Wireless Netw (2009) 15:831–844
DOI 10.1007/s11276-007-0077-y

An efficient delay constrained scheduling scheme for IEEE 802.16 networks

Fen Hou · Pin-Han Ho · Xuemin (Sherman) Shen

Published online: 3 November 2007

Abstract In this paper, we propose a simple yet efficient delay-constrained scheduling scheme for IEEE 802.16 networks. The proposed scheduling scheme not only can satisfy delay constraints of different service types by manipulating a simple operation parameter but also can achieve a good fairness performance. An analytical model is developed to evaluate the performance in terms of inter-service time, average queue length, and mean waiting time, and is verified through extensive simulations. Furthermore, an implementation procedure of the proposed scheme is given, which reflects the scheme's good features of the fast re-configurability and adaptability.

Keywords Scheduling · Proportional fairness · Preference metric · IEEE 802.16

1 Introduction

As a promising broadband wireless access standard, IEEE 802.16 [1, 2] has attracted extensive attentions from both the industry and academia. The features of easy deployment, wide coverage, and high data rate make IEEE 802.16 network a promising alternative to Digital Subscriber Line (DSL) and cable modem services to provide broadband wireless access in the suburban and rural areas for supporting real-time applications such as voice over Internet Protocol (VoIP), video conference, and Internet Protocol TV (IPTV). An IEEE 802.16 network is usually composed of a base station (BS) and multiple subscriber stations (SSs), as shown in Fig. 1. Each SS could be either a residential customer or an office building. IEEE 802.16 standard supports two modes: mesh mode and Point-to-Multi-Point (PMP) mode. In the mesh mode, SSs can communicate with each other directly, extending the coverage of an IEEE 802.16 network; in the PMP mode, each SS directly communicates with the corresponding BS through wireless links, and the BS is connected to a core network through a wired link. Compared to the mesh mode, the PMP mode can achieve better simplicity and easier deployment due to the centralized control, and is expected to serve as an important role in the Internet market of service-oriented wireless metropolitan area network (WMAN). In this paper, we focus on the PMP mode.

IEEE 802.16 standard defines four service types: Unsolicited Grant Service (UGS), Real-time Polling Service (rtPS), Non-real-time Polling Service (nrtPS), and Best Effort (BE). The UGS is designed to support real-time constant-bit-rate (CBR) applications such as VoIP, which has stringent constraint on delay and delay jitter; The rtPS is designed to support real-time variable–bit-rate applications such as IPTV, gaming, and video conferencing, where delay and minimum throughput requirements are addressed; The nrtPS is used to support non-real-time applications such as file transfer protocol (FTP), where minimum achievable throughput is defined; and BE, the lowest priority service, is subject to no quality of service (QoS) requirement.

F. Hou (✉) · P.-H. Ho · X. (Sherman) Shen
Department of Electrical and Computer Engineering, University of Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
e-mail: fhou@bbcr.uwaterloo.ca

