I. J. Communications, Network and System Sciences, **2008, 3, 207-283**
Published Online August 2008 in SciRes (http://www.SciRP.org/journal/ijcns/).

Priority-Based Resource Allocation for Downlink OFDMA Systems Supporting RT and NRT Traffics

Hua WANG, Lars DITTMANN
Department of Communications, Optics & Materials
Technical University of Denmark, Lyngby, Denmark
E-mail: {huw, ld}@com.dtu.dk
Received on March 11, 2008; revised and accepted on May 22, 2008

Abstract

Efficient radio resource management is essential in Quality-of-Service (QoS) provisioning for wireless communication networks. In this paper, we propose a novel priority-based packet scheduling algorithm for downlink OFDMA systems. The proposed algorithm is designed to support heterogeneous applications consisting of both real-time (RT) and non-real-time (NRT) traffics with the objective to increase the spectrum efficiency while satisfying diverse QoS requirements. It tightly couples the subchannel allocation and packet scheduling together through an integrated cross-layer approach in which each packet is assigned a priority value based on both the instantaneous channel conditions as well as the QoS constraints. An efficient suboptimal heuristic algorithm is proposed to reduce the computational complexity with marginal performance degradation compared to the optimal solution. Simulation results show that the proposed algorithm can significantly improve the system performance in terms of high spectral efficiency and low outage probability compared to conventional packet scheduling algorithms, thus is very suitable for the downlink of current OFDMA systems.

Keywords: OFDMA, Radio Resource Management, Quality of Service, Real-time and Non-real-time Traffics

1. Introduction

Orthogonal Frequency Division Multiple Access (OFDMA) is an attractive multiple access scheme for future wireless and mobile communication systems, which has been developed to support a variety of multimedia applications with different Quality-of-Service (QoS) requirements. OFDMA builds on Orthogonal Frequency Division Multiplexing (OFDM), which is immune to intersymbol interference and frequency selective fading, as it divides the frequency band into a group of mutually orthogonal subcarriers, each having a much lower bandwidth than the coherence bandwidth of the channel. In multi-user environment, each user is dynamically assigned to a subset of subcarriers in each frame, to take advantage of the fact that at any time instance, channel responses are different for different users and at different subcarriers [1]. This capability of OFDMA systems enables the network to perform a flexible radio resource management, such as dynamic subcarrier assignment (DSA), adaptive power allocation (APA), and adaptive modulation and coding (AMC) scheme to improve the system performance significantly under different traffic loads and time-varying channel conditions.

Recently, radio resource management for OFDMA systems has attracted enormous research interests in both academia and industry. Many scheduling algorithms have been proposed which can adapt to changes in users' channel conditions and QoS requirements. In the literature, the resource allocation problem can be divided into two categories with different design objectives. The objective of the first category is to minimize the total transmit power subject to individual data rate constraints, see [2–4]. The objective of the second category aims at maximizing the overall (weighted) transmission rate subject to power constraints, see [5–7]. In either case, the optimal resource allocation solutions are difficult to get due to high computational complexity of non-linear

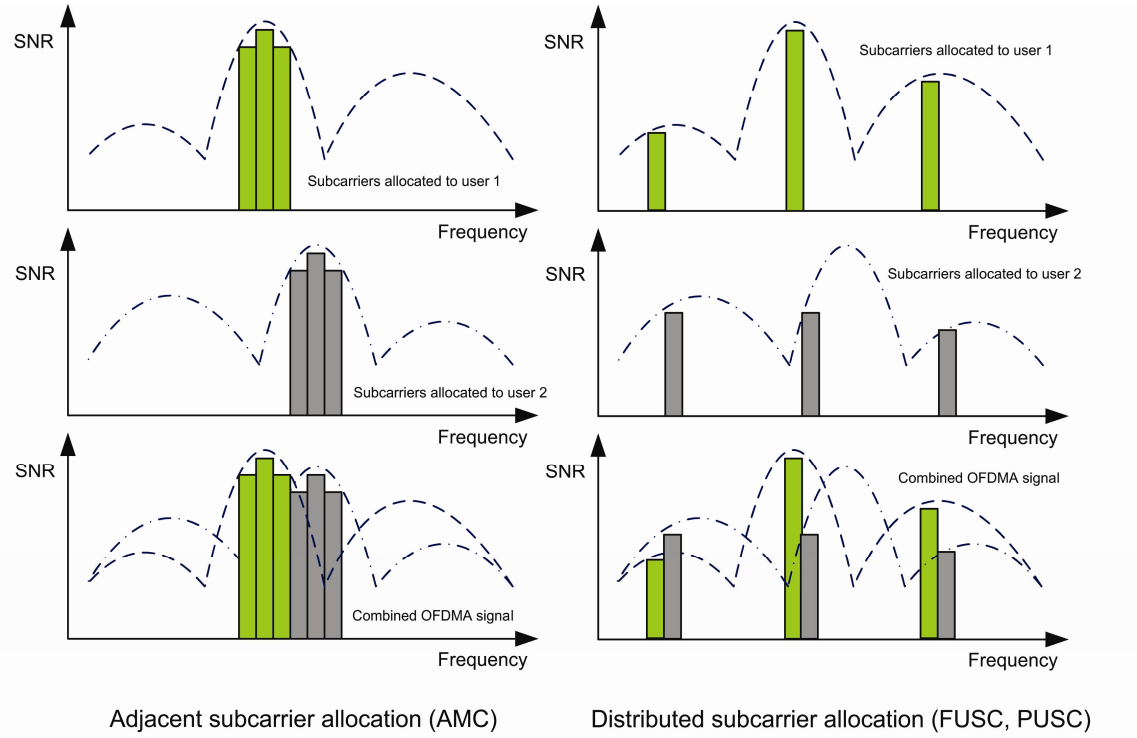

Figure 1. Adjacent and distributed subcarrier allocation.

optimization with integer variables. Instead, suboptimal solutions based on relaxation, problem splitting, or heuristic algorithms are proposed to reduce computational complexity [8]. Such algorithms are often refereed to as *loading algorithms*.

