transform (FFT). \( N_0 \) is the power spectral density (PSD) of white noise and \( B \) denotes the bandwidth of single subcarrier. The system model is detailed in Fig.1. Each user in the wireless cell make a service request to the BS. We assume that the BS is connected to \( K \) remote servers via fiber optics. Each of this server contains the data or the service demanded by that the BS is connected to \( K \) remote servers via fiber optics.

A flow controller derives the output data rates from these servers during each time slot based on the average capacity of the channel. This data rate is fed to the physical layer resource allocation block at the BS for power and subcarriers assignment to the users. These allocation decisions are sent to the users via separate control channels which allow them to recover their data.

III. PROBLEM FORMULATION

Let the length of \( k \)th sever queue be denoted by \( Q(k,t) \) at time \( t \). The queue state process evolves according to the following equation,

\[
Q(k,t + 1) = Q(k,t) + A(k,t) - R(k,t)
\]  

(1)

\( A(k,t) \) and \( R(k,t) \) are the arrival and departure processes from \( k \)th server queue at time \( t \). The arrival process is basically the traffic generation rate at the application layer as a result of a particular demand of the user. Let, \( R = [R_1, R_2, \ldots , R_K]^T \) be the data rate vector, then under a given channel state \( \mathbf{H} \), the capacity region of an OFDMA system is given by,

\[
\mathcal{C}(\mathbf{H}) = \left\{ \mathbf{R}(\mathbf{I}) : I_m \cap I_n = \emptyset \forall m \neq n, \bigcup_{k=1}^{K} I_k \subseteq F, \sum_{k=1}^{K} \sum_{f \in I_k} p(k,f) \leq P_{\text{max}} \right\}
\]  

(2)

where, \( \mathbf{I} = [I_1, \ldots , I_K] \) are the subcarriers assigned to the users and \( p(k,f) \) is the power allocated to user \( k \) on subcarrier \( f \). The stability of the system demands that the output rate from the queue should be greater than the traffic generation rate for delay sensitive traffic. From [14], it is clear that the necessary condition for stability demands that, the input arrival rate vector \( \mathbf{A} = [A_1, A_2, \ldots , A_K]^T \in \mathcal{C} \), where \( \mathcal{C} \) is the ergodic capacity region of the channel. The joint flow control and physical resource allocation problem is to maximize the sum of concave utility functions of the queue state process in such a way that the achieved data rate vector should lie within the ergodic capacity region. We use different utility functions depending on the type of service demanded by the user. This optimization problem can be written as,

\[
\max_{\mathbf{R}} \sum_{k=1}^{K} U_k(Q_k + A_k - R_k)
\]  

subject to,

\[
\sum_{k=1}^{K} R_k \in \mathcal{C}
\]  

(4)

Due to time-varying nature of wireless channel, the channel capacity region cannot be known at the Transport layer ahead of the physical resource allocation decisions. Therefore in order to solve our problem in distributed fashion at these two layers, we decouple it into two sub-problems. At the remote server stations we consider the utility maximization problem subject to the average channel capacity \( \overline{C} \) to develop a flow control algorithm,

\[
\max_{\mathbf{R}} \sum_{k=1}^{K} U_k(Q_k + A_k - R_k)
\]  

(5)
subject to,  
\[ \sum_{k=1}^{K} R_k \in \mathcal{C} \]  
(6)

\( \mathcal{C} \) is updated using a low pass filter. In order to update \( \mathcal{C} \), BS communicates the actual total achieved data rate at the physical layer during the last time slot to the flow control algorithm. The flow control algorithm derives the output data rates. These data rates are passed as constraints in the physical layer resource allocation problem. At the physical layer the problem is to maximize the system throughput subject to the data rate constraints imposed by the flow control algorithm. Let \( r(k,f) \) be the data rate allocated to user \( k \) on subcarrier \( f \) then the sum-rate maximization problem at the physical layer is as follows,

\[ \max_{k=1}^{K} \sum_{f \in I_k} r(k,f) \]  
(7)

subject to following constraints,

\[ \sum_{f \in I_k} r(k,f) \geq R_k, \forall k \]
\[ \sum_{k=1}^{K} \sum_{f \in I_k} r(k,f) \leq P_{max} \]

\[ I_m \cap I_n = \Phi, \forall m \neq n, \bigcup_{k=1}^{K} I_k \subseteq \{1,\ldots,F\} \]

By solving this problem we get the actual subcarrier and power allocation decisions. In the next section we develop the flow control algorithm based on optimization problem (5).