P.-H. Ho
e-mail: pinhan@bbcr.uwaterloo.ca

X. (Sherman) Shen
e-mail: xshen@bbcr.uwaterloo.ca

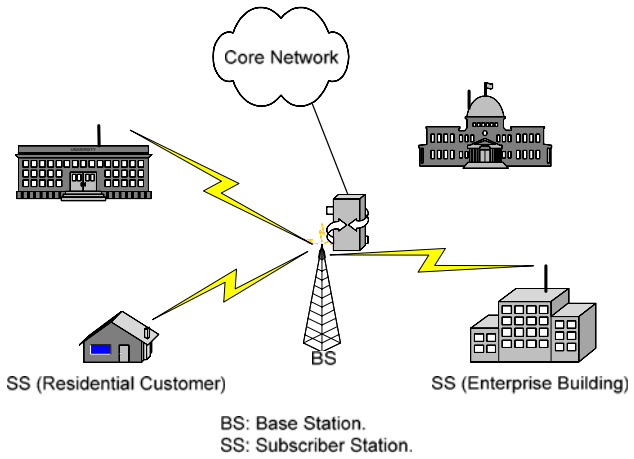


Fig. 1 An IEEE 802.16 network with PMP mode

Scheduling plays a key role in providing service differentiation and QoS satisfaction for the four types of services [3, 4], particularly when the system is heavily loaded possibly due to bursty arrival of traffic and/or deep fading of wireless channels. Some simple scheduling schemes, such as Round-Robin and strict priority scheme, may not be able to efficiently satisfy delay requirements of different service types. On the other hand, a precise and complicated model that considers the delay constraint of each service type may require too much computation time in a dynamic implementation, which leads to a scalability problem in a system with frequently changing traffic and channel conditions. Thus, it is a challenging issue to design an efficient and simple QoS scheduling scheme that not only can satisfy delay constraints of different service types but also has the features of adaptability and easy implementation.

In this paper, we propose a novel scheduling scheme, which can provide delay constraint satisfaction and service differentiation. In addition, the proposed scheduling scheme is easy to implement. The main contributions of this paper are in threefold. Firstly, the proposed scheduling scheme can satisfy the delay of each service type by manipulating a simple parameter. Instead of using a different preference metric for each service type or a complicated optimization process for obtaining priority of each queue, we define a uniform preference metric with a simple parameter for each queue in the system. Secondly, the proposed scheme achieves perfect fairness in terms of satisfying the bandwidth requirement of each queue. Thirdly, an analytical model is developed to evaluate the inter-service time, mean waiting time, and average queue length, and a detailed implementation procedure is given. Simulation results demonstrate that the proposed scheduling scheme can meet our design requirements, which also verify the accuracy of the analytical model.

The remainder of the paper is organized as follows. Section 2 gives a survey on the related work. Section 3 describes the proposed scheduling scheme. Section 4 presents the analysis of some important performance metrics. The implementation of the proposed scheme is given in Sect. 5. Extensive simulations are conducted to verify the analytical model and demonstrate the efficiency of the proposed scheme in Sect. 6. Section 8 concludes the paper.

2 Related works

In general, there are four main design requirements for a scheduling scheme: easy-implementation, the ability of service differentiation, a high system throughput, and a good fairness performance. Round-Robin is known as the most straightforward scheduling scheme, which is focused on the bandwidth allocation, rather than delay constraints. Strict priority (SP) scheme was introduced in [5], where UGS queue was assigned with the highest priority, followed by rtPS queue, nrPS queue, and BE queue. SP is obviously subject to a number of challenges in the aspect of fairness. Similar approaches were proposed to give a relatively higher priority to the queues with stringent delay and throughput constraints, in which the system reserves relatively more resource to these queues. With such a design principle, the study in [6] evaluated the average delay and packet loss rate through extensive simulations. The problem of the scheme lies in the difficulty to determine the proportion of system capacity allocated to each queue when the network environments are dynamically changed. Opportunistic scheduling (OS) [7] aims at maximizing the system throughput by taking the best advantage of multi-user channel diversity, in which the resource is assigned to the queue with the best channel condition at a given time slot. However, OS may result in poor fairness due to the starvation of end users experiencing poor channel conditions for a long time. An opportunistic fair scheduling scheme was proposed in [8] to maintain the long-term fairness and achieve a smooth service rate.

A credit-based code division generalized process sharing (C-CDGPS) scheme was proposed in [9], where the scheduler assigns the resource based on both the GPS discipline and each user's credit to achieve high resource utilization as well as the long-term fairness. A cumulative distribution function based scheduling scheme was proposed in [10], where the probability of selecting a user to be served depended only on the distribution of its own channel condition so that the performance of fairness can be improved. Another well-known scheduling scheme to initiate a graceful tradeoff between the system throughput and fairness is proportional fairness (PF) scheduling