In most loading algorithms, the QoS requirement of each user is usually defined in terms of a fixed number of transmission bits per frame. However, in practical communication systems, it is neither sufficient nor efficient to represent different QoS requirements solely by a fixed data rate per frame. The resource allocation problem for systems supporting both realtime (RT) and non-real-time (NRT) multimedia traffic becomes much more complicated when diverse QoS requirements have to be considered. The transmission of RT packets can be delayed as long as the delay constraint is not violated, and the transmission of NRT packets can be more elastic. Furthermore, most loading algorithms assume that users always have data to transmit, which is not the case in real systems. Instead, appropriate traffic models should be taken into account in the design of scheduling algorithms. Therefore, efficient *packetbased scheduling algorithms* are of interest. Many packet scheduling algorithms with different design objectives have been proposed in the literature [9–11].

In this paper, we propose a novel resource allocation algorithm for downlink OFDMA systems supporting both RT and NRT multimedia traffic. Unlike the conventional approaches, which decompose the resource allocation into two steps: packet scheduling and subcarrier-and-power allocation [4,11], the proposed algorithm tightly couples these two steps together through an integrated cross-layer approach to take advantage of the inter-dependencies between PHY and MAC layers. The basic idea is that if a packet is scheduled for transmission on a specific subchannel, it will get a priority value based on both the instantaneous channel conditions as well as the QoS requirements. Then we can formulate the resource allocation problem into an optimization problem with the objective to maximize the total achievable priority values. A suboptimal heuristic algorithm is also proposed to reduce the computational complexity. Simulation results show that the proposed algorithms can achieve high spectral efficiency with satisfying QoS performance in each service class.

The rest of the paper is organized as follows. We first give a brief introduction of the system model in Section 2. The resource allocation problem is formulated in Section 3. Section 4 presents a suboptimal heuristic algorithm with low computation complexity. Simulation environments and results are outlined and discussed in Section 5. Finally, conclusions and future work are drawn in Section 6.

2. System Model

OFDMA is a multiple access scheme based on OFDM. While OFDM employs fast Fourier transform (FFT) of

size 256 (subcarriers) in fixed WiMAX, OFDMA employs a larger FFT space (2048 and 4096 subcarriers) which are further grouped into subchannels. The subchannels are assigned to different users and may employ different modulation and coding schemes to exploit frequency diversity as well as time diversity [12]. There are two approaches of allocating subcarriers to form a subchannel in OFDMA: distributed subcarrier permutation and adjacent subcarrier permutation. The two approaches are shown in Figure 1. In distributed subcarrier permutation, a subchannel is formed with different subcarriers randomly distributed across the channel spectrum. This approach maximizes the frequency diversity and averages inter-cell interference. It is suitable for mobile environment where channel characteristics change fast. Both partial usage of subchannels (PUSC) and full usage of subchannels (FUSC) schemes employ distributed subcarrier permutation. In adjacent subcarrier permutation, a subchannel is formed by grouping adjacent subcarriers. This approach creates a 'loading gain' and is easy to use with beam-forming adaptive antenna system (AAS). It is suitable for stationary or nomadic environment where channel characteristics change slowly. The AMC scheme employs adjacent subcarrier permutation.

In this paper, we assume that subscriber stations are stationary or nomadic users with slowly varying channel conditions. Therefore, adjacent subcarrier permutation strategy is employed to support AMC. In OFDMA, radio resource is partitioned in both frequency domain and time domain, which results in a hybrid frequency-time domain resource allocation. It provides an added dimension of flexibility in terms of higher granularity compared to OFDM/TDM systems.

We consider the downlink scenario of an infrastructure-based OFDMA system with U_s subcarriers and K users. At the physical layer, the time axis is divided into frames with fixed length, each of which consists of a downlink (DL) and an uplink (UL) subframe to support TDD operation. In each DL subframe, there are U_t time slots available for downlink transmissions, each of which may contain one or several OFDM symbols. To reduce the resource addressing space, channel coherence in frequency and time is exploited by grouping I_s adjacent subcarriers and I_t time slots to form a basic resource unit (BRU) for resource allocation. A BRU is the minimum resource allocation unit as shown in Figure 2. The size of a BRU is adjusted so that the channel experiences flat fading in both frequency and time domain. Thus in each DL subframe, there are $S = U_s/I_s$ subchannels in frequency domain and $N = U_t/I_t$ slots in time domain, which corresponds to a total of $S * N$ BRUs available in frequency-time domain for DL transmissions. Each BRU can be assigned to different users and be independently bit and power loaded. In principle, adaptive power allocation in each BRU can improve the system performance. However, some studies show that performance improvements are only marginal over a wide

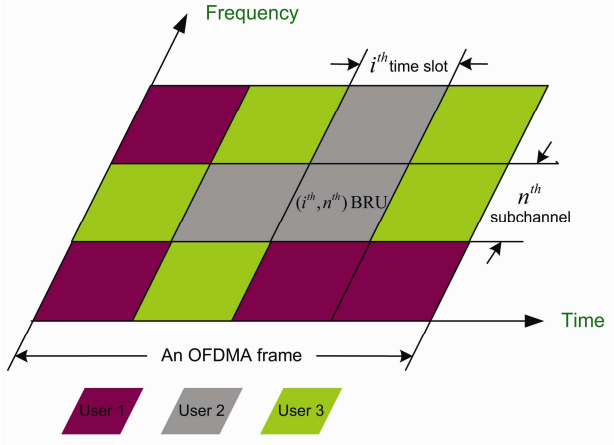

Figure 2. Frequency-time domain redio resource allocation in OFDMA systems.

range of SNRs due to the statistical effects [1]. Therefore, we assume that the total transmission power is equally distributed among all subchannels.