IV. FLOW CONTROL ALGORITHM

In this section we develop the flow control algorithm by solving the optimization problem (5),

\[ \max_{k=1}^{K} U_k(Q_k + A_k - R_k) \]  
(8)

subject to,

\[ \sum_{k=1}^{K} R_k \leq \mathcal{C} \]  
(9)

This constraint ensures the feasibility of the resource allocation problem at the physical layer since BS is power limited. By feasibility in this paper we mean that the derived data rates should be such that they can be achieved within the BS power constraint. We update the average channel capacity by using a low pass filter,

\[ \mathcal{C}_{new} = (1-\rho)\mathcal{C}_{old} + \rho \sum_{k=1}^{K} \sum_{f \in I_k} r(k,f) \]  
(10)

where, \( 0 < \rho < 1 \) and \( \sum_{k=1}^{K} \sum_{f \in I_k} r(k,f) \) is the actual achieved data rate at the physical layer after the application of physical resource allocation algorithm. Since we assume that the utility functions are concave this optimization problem can be solved using the convex optimization theory [15], [16]. With \( \lambda \) as the lagrange multiplier associated with the constraint and solving the lagrange KKT conditions we obtain,  
\[ \lambda = U_k'(Q_k + A_k - R_k) \]  
(11)

where, \( U_k' = \partial U_k/\partial R_k \). From equation (11) we get,

\[ R_k = Q_k + A_k - U_k'-1(\lambda) \]  
(12)

Our scheduler propose data rates for the physical layer according to the following algorithm,

1. Initialize \( \lambda > 0 \).
2. Update \( \mathcal{C} = (1-\rho)\mathcal{C}_{old} + \rho \sum_{k=1}^{K} \sum_{f \in I_k} r(k,f) \).
3. While \( \sum_{k=1}^{K} R_k < \mathcal{C} \) repeat,  
a) \( \lambda = \lambda + \Delta \), where \( \Delta \) is a small step size.  
b) Calculate \( R_k \) using eq (6).

This algorithm converges after some iterations. Once the data rates are obtained they are passed on to the physical layer resource allocation algorithm for subcarrier and power allocation decisions.

V. UTILITY FUNCTIONS

Since different traffic types require different treatment based on their QoS objectives, therefore, multiple services are accommodated by selecting the appropriate utility functions. We can broadly categorize all the traffic into two types; a) Real Time (RT) traffic which is delay sensitive such as video/audio streaming services, voice transmission and b) Non-Real Time (NRT) traffic such as file transfer, web browsing etc. where the delay requirements are not so strict. We use the concave utility function designed in [17] for both these traffic types. In [17], a z-shaped function, \( U(Q) = 1 - \frac{1}{1+e^{Q-a_R}} \) is used for RT traffic. \( a_R > 0 \) and \( b_R > 0 \) determine the slope and the inflection point of the functions and \( b_R \) is set to the deadline of RT QoS. Similarly, \( U(Q) = 1 - c_N e^{b_N(Q-b_N)} \) is used for NRT services. \( a_N > 0 \), \( b_N > 0 \) and \( c_N > 0 \) determine the slope, the QoS requirement and the amplitude of the utility function respectively. These utility functions are shown in Fig.2. A detailed discussion about these utility functions is given in [17].

VI. PHYSICAL LAYER RESOURCE ALLOCATION ALGORITHM

In this section we discuss the resource allocation problem. Since this algorithm is the same during each time slot we drop the time index for the sake of continence. Our sum-rate maximization problem is as follows,

\[ \max_{k=1}^{K} \sum_{f \in I_k} r(k,f) \]  
(13)