[11, 12]. It can achieve a good fairness performance since the long-term fractions of overall system resource obtained by all queues are identical. In addition, it takes advantage of multi-user channel diversity to obtain a high system throughput. However, it is generally difficult to conduct a quantitative analysis. In other words, important performance metrics such as delay and queue length cannot be analytically evaluated in advance. In [13], a modified proportional fairness scheduling was developed, where the long-term throughput of each queue was replaced with its average channel condition. However, the scheme cannot deal with the issues of service differentiation and QoS satisfaction. The study in [14] defined *delay satisfaction indicator* and *throughput satisfaction indicator* in the design of the preference metric for real-time and non-real-time queues, respectively. With the help of several additional parameters in the preference metric, the scheme can achieve satisfied performance in a given network condition. However, the study does not explore the relationship between throughput and delay for a specific queue. It is also difficult to adaptively configure/reconfigure those custom-designed parameters for each service type.

Largest Weighted Delay First (LWDF) [15] is considered as a more advanced and well known scheme in QoS scheduling, where the head-of-line delay of each real-time queue is manipulated in the preference function. Two modified LWDF schemes were proposed in [16, 17]. In each time slot, queue j is served so as to maximize the term $\gamma_j W_j r_j$, where γ_j is an arbitrary positive constant, W_j is the head-of-line packet delay, and r_j is the channel capacity for queue j. However, none of them can analytically determine an appropriate value for γ_j. The scheme proposed in [18] made use of exponential rule for delay-constrained traffic and proportional fair scheduling for the BE traffic.

3 Delay-constrained scheduling scheme

The proposed delay-constrained scheduling scheme deals with the drawback of the PF scheme by considering the waiting time experienced by each queue in the design of the preference metric. In the following, we first remind the PF scheme, followed by the proposed scheduling scheme.

3.1 Proportional fairness and its drawback

PF [11, 12] scheme takes the ratio between the channel capacity (denoted as $r_i(t)$) and the long-term throughput of queue i (denoted as $R_i(t)$) as its preference metric:

$$C_i(t) = \frac{r_i(t)}{R_i(t)} \tag{1}$$

where $R_i(t)$ is calculated using exponential moving average in terms of T_i,

$$\begin{aligned} R_i(t) &= \left(1 - \frac{1}{T_i}\right) R_i(t-1) + \frac{r_i(t)}{T_i} \\ &= \begin{cases} \left(1 - \frac{1}{T_i}\right)^{n_i(t)} R_i(t - n_i(t)) \\ \forall t | n_i(t) \neq 0 \\ \left(1 - \frac{1}{T_i}\right)^{n_i(t-1)} R_i(t - n_i(t-1)) + \frac{r_i(t)}{T_i} \\ \forall t | n_i(t-1) \neq 0,\ n_i(t) = 0 \end{cases} \end{aligned} \tag{2}$$

where $r_i(t)$ is the channel capacity for queue i at time t, T_i decides the contribution of $r_i(t)$ to $R_i(t)$, which is taken as the number of time duration from the very beginning to the current timeslot, and $n_i(t)$ is the number of time slots that have elapsed up to time t since the queue was previously visited. With the preference metric in (1), the scheduler will visit the queue with the largest value of the preference metric in the next time slot. The PF scheme is designed to balance the system throughput and fairness performance, but it does not consider the delay requirement of each queue. In the following, we propose a delay-constrained scheduling scheme by considering the delay experienced by each queue from its previous service such that service differentiation and delay requirements can be handled.

3.2 Delay-constrained scheduling scheme

It is identified that the selection of T_i in (2) addresses a fundamental impact on the delay performance of each queue under the preference metric, where two queues with different values of T_i may experience different average delay in spite of the same perceived throughputs. Furthermore, the queue with a smaller T_i will have its preference metric fluctuate more seriously, which makes the queue to be served more frequently. Such a characteristic is critically desired for the queues with stringent delay constraints. Therefore, T_i can be manipulated for differentiating the delay behavior of each queue, which serves as the basic operational principle in the proposed delay-constraint scheduling scheme. In other words, instead of making T_i as large as possible, the proposed scheme sets a specific value of T_i for each queue in order to satisfy the delay requirement for each queue.