We further assume that in each frame the base station (BS) has perfect knowledge of channel state information (CSI) for each subchannel of each user. This can be obtained by piggybacking such information in each uplink packet, which is suitable for slowly varying channels. Based on CSI, adaptive modulation and coding scheme is employed to adjust the transmission mode dynamically according to the time-varying channel conditions. Multiple transmission modes are available, with each mode representing a pair of specific modulation format and a forward error correcting code. The transmission mode is determined by the instantaneous signal-to-noise ratio (SNR). To utilize the PHY layer resources more efficiently, fragmentation at the MAC layer is enabled. A separate queue with a finite queue length of L MAC protocol data units (PDUs) is maintained for each connection at the base station. We assume that the MAC PDUs are of fixed size, each of which contains d information bits.

3. Resource Allocation Model

The resource allocation at the BS involves the allocation of subchannels, time slots, and modulation order and coding rate assignment. It is executed at the beginning of every frame to properly allocate radio resources to the demanding users according to their queue status, CSI, and QoS requirements.

The real-time traffic is delay-sensitive and has strict delay requirement. The non-real-time traffic can tolerate longer delays, but requires a minimum throughput. We propose a novel priority-based packet scheduling algorithm to support both RT and NRT multimedia traffic with high spectral efficiency and good QoS satisfaction. The basic idea behind the proposed algorithm is that the transmission is scheduled on a

packet-bypacket basis. Specifically, at each scheduling interval, if a PDU was scheduled for transmission on a specific subchannel, it is assigned a priority value based on the instantaneous channel condition (PHY layer issue), as well as the QoS constraint (MAC layer issue). Then we can formulate the scheduling problem into a mathematical optimization problem with the objective to maximize the total achievable priority values.

We apply an extended EXP algorithm as our priority function for both RT and NRT traffics. The EXP rule was proposed to provide QoS guarantees over a shared wireless link in terms of the average packet delay for RT traffic and a minimum throughput for NRT traffic [15].

For RT traffic, if the i^{th} PDU from the k^{th} connection is scheduled for transmission on subchannel n, its priority value is calculated as:

$$\mathbf{P}(k,i,n) = a_k \cdot \frac{\mu_{k,n}(t)}{\overline{\mu}_k(t)} \cdot \exp\left(\frac{a_k W_{k,i}(t) - \overline{aW}}{1+\sqrt{\overline{aW}}} \right) \quad \textbf{(1)}$$

where $\overline{aW} = \frac{1}{k}\sum_k a_k W_{k,1}(t)$,and $a_k = -\log \delta_k / T_{k,\max} \cdot W_{k,i}(t)$ is the i^{th} PDU delay of connection k at time t, $T_{k,max}$ is the maximum allowable delay of connection k, δ_k is the maximum outage probability of connection k, $\mu_{k,n}(t)$ is the instantaneous channel rate with respect to the signal-to-noise ratio and a predetermined target error probability if subchannel n is assigned to connection k at time t, and $\overline{\mu}_k(t)$ is the exponential moving average (EMA) channel rate of connection k with a smoothing factor t_c, calculated as:

$$\overline{\mu}_k(t) = \left(1 - \frac{1}{t_c}\right)\overline{\mu}_k(t-1) + \frac{1}{t_c}\mu_k(t) \quad \textbf{(2)}$$

where $\mu_k(t) = \sum_{n=1}^{N} c_{k,n} \cdot \mu_{k,n}(t)$ is the total channel rate of connection k at time t. If subchannel n is assigned to connection k, $c_{k,n}=1$, otherwise $c_{k,n}=0$.

For NRT traffic, the extended EXP algorithm is used in conjunction with a token bucket control to guarantee a minimum throughput [15]. We associate each NRT queue with a virtual token bucket. Tokens in each bucket arrive at a constant rate $r_{k,req}$, which is the required minimum throughput of connection k. Let us define $V_{k,i}(t)$ to be the virtual waiting time of the i^{th} PDU from connection k:

$$V_{k,i}(t) = \frac{\max\{0, Q_k(t) - (i-1)\cdot d\}}{r_{k,req}} \quad k \in NRT \quad \textbf{(3)}$$

where $Q_k(t)$ is the number of tokens associated with connection k at time t, and d is the fixed size of a MAC PDU. Note that we do not need to actually maintain the virtual waiting time, as the arrival rates of tokens are constant. Then, the calculation of the priority for a NRT PDU is similar to Exp.(1), with $W_{k,i}(t)$ being replaced by $V_{k,i}(t)$. After a PDU is scheduled for transmission, the number of tokens in the corresponding token queue is reduced by the actual amount of data transmitted.

Let $\mathbf{u}(k,i,n)$ be defined as a binary random variable indicating subchannel allocation. That is, $\mathbf{u}(k,i,n)=1$ means that the i^{th} PDU from connection k is allocated for transmission on subchannel n, and $\mathbf{u}(k,i,n)=0$ otherwise. Also let us define $\mathbf{m}(k,i,n)$ be the number of time slots occupied on subchannel n if the i^{th} PDU from connection k is scheduled for transmission on subchannel n, calculated as:

$$\mathbf{m}(k,i,n) = \left\lceil \frac{d}{\mu_{k,n}(t)} \right\rceil \quad \textbf{(4)}$$

where $\lceil x \rceil$ denotes the smallest integer larger than x.

Then, the scheduling problem can be mathematically formulated as follows:

$$\arg\max_{\mathbf{u}(k,i,n)} \sum_{k=1}^{K}\sum_{i=1}^{L}\sum_{n=1}^{S} \mathbf{u}(k,i,n)\cdot \mathbf{P}(k,i,n) \quad \textbf{(5)}$$

Subject to:

$$\sum_{k=1}^{K}\sum_{i=1}^{L} \mathbf{u}(k,i,n)\cdot \mathbf{m}(k,i,n) \le N \quad \forall n \quad \textbf{(6)}$$

$$\sum_{n=1}^{S} \mathbf{u}(k,i,n) \le 1 \quad \forall k,i \quad \textbf{(7)}$$

$$\mathbf{u}(k,i,n) \in \{0,1\} \quad \forall k,i,n \quad \textbf{(8)}$$

where S denotes the total number of subchannels, N denotes the total number of time slots, K denotes the total number of connections, and L denotes the maximum queue size.