subject to following constraints,

\[ \sum_{f \in I_k} r(k,f) \geq R_k, \forall k \]
We further define power allocated to user \( k \) of variable we have, \( \gamma_{k,f} \) can be converted into a classical convex optimization problem there are introduces the following constraint, eq.(15) represents a concave function which can be verified from its Hessian which is negative semi-definite when \( \sum_{k=1}^K \sum_{f=1}^F o(k,f) \leq P_{max} \) \( \gamma_{k,f} \) becomes\(^1\), \begin{align*}
 r(k,f) &= \gamma_{k,f} \log (1 + \frac{\alpha}{\gamma_{k,f}} g(k,f)) \\
 \text{We further define } o(k,f) &= \gamma_{k,f} p(k,f), \text{ as the average power allocated to user } k \text{ on subcarrier } f. \text{ With this change of variable we have,}
 \end{align*}
Eq (15) represents a concave function which can be verified from its Hessian which is negative semi-definite when \( \gamma_{k,f} \geq 0 \) and \( o(k,f) \geq 0 \). Finally the optimization problem can now be written as,
\begin{align*}
\max_{\sum_{k=1}^K \sum_{f=1}^F \gamma_{k,f} \log \left(1 + \frac{o(k,f)\gamma_{k,f}}{\gamma_{k,f}}\right)}
\text{subject to,}
\sum_{f=1}^F \gamma_{k,f} \log \left(1 + \frac{o(k,f)\gamma_{k,f}}{\gamma_{k,f}}\right) \geq R_k & \quad \forall k \\
\sum_{k=1}^K \gamma_{k,f} \leq 1 & \quad \forall f
\end{align*}
\( o(k,f) \) and \( \alpha \) as the Lagrange multipliers associated with constraints (17), (18) and (19) respectively and solving the appropriate Lagrange-KKT optimality conditions we get,
\begin{align*}
p(k,f) &= \left(1 + \frac{\delta(k)}{\alpha} - \frac{1}{g(k,f)}\right)^+ \\
\left(1 + \delta(k)\right) \left(\log \left(1 + \frac{\delta(k)}{\alpha} g(k,f)\right)\right)^+ - \left(1 - \frac{\alpha}{\left(1 + \delta(k)\right)\gamma_{k,f}}\right)^+ = \mu_f
\end{align*}
Based on the equations (20) and (21), efficient resource allocation algorithms can be developed. For this paper we consider the optimal algorithm developed in [13]. This algorithm achieves the rate and power constraints simultaneously. This algorithm consists of an inner loop and an outer loop. The outer loop starts with a small value of \( \alpha > 0 \) and increments it in small steps. For each value of \( \alpha \), the inner loop allocates the power and subcarriers to the users according to their data rate constraints. This process is repeated till all the constraints are satisfied. We produce this algorithm from [13] in Table I.

**VII. SIMULATION RESULTS**

We consider a downlink OFDMA system with 10 users and 24 subcarriers. Base station has a perfect channel state information and a peak power constraint of 43dBm. We consider a frequency selective Rayleigh fading channel with exponential delay profile. Path losses are calculated according to Cost-Hata Model [18]. The power spectral density of noise
is -174 dbm/Hz. Time is divided into slots and duration of each Transmission Time Interval (TTI) is 1 ms. We assume that packets are generated according to poisson distribution with packet size of 1 Kbits. The users are uniformly distributed in a cell of radius 700m. Moreover, the bandwidth of each subcarrier is 375 KHz. The parameters for RT utility functions are \( a_R = 1.5 \) and \( b_R = 5 \text{ ms} \), while those for NRT utility function are \( a_N = 0.1, b_N = 25 \text{ ms} \) and \( c_N = 0.5 \). We consider that there are 5 RT and 5 NRT users.

In Fig. 3, we plot the average backlog of the system vs the traffic generation rate at the application layer. We are interested in determining the maximum traffic generation rate that can ensure that the data rates demanded by the flow control algorithm remain within the ergodic capacity region. From the figure, we can see that the traffic generation rate has to be constrained to 5.5 packets/TTI/user in order to achieve empty queues all the time. However, when this rate is further increased, queue backlogs start to grow because the data rates by the flow control algorithm cannot be achieved at the physical layer.

![Fig. 3. Average backlog vs Traffic generation rate (packets/TTI/user) for 10 user 24 subcarrier system.](image)

In Fig. 4, we compare the throughput achieved by RT and NRT users in the system. When the traffic generation rate lies within the ergodic capacity region, we can see that our scheme does not discriminate between the two traffic types and all the demanded data rates are achieved. However, when the traffic generation rate is further increased, RT users achieve higher data rates compared to NRT users. This is due to the steeper slope of RT utility function which results in higher priority for RT users and hence higher throughput.

![Fig. 4. Total Achieved Throughput for 5 RT and 5 NRT users.](image)

**REFERENCES**