In addition to manipulating parameter T_i, one more parameter $R_i^{\exp}(t)$ is added into (1) to reflect the fact that

the queues are subject to connection requests with a fixed bandwidth demand instead of saturated traffic, i.e.,

$$C_i(t) = \frac{w_i(t)}{R_i(t)/R_i^{\exp}(t)} \tag{3}$$

where $w_i(t)$ is the channel condition weighting on queue i at time t, $R_i^{\exp}(t)$ is the expected throughput for queue i at time t, which is constant as long as the number of connections admitted to the queue is not changed. Equation 3 can be rewritten as (4) to capture the behavior of preference metric as the queue is waiting for service:

$$C_i(t) = \frac{w_i(t)}{(R_i^s(t)/R_i^{\exp}(t))(1-1/T_i)^{n_i(t)}} \tag{4}$$

where $n_i(t)$ is the number of time slots that have elapsed up to time t since the queue was previously visited, $R_i^s(t)$denotes the long-term throughput of queue i at the instant when the queue was previously visited at time t, i.e., $R_i(t) = R_i^s(t)(1-1/T_i)^{n_i(t)}$.

For a stable system, the long-term throughput of queue i must match the expected throughput of the queue. By approximately taking term $(R_i^s(t)/R_i^{\exp}(t))$ identical for all queues, we can further abstract the preference metric as

$$C_i(t) = \frac{w_i(t)}{(1-1/T_i)^{n_i(t)}} \tag{5}$$

Based on the IEEE 802.16d standard, the channel condition of each SS is dominated by its geographic environment and its distance from the BS. Therefore, for reducing the complexity, $w_i(t)$ in (5) is replaced with the average channel condition of the SS to which the queue i belongs, denoted as $\bar{w}_i$. Consequently, the preference metric is expressed as

$$C_i(t) = \frac{\bar{w}_i}{(1-1/T_i)^{n_i(t)}} \tag{6}$$

Thus, the proposed scheduling scheme is to rank each queue by a unified preference metric shown in (6), where T_i is the parameter for manipulation in order to achieve service differentiation, and $n_i(t)$ is the number of waiting timeslots up to time t since the queue was previously visited. Note that the value of the parameter T_i is larger than 1 and without unit. The proposed scheduling scheme can provide the delay satisfaction for each queue by selecting a specific set of T_i.

4 Performance analysis

The performance of the proposed scheduling scheme is analyzed in terms of three important performance measures: inter-service time N_i (denoted as N_i), average queue length (denoted as $E[L]$), and mean waiting time (denoted as $E[W]$). The fairness issue is also investigated.

4.1 Analysis of inter-service time

When a queue is waiting for being served, the value of its preference metric $C_i(t)$ increases for every time slot, which is similar to *charging* process of a capacitor. After the queue is served, $C_i(t)$ will sharply decrease during the next time slot, which is similar to a *discharging* process. Thus, the value of preference metric of each queue upon being serviced (or termed the *winning metric* of the queue) is expected to be identical in the stationary state. That is, the following relationship is held for the proposed scheduling scheme:

$$\frac{\bar{w}_1}{(1-1/T_1)^{N_1}} \approx \cdots \approx \frac{\bar{w}_i}{(1-1/T_i)^{N_i}} \approx \cdots \approx \frac{\bar{w}_k}{(1-1/T_K)^{N_K}} \tag{7}$$

where N_i is the inter-service time of queue i.

With the proposed scheme, whether or not a queue wins the chance of service depends on all previous preference metric values of all queues under consideration. Due to the dependence of all other queues, to mathematically verify the relationship given by (7) is too complicate to be tractable. Therefore, we resort to the way of conducting extensive simulations. This relationship is verified by extensive simulation results under a variety of parameters, which are shown in Fig. 2.

Figure 2(a)–(c) illustrates the mean and standard deviation of the value of each queue's preference metric when it is selected to be served given different T_i and an identical weighting $\bar{w}_i$ $(i = 1,2,\ldots,K)$, where the total number of queues is 40, 100, and 200, respectively. Figure 2(d)–(f) is the same as Fig. 2(a)–(c) except for different channel weighting $\bar{w}_i$ for each queue. From Fig. 2(a)–(f), it is observed that the means of the winning metric of all queues are almost identical under each simulation scenario, and the standard deviation is generally less than 3% of the mean and can be negligible.