The first constraint ensures that the allocated bandwidth does not exceed the total available bandwidth in terms of time slots on each subchannel. The second constraint says that a PDU can only be transmitted via one subchannel. The instantaneous channel conditions and the QoS related parameters are embodied into the priority function $\mathbf{P}(k,i,n)$ with the objective of maximizing the total achievable priority values, thus improving the spectral efficiency while maintaining QoS guarantees.

The above optimization problem can be solved by determining the values of binary variable $\mathbf{u}(k,i,n)$ through standard linear integer programming (LIP)[1]. The solution to the problem provides an optimal resource allocation.

[1]The optimal solution of the LIP problem formulated in this paper is obtained by using the General Algebraic Modeling System(GAMS).

However, the computation complexity of the optimal solution is too high to be applied in practical systems. To reduce the computational complexity, we propose a suboptimal heuristic algorithm with low complexity in the next section.

4. Proposed Suboptimal Scheme

In the suboptimal algorithm, we allocate radio resources on a packet-by-packet basis. The general idea is that, at each scheduling interval, the packet with the highest priority value from all queues is scheduled for transmission, and this procedure continues until either there is no radio resource left or there is no packet remaining unscheduled in the queue. A detailed description of the proposed scheduling algorithm is listed in pseudocode 1, where Ω_s^k is the set of subchannels that are available for data transmission of connection k, t_n is the number of residual time slots on subchannel n, q_k is the current queue size of connection k, and i_k is a pointer to the next PDU to be scheduled of connection k.

It works as follows: If connection k has pending traffic in the queue, the proposed algorithm first pre-allocates the best subchannel n in terms of the instantaneous channel quality to connection k from its available subchannel set Ω_s^k (see Step 14). If there is not enough capacity left on the best subchannel n to accommodate one PDU from connection k's queue, subchannel n will be removed from connection k's available subchannel set, and the second best subchannel n' will be selected. This procedure continues until a best possible subchannel is pre-allocated to connection k (see Step 13-22). Otherwise, connection k is removed from the scheduling list. After the subchannel pre-allocation process for all connections is complete, the algorithm calculates the priority value of the head-of-line (HOL) PDU in each nonempty queue, and schedule the PDU with the highest priority value for transmission on subchannel n^* (see Step 16 & 24). The scheduled PDU is removed from the corresponding queue and the consumed radio resources in terms of time slots are subtracted on subchannel n^* (see Step 25 & 26-30). Then it starts from the beginning and continues until either there is no radio resource left or there is no PDU pending in the queue. A detailed flowchart of the proposed suboptimal algorithm is given in Appendix I.

5. Simulation Results and Discussions

To evaluate the performance of the proposed resource allocation algorithm for downlink OFDMA systems supporting both RT and NRT multimedia traffic, a system-level simulation is performed in OPNET.

Algorithm 1 Suboptimal Packet Scheduling Algorithm for Downlink OFDMA Systems

1: Set $t_n \leftarrow N$ for $\forall n$ {initialize t_n}
2: Set $i_k \leftarrow 1$ for $\forall k$ {initialize i_k}
3: Get q_k for $\forall k$ {get the queue size of connection k}
4: **for** k =1 to K **do**
5: **if** $q_k > 0$ **then**
6: Set $\Omega_s^k \leftarrow$ {1,…,S} {initialize Ω_s^k}
7: **else**
8: Set $\Omega_s^k \leftarrow \phi$ {set Ω_s^k to be null}
9: **end if**
10: **end for**
11: **while** $\exists x, \Omega_s^x \neq \phi$ **do**
12: **for** k =1 to K **do**
13: **while** $\Omega_s^k \neq \phi$ **do**
14: Select $n \leftarrow \arg\max_{n \in \Omega_s^k} \mu_{k,n}(t)$ {assign the best subchannel from the available subchannel set}
15: **if** $t_n \geq \left\lceil \frac{d}{\mu_{k,n}(t)} \right\rceil$ **then**
16: Calculate P(k,i_k,n) in Exp. (1)
17: BREAK
18: **else**
19: $\Omega_s^k \leftarrow \Omega_s^k -$ {n} {remove n from the available subchannel set if there is not enough capacity left}
20: CONTINUE
21: **end if**
22: **end while**
23: **end for**
24: Schedule the i_{k^*} th PDU of connection k^* on subchannel n^*, where $(k^*, i_{k^*}, n^*) \leftarrow \arg\max \mathbf{P}(k, i_k, n)$
25: $t_{n^*} \leftarrow t_{n^*} - \left\lceil \frac{d}{\mu_{k^*,n^*}(t)} \right\rceil$ {update the residual time slots}
26: **if** $i_{k^*} = q_k$ **then**
27: $\Omega_s^{k^*} \leftarrow \phi$ {set $\Omega_s^{k^*}$ to be null when all pending PDUs of connection k^* have been scheduled for transmission}
28: **else**
29: $i_{k^*} \leftarrow i_{k^*} + 1$ {point to the next pending PDU}
30: **end if**
31: **end while**

5.1. System Model

We consider the downlink of a single-cell OFDMA system with TDD operation. The cell radius is 2 km, where subscriber stations are randomly placed in the cell with uniform distribution. The total bandwidth is set to be 5 MHz, which is divided into 10 subchannels. The BS transmit power is set to 20W (43 dBm) which is evenly distributed among all subchannels. The duration of a frame is set to be 1 ms so that the channel quality of each connection remains almost constant within a frame, but may vary from frame to frame. The propagation model is derived from IEEE 802.16 SUI channel model [20]. Path loss is modeled according to terrain Type A suburban macro-cell. Large-scale shadowing is modeled by log-normal distribution with zero mean and standard deviation of 8 dB. The rms delay spread is 0.5μs, typical of an urban environment. The effect of small scale

Table 1. A summary of system parameters.

Parameters	Value
System	OFDMA/TDD
Central frequency	3500 MHz
Channel bandwidth	5 MHz
Number of subchannels	10
Length of OFDM symbol	156.25 μs
User distribution	Uniform
Beam pattern	Omni-directional
Cell radius	2 km
Frame duration	1 ms
BS transmit power	20 W
Thermal noise density	–174 dBm/Hz
Propagation model	802.16 SUI-5 Channel model
Maximum MAC PDU size	256 bytes

Table 2. Modulation and Coding Schemes for 802.16 [16].

Modulation scheme	Coding rate	Bits/ symbol	Target SNR for 1% PER (dB)
BPSK	1/2	0.5	1.5
QPSK	1/2	1	6.4
QPSK	3/4	1.5	8.2
16QAM	1/2	2	13.4
16QAM	3/4	3	16.2
64QAM	1/2	4	21.7
64QAM	3/4	4.5	24.4

Table 3. A summary of traffic paramenters.

Type	Characteristics	Distribution	Parameters
VoIP	ON period	Exponential	Mean=1.34 sec
VoIP	OFF period	Exponential	Mean=1.67 sec
VoIP	Packet size	Constant	66 bytes
VoIP	Inter-arrival time between packets	Constant	20 ms
Video	Packet size	Log-normal	Mean=4.9 bytes Std.dev.=10 ms
Video	Inter-arrival time between packets	Normal	Mean=33 ms Std.dev.=10 ms
Web	Reading time between sessions	Exponential	Mean=5 sec
Web	Number of packets within a packet call	Geometric	Mean=25 packets
Web	Inter-arrival time between packets	Geometric	Mean=0.0277 sec
Web	Packet size	Truncated Pareto	k=81.5 bytes α=1.1 m=2 M bytes

multipath fading is modeled by a tapped delay line (TDL) with exponential power delay profile as follows:

$$h(\tau,t)=\sum_{i=0}^{N}\beta_i(t)\delta(\tau-\tau_i(t)) \quad \textbf{(9)}$$

where N is the total number of paths, $\delta(\cdot)$ is the Dirac impulse, $\beta_i(t)$ and $\tau_i(t)$ are the time-variant gain and delay of the i^{th} path, respectively. The channel gains $\beta_i(t)$ are zero mean mutually independent Gaussian stationary processes with an exponentially decaying power profile and a classical Jake's spectrum. The thermal noise density is assumed to be –174 dBm/Hz.

Table 1 summarizes the system parameters used in the simulation. We assume that all MAC PDUs are transmitted and received without errors and the transmission delay is negligible. The modulation order and coding rate in the AMC scheme is determined by the instantaneous SNR of each user on each subchannel. We follow the AMC table shown in Table 2, which specifies the minimum SNR required to meet a target packet error rate, e.g., 1%.

5.2. Traffic Model

In the simulation, three types of traffic sources are generated:

- **Real-time (RT) voice**: RT voice traffic is assumed to be VoIP that periodically generates packets of fixed size. Assuming that silence suppression is used, VoIP traffic can be modeled as a two-state Markov ON/OFF source [17].
- **Real-time (RT) video**: RT video traffic is assumed to be the videoconference which consists of a VoIP source and a video source [17]. A video source periodically generates packets of variable size.
- **Non-real-time (NRT) data**: NRT data traffic is assumed to be Internet traffic such as web browsing that requires large bandwidth and generates bursty data of variable size. We apply the Web browsing model for the Internet traffic [18].

It is assumed that each user has a connection pair consisting of a RT connection and a NRT connection. VoIP and video traffic is served in RT connection while data traffic is served in NRT connection. Each connection alternates between the states of idle and busy, which are both exponentially distributed, and is loaded with corresponding traffic source when the connection is in busy state. A summary of traffic parameters of different traffic types are listed in Table 3.

5.3. Performance Evaluation

We evaluate and compare the performance of the proposed priority-based scheduling algorithm with other conventional algorithms in terms of the average packet delay, the throughput, the outage probabilities, and the modulation efficiency via extensive computer simulations.

For delay-sensitive RT traffic, the *average packet delay* and the *delay outage probability* are the main performance metrics. The delay constraint for RT traffic is set to be 50ms. For loss-sensitive NRT traffic, the *average throughput* and the *throughput outage probability* are the main performance metrics. The minimum throughput constraint for NRT traffic is set to be 100 Kbits/sec. The outage probabilities for both RT

and NRT traffic should be less than 3%. In order to evaluate the spectral efficiency, the *modulation efficiency* is also considered in the performance evaluation.

For comparisons, we include the simulation results of two conventional scheduling algorithms proposed for OFDMA systems. The first one is maximum SNR, where users are selected for transmission over each subchannel according to their CSI. The second one is proportional fair (PF) [14], where users are selected for transmission over subchannel *n* according to the following criteria:

$$i_n^* = \arg\max_i \frac{\mu_{i,n}(t)}{\overline{\mu}_{i,n}(t)} \quad \textbf{(10)}$$

where $\overline{\mu}_{i,n}(t)$ is the average data rate of the n^{th} subchannel of user *i*. To compare the performance between OFDM/TDM and OFDMA based systems, simulation results of the EXP rule applied in OFDM/TDM systems are also included. The EXP rule is considered to be one of the best scheduling algorithms in OFDM/TDM based systems [15], of which each user transmits in the assigned time slots over all subchannels.