By taking the approximation, the relationship between the N_i, T_i, and $\bar{w}_i$ of each queue can be reformulated as

$$\frac{(1-1/T_i)^{N_i}}{(1-1/T_1)^{N_1}} = \frac{\bar{w}_i}{\bar{w}_1}, \quad i = 1,2,\ldots,K \tag{8}$$

Thus, the relation between N_i and N_1 is given by:

$$N_i = \frac{\log(\bar{w}_i/\bar{w}_1) + N_1\log(1-1/T_1)}{\log(1-1/T_i)}, \quad i = 1,2,\ldots,K \tag{9}$$

The relation between T_i and T_1 is given by

$$T_i = \frac{1}{1-(\bar{w}_i/\bar{w}_1)^{1/N_i}(1-1/T_1)^{N_1/N_i}}, \quad i = 1,2,\ldots,K \tag{10}$$

In addition, the summation of the probabilities of all queues to be served at a given time slot is 1, which is expressed as

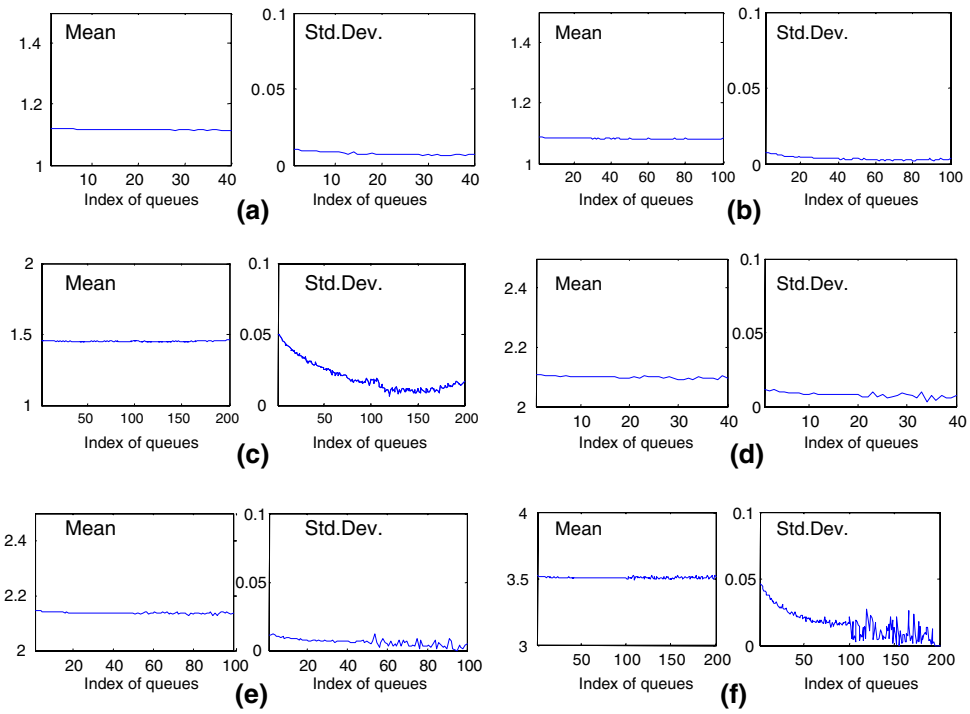


Fig. 2 The mean and standard deviation of preference metric value of each queue

$$\sum_{i=1}^{K} \frac{1}{N_i + 1} = 1 \tag{11}$$

where $1/(N_i + 1)$ denotes the probability that queue i is chosen for service at a given time slot.

Based on (9) and (11), N_1 can be derived as

$$\frac{1}{N_1} + \sum_{i=2}^{K} \frac{\log(1 - 1/T_i)}{\log(\bar{w}_i/\bar{w}_1) + N_1 \log(1 - 1/T_1)} = 1 \tag{12}$$

Therefore, the inter-service time for queue i can be derived based on (9), (11), and (12).