Figure 3 shows the average packet delay of RT traffic versus the number of users for different scheduling algorithms. When the number of users is below 48, the average packet delay of the proposed scheme increases marginally and it is well kept below the maximum allowable delay, which is 50 ms in our scenario. After that point, the system is overloaded and the average packet delay increases sharply. Similar phenomenon of the proposed scheme can be observed for the delay outage probability shown in Figure 4. However, the average packet delay of the PF scheme and the MAX-SNR scheme is much larger compared to our proposed scheme, which consequently results in a higher delay outage probability when the number of users is below 48. Furthermore, it can be seen from Figure 4 that when the number of users is above 48, the delay outage probability

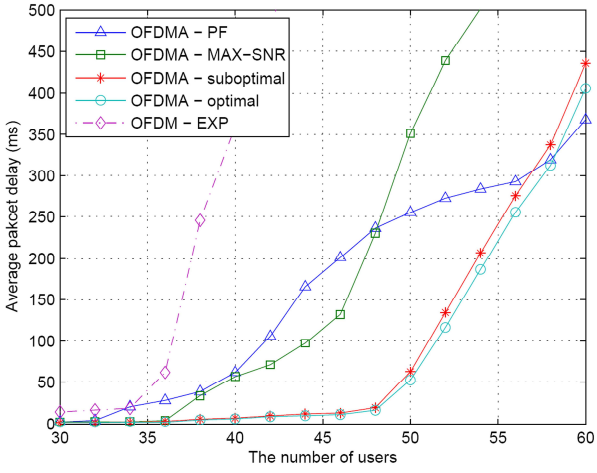

Figure 3. Average packet delay in RT.

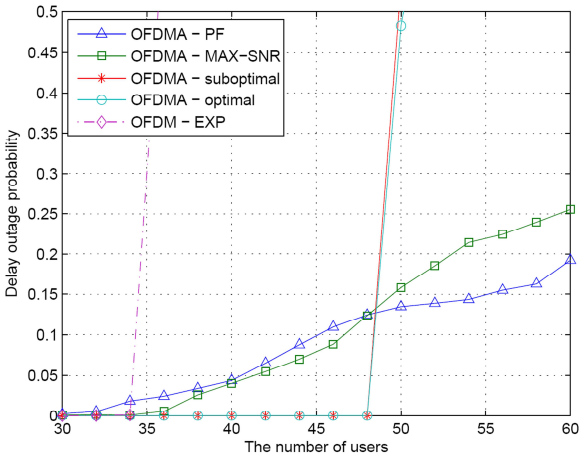

Figure 4. Delay outage probability in RT.

of the proposed scheme increases rapidly to one, which means that the system is overloaded and almost no RT connections can maintain the required delay constraint. On the other hand, some RT connections in the PF and MAX-SNR schemes can still maintain the required delay constraint as the delay outage probabilities in these two schemes increase steadily with respect to the number of users. This is because in the proposed scheme, it not only takes the instantaneous channel conditions, but also the delay requirement into consideration when scheduling packets. RT connections with larger packet delay are assigned higher priorities in an effort to average out the packet delay among all RT connections. As a result, each RT connection will have similar average packet delay regardless of its channel conditions. When the system is overloaded, congestion occurs and all RT connections will experience bandwidth starvation, which results in a sharp increase of the average packet delay and the delay outage probability. However, in the PF and MAX-SNR schemes, the scheduler selects a connection for transmission only based on instantaneous channel conditions. As a consequence, connections with good channel conditions will always experience very short delay at the cost of bandwidth starvation for connections with poor channel conditions. Therefore, the delay outage probability in the PF and MAX-SNR schemes increases much more smoothly compared to the proposed scheme when the number of users is above 48. As for the EXP rule applied in OFDM/TDM systems, the dotted line in Figure 3 & 4 indicates that the performance of OFDM-based system is much worse than OFDMA-based system.

Figure 4 shows the delay outage probability of RT traffic versus the number of users for different scheduling algorithms. It is obvious that the proposed scheme outperforms over the other conventional schemes. The maximum number of supportable RT users under a predefined 3% outage probability in PF, MAX-SNR and the proposed scheme are 38, 38, and 48 respectively.

Figure 5 shows the throughput of NRT traffic versus the number of users for different scheduling algorithms. The throughput of the MAX-SNR scheme achieves the highest value among all schemes. It increases proportional to the number of users. In the proposed scheme, the throughput increases proportional to the number of users when there are less than 50 users. After that point, the throughput remains on a steady level regardless of the number of users. While in the PF scheme, the throughput is significantly lower than the other schemes. It can be explained as follows: In the MAXSNR scheme, the scheduler simply selects the connection with the best CSI for transmission. When the number of users increases, the scheduler has more chance to serve a user in good channel conditions (multi-user diversity gain) which results in a high throughput. That's why the spectral efficiency of the MAX-SNR scheme increases with respect to the number of users shown in Figure 7. In the proposed scheme, both the CSI and the QoS constraints are taken into account to guarantee the required QoS performance (i.e., a minimum throughput of 100Kbpbs for each NRT connection). When the system is underloaded (the number of users is less than 50), the bandwidth is large enough to satisfy the QoS requirements of all connections and the scheduling criterion mainly concerns with the CSI of each connection. As a result, the throughput as well as the spectral efficiency increases proportional to the number of users. However, when the system is overloaded (the number of users is above 50), the bandwidth is not sufficient to satisfy the QoS requirements of all connections. Thus congestion occurs and the throughput reaches at a steady level. When congestion occurs, the proposed algorithm tends to put more weight on the QoS constraint than the CSI in an effort to provide equal opportunities of QoS satisfaction among all NRT connections. In other words, the throughput of each NRT connection in the proposed scheme decreases proportionally to the number of users when the system is overloaded. That explains a sharp increase of the throughput outage probability shown in Figure 6. In the PF scheme, the throughput is relatively low due to the reason that the spectral efficiency is significantly lower than the MAX-SNR and the proposed schemes. Again, from Figure 5 & 6, we can see that OFDMA based scheduling algorithms have better performance than OFDM/TDM based scheduling algorithm.

Figure 6 shows the throughput outage probability of NRT traffic versus the number of users for different scheduling algorithms. It is obvious that the proposed scheme outperforms over the other conventional schemes. The maximum number of supportable NRT users under a predefined 3% outage probability in PF, MAX-SNR and the proposed scheme are 34, 34, and 50 respectively.

Figure 7 depicts the normalized spectral efficiency, which is defined as the ratio between the achieved

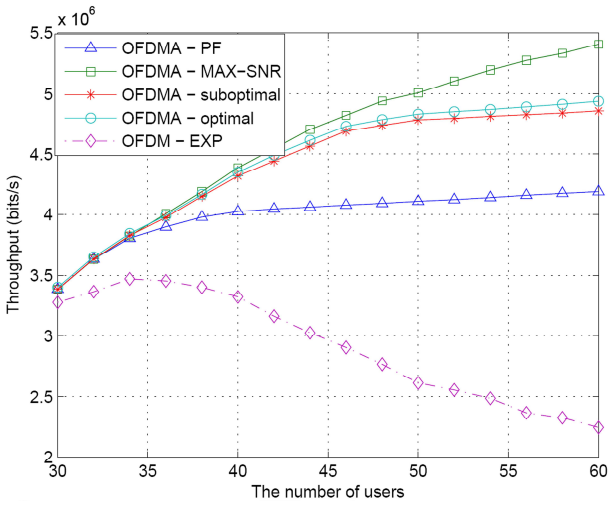

Figure 5. Average throughput in NRT.

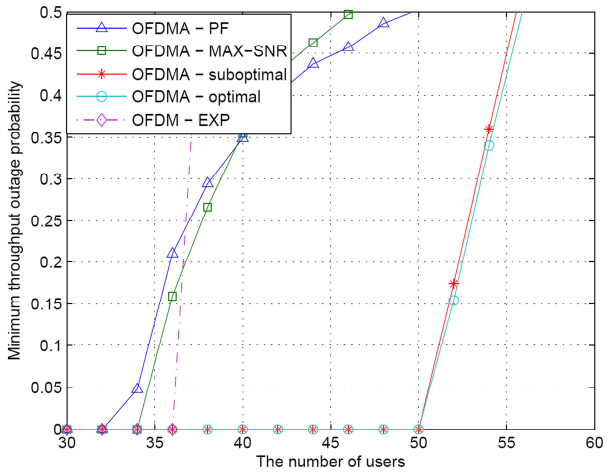

Figure 6. Throughput outage probability in NRT.

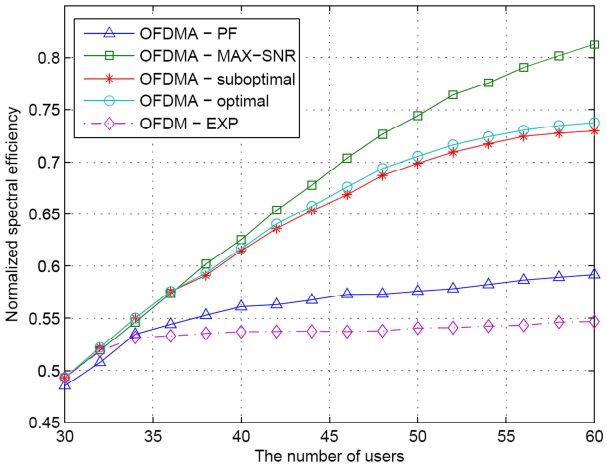

Figure 7. Normalized spectral efficiency in RT and NRT.

modulation and the highest modulation, under different schemes. It can be seen that the MAX-SNR scheme achieves the highest spectral efficiency due to the reason that in the the MAX-SNR, the connection with the best CSI is selected for transmission. The proposed scheme can also achieve a relatively high spectral efficiency as it takes both the channel condition as well as the QoS constraints into account when scheduling packets. While the spectral efficiency in the PF scheme is relatively low compared to the MAX-SNR and the proposed schemes.

From the above figures, we can see that the performance of the proposed suboptimal scheme is close to the optimal scheme, but with considerably low computation complexity. We can also see that the OFDMA based scheduling algorithms outperform the OFDM/TDM based scheduling algorithm as expected.

This is because in OFDMA systems, we can not only exploit multiuser diversity in the time domain, but also in the frequency domain.

6. Conclusions and Future Work

This paper addresses the problem of QoS scheduling and resource allocation for downlink OFDMA systems supporting both real-time (RT) and non-real-time (NRT) multimedia traffic. The proposed algorithm assigns a priority to each packet based on the extended EXP rule which tightly couples the PHY layer issue (instantaneous channel conditions) and MAC layer issue (QoS requirements) together. To reduce the computational complexity of a linear integer optimization problem, a suboptimal heuristic algorithm is proposed. Through systemlevel simulation, it is shown that the performance of the suboptimal algorithm is slightly different from the optimal algorithm, and both the optimal and suboptimal algorithms outperform the conventional OFDMA scheduling algorithms in terms of high spectral efficiency and better QoS satisfaction. It is also shown that OFDMA based scheduling algorithms outperform the OFDM/TDM based scheduling algorithm due to an added dimension of multiuser diversity in frequency domain in OFDMA systems.

Base stations are usually equipped with multiple transmit antennas. Hence, space-division multiple access in the form of linear beam-forming provides additional degrees of freedom for user scheduling. Regarding future work, the proposed algorithm could be extended to this more general setup, wherein the radio resource is partitioned in both time-frequency domain and space domain.

Appendix I: Flowchart of the Suboptimal Algorithm.

The diagram of the proposed suboptimal heuristic algorithm is shown in Figure 8.