4.2 Analysis of average queue length and waiting time

The adaptive modulation and coding (AMC) is adopted in IEEE 802.16 networks to provide a better tradeoff between system capacity and coverage. The modulation level adopted at each SS mainly depends on its geographic limitation and distance from the BS. As shown in Fig. 3, the SSs near the BS usually use higher modulation level than that of SSs far from the BS. Therefore, in an IEEE 802.16 network, each SS uses a specific modulation level, and different SSs may have different modulation levels. Assume that the packet arrival of each queue follows Poisson process, and observe the number of packets in queue i at the time slot the queue is visited. Let the random variable $X_n \in \{0,1,2,\ldots\}$ denote the number of packet in queue i just before the queue is visited at the n-th time. The process $\{X_n: n = 1,2,3,\ldots\}$ forms a Markov chain on the state space $\{0,1,2,\ldots\}$ [19]. The evaluation of the X_n is given by

$$X_{n+1} = \begin{cases} A_{n+1} & X_n = 0, 1, \ldots, v_i \\ X_n - v_i + A_{n+1} & X_n > v_i \end{cases} \tag{13}$$

where A_{n+1} represents the number of arrival packets during a visited time slot plus a inter-service time of queue i. v_i is the maximum number of packets delivered in a timeslot for queue i. It is relative to the modulation level used by the SS to which queue i belongs. The one-step transition probability matrix of the Markov model is given by

$$\underline{\underline{P}} = \begin{bmatrix} a_0 & a_1 & a_2 & a_3 & \cdots & a_{L_i} \\ a_0 & a_1 & a_2 & a_3 & \cdots & a_{L_i} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ a_0 & a_1 & a_2 & a_3 & \cdots & a_{L_i} \\ 0 & a_0 & a_1 & a_2 & \cdots & a_{L_i-1} \\ 0 & 0 & a_0 & a_1 & \cdots & a_{L_i-2} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \end{bmatrix} \begin{matrix} \leftarrow \text{The 1st row} \\ \leftarrow \text{The 2nd row} \\ \\ \leftarrow \text{The } v_i\text{th row} \\ \\ \\ \leftarrow \text{The } (L_i+1)^{\text{th}} \text{ row} \end{matrix}$$

$$a_k = \Pr[A = k], \quad k = 1, 2, 3, \ldots, L_i$$

where a_k is the probability that k packets arrive at the ith queue during a visited time plus a inter-service time, and L_i is the buffer capacity of queue i in unit of packets.

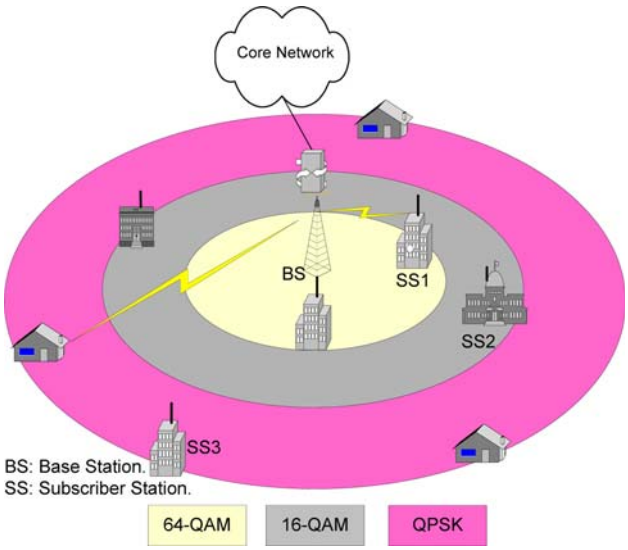


Fig. 3 Adaptive modulation in an IEEE 802.16 network

Let $\underline{\Pi} = (\pi_1, \pi_2, \pi_3, \ldots, \pi_{L_i})$ denote the steady-state probability vector, which can be derived from the following balance equations

$$\begin{cases} \underline{\Pi} = \underline{\Pi}\,\underline{P} \\ \sum_{j=0}^{L_i} \pi_j = 1 \end{cases} \tag{14}$$

Let Z denote the number of packet observed by a packet when it finishes its service and departs the queue. From the steady-state probability, the distribution of the random variable Z is given by

$$p_n = \frac{1}{1-\pi_0}\left\{\sum_{x=1}^{v_i} \Pr[Z = n, X = x] + \Pr[Z = n, X > v_i]\right\},$$
$$n = 0, 1, \ldots, L_i \tag{15}$$

The average queue length and the mean waiting time are given by

$$E[L] = \sum_{n=0}^{L_i} n p_n \tag{16}$$

$$E[W] = \frac{E[L]}{\lambda_i} \tag{17}$$

where λ_i is the arrival rate of queue i.

4.3 Fairness

In IEEE 802.16 networks, system resource is allocated to different service types based on their demands and priorities. Therefore, fairness is a metric to measure how closely the allocation is to the demand of each service type. The *Jain fairness* index [20, 21] is a commonly adopted fairness index to measure the fairness among all the queues under consideration, which is defined as

$$I = \frac{\left|\sum_{n=1}^{K} x_k\right|^2}{K\sum_{k=1}^{K}(x_k)^2} \tag{18}$$

where K is the total number of service types, and x_k is defined as

$$x_i = \begin{cases} \frac{a_i}{d_i} & \text{if } a_i < d_i \\ 1 & \text{Otherwise} \end{cases} \tag{19}$$

where d_i is the resource demanded by service type i, and a_i is the resource allocated to service type i.

The proposed scheduling scheme can satisfy the inter-service time requirement for each service type by setting a specific T_i. Since the inter-service time determines the obtained resource, satisfying the inter-service requirement equivalently means satisfying the resource requirement of each service type. Therefore, it is concluded that the proposed scheduling scheme can achieve satisfying fairness performance, which is verified by the simulation given in Sect. 6.

5 Implementation

Without loss of generality, we assume the number of SSs is M, and there are four logic queues for each SS, which correspond to four service types, respectively, as shown in Fig. 4. Therefore, the total number of queues (denoted as K) equals $4M$. Arrival packets are buffered at the corresponding queue first based on their destination SS and service type. The traffic arrival rate of a queue depends on the number of connections admitted to the queue and the arrival rate of each admitted connection. Each queue is given a specific delay constraint denoted as D_i. A network state vector $\underline{\Omega}(t) = [\Omega_1(t), \ldots, \Omega_K(t)]$ is maintained at the scheduler, where $\Omega_i(t) = (\lambda_i(t), \bar{w}_i)$, and $\lambda_i(t)$ and $\bar{w}_i$ are the traffic arrival rate and the channel weighting at queue i, respectively. A preference metric vector $\underline{C}(t) = [C_1(t), C_2(t), \ldots, C_K(t)]$ is also maintained at the scheduler. Vector $\underline{C}(t)$ is updated at the beginning of each frame, based on (6) and the network state vector $\underline{\Omega}(t)$. According to the updated $\underline{C}(t)$, the scheduler at the BS assigns each time slot to a queue with the largest preference metric value at that time slot, and allocates all bandwidth to it during that time slot.

From the perspectives of system designers and operators, two most important issues are: (1) how does the system select the operation parameter T_i such that the given delay requirement can be satisfied? (2) How often or in