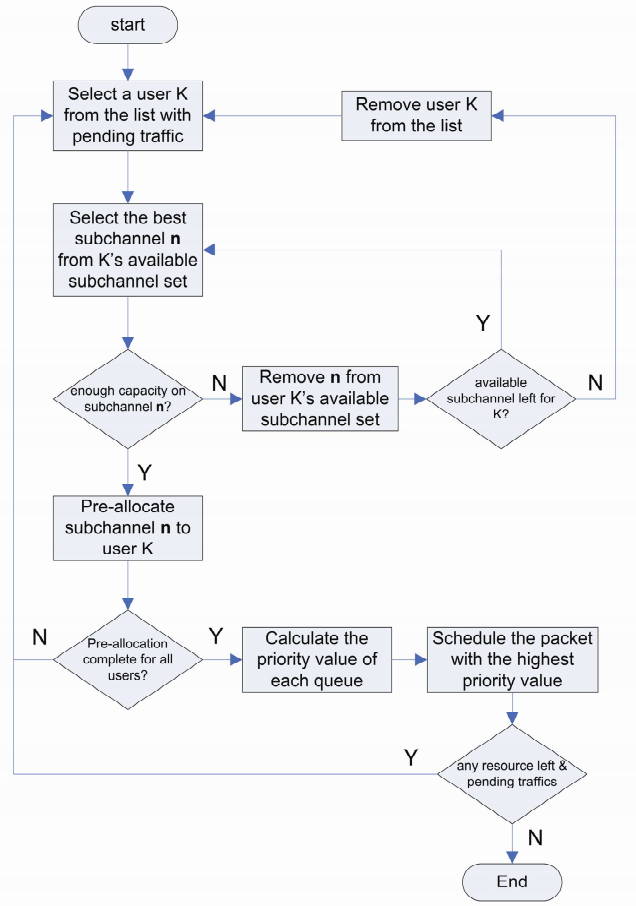

Figure 8. Flowchart of the proposed suboptimal scheduling algorithm.

7. References

[1] S. H. Ali, K. D. Lee, and V. C. M. Leung, "Dynamic resource allocation in OFDMA wireless metropolitan area networks," IEEE Wireless Communications, Vol. 14, No. 1, pp. 6–13, 2007.

[2] Y. J. Zhang and K. B. Letaief, "Energy-efficient MAC-PHY resource management with guaranteed QoS in wireless OFDM networks," ICC 2005, Vol. 5, pp. 3127–3131, May 2005.

[3] Y. J. Zhang and K. B. Letaief, "Adaptive resource allocation and scheduling for multiuser packet-based OFDM networks," ICC 2004, Vol. 5, pp. 2949–2953, June 2004.

[4] A. Todini, M. Moretti, A. Valletta, and A. Baiocchi, "A modular cross-layer scheduling and resource allocation architecture for OFDMA systems," GLOBECOM 2006.

[5] X. Zhang, E. Zhou, R. S. Zhu, S. M. Liu, and W. B. Wang, "Adaptive multiuser radio resource allocation for OFDMA systems," GLOBECOM 2005, Vol. 6, December 2005.

[6] X. Zhang and W. B. Wang, "Multiuser frequency-time domain radio resource allocation in downlink OFDM systems: Capacity analysis and scheduling methods," Computers and Electrical Engineering, Vol. 32, No. 1–3, pp. 118–134, 2006.

[7] S. S. Jeong, D. G. Jeong, and W. S. Jeon, "Cross-layer design of packet scheduling and resource allocation in OFDMA wireless multimedia networks," VTC 2006, Vol. 1, pp. 309–313, 2006.

[8] M. Bohge, J. Gross, M. Meyer, A. Wolisz, and T. U. Berlin, "Dynamic resource allocation in OFDM systems: an overview of cross-layer optimization principles and techniques," IEEE Network, Vol. 21, No. 1, pp. 53–59, 2007.

[9] S. Ryu, B. Ryu, H. Seo, and M. Shin, "Urgency and efficiency based packet scheduling algorithm for OFDMA wireless system," ICC 2005, Vol. 5, pp. 2779–2785, May 2005.

[10] C. F. Tsai, C. J. Chang, F. C. Ren, and C. M. Yen, "Adaptive radio resource allocation for downlink OFDMA/SDMA systems," ICC 2007, pp. 5683–5688, June 2007.

[11] K. F. Ahmed and M. F. Khaled, "Opportunistic scheduling of delay sensitive traffic in OFDMA-based wireless networks," WoWMoM 2006, Vol. 2006, pp. 279–288, June 2006.

[12] B. Rong, Y. Qian, and K. J. Lu, "Integrated downlink resource management for multiservice WiMAX networks," IEEE Transactions on Mobile Computing, Vol. 6, Issue. 6, pp. 621–632, 2007.

[13] Y. M. Ki, E. S. Kim, S. I. Woo, and D. K. Kim, "Downlink packet scheduling with minimum throughput guarantee in TDD-OFDMA cellular network," Lecture Notes in Computer Science, Vol. 3462, pp. 623–633, 2005.

[14] Y. M. Ki and D. K. Kim, "Packet scheduling algorithms for throughput fairness and coverage enhancement in TDD-OFDMA downlink network," IEICE - Transactions on Communications, Vol. E88-B, No. 11, pp. 4402–4405, 2005.

[15] S. Shakkottai and A. L. Stolyar, "Scheduling algorithms for a mixture of real-time and non-real-time data in HDR," Proceedings of International Teletraffic Congress (ITC), 2001.

[16] C. Hoymann, "Analysis and performance evaluation of the OFDM-based metropolitan area networks IEEE 802.16," Computer Networks, Vol. 49, No. 3, pp. 341–363, 2005.

[17] C. Cicconetti, L. Lenzini, E. Mingozzi, and C. Eklund, "Quality of service support in IEEE 802.16 networks," IEEE Network, Vol. 20, No. 2, pp. 50–55, 2006.

[18] D. H. Kim, H. R. Byung, and C. G. Kang, "Packet scheduling algorithm considering a minimum bit rate for non-real-time traffic in an OFDMA/FDD-based mobile internet access system," ETRI Journal, Vol. 26, No. 1, pp. 48–52, 2004

[19] H. Wang, "Priority-based resource allocation for RT and NRT traffics in OFDMA systems," The 3rd IEEE International Conference on Wireless Communications, Networking and Mobile Computing (IEEE WiCOM), Vol. 1, pp. 791–794, 2007.

[20] IEEE 802.16.3c-01/29r4, "Channel models for fixed wireless applications," IEEE 802.16 Broadband Wireless Access Working Group, July 2001.

[21] T. K. Sarkar, Z. Ji, K. Kim, A. Medouri, and M. Salazar-Palma, "A survey of various propagation models for mobile communication," IEEE Antennas and Propagation Magazine, Vol. 45, No. 3, pp. 51–82, 2003.