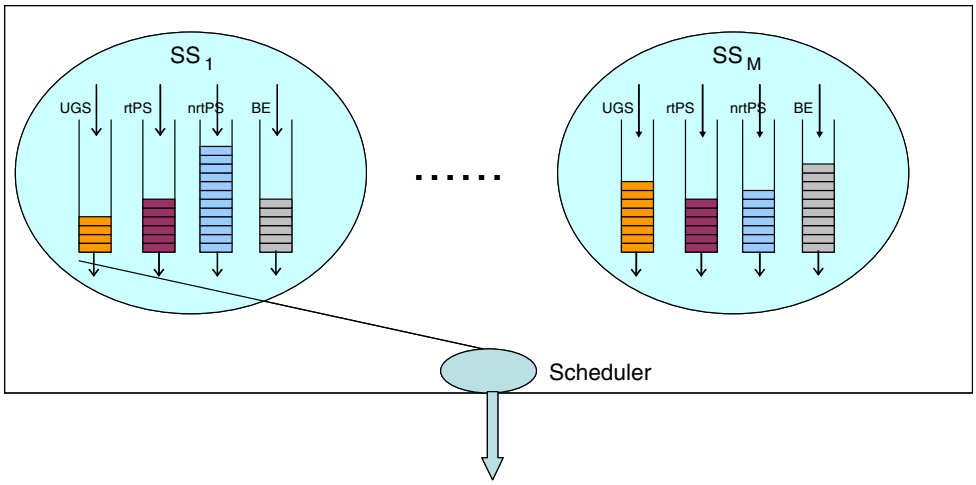


Fig. 4 An illustration of the scheduler at BS

what circumstance should the operation parameter T_i be reconfigured? Since the network state may change with time, the operation parameter T_i should be adjusted accordingly in order to keep the average delay of each queue constrained. A generic reconfiguration process for T_i is illustrated in Fig. 5. The input parameters of each reconfiguration process are $\underline{\Omega}(t)$ and D_i for $i = 1,\ldots,K$, while the outcome is the operation parameter T_i for each queue.

More specific reconfiguration process is illustrated in Fig. 6. When any network state variation is sensed and if the ratio of the change is larger than a threshold ρ_{th}, a reconfiguration process is activated. If the change of the network state is caused by fluctuation of wireless channel conditions, a new operation parameter T_i is obtained based on the relation between T_i and weighting $\bar{w}_i$ which is given by (10). If the change of the network state is caused by the fluctuation of arrival rate of some queue, the derivation of a new parameter T_i is divided into two steps. The first step is to obtain the corresponding maximum inter-service time requirement on the given delay requirement and arrival rate of each queue. After having the new inter-service time requirement, a set of new operation parameter T_i can be obtained based on the relation between N_i and T_i which is given by (10).

In summary, given the arrival rate, channel condition weighting of each SS, and a set of delay constraints for all queues, the value of operation parameter T_i ($i = 1,2,3,\ldots,K$)

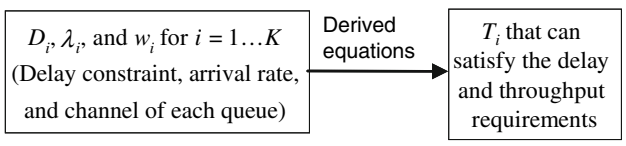


Fig. 5 The reconfiguration process of T_i

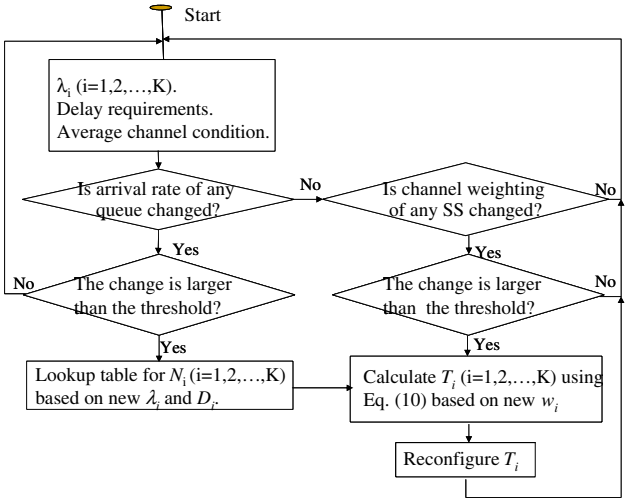


Fig. 6 The flowchart of the reconfiguring of T_i

can be obtained in a simple yet efficient manner. Dynamic reconfiguration on the key system parameter, T_i, can be performed efficiently through a set of close-form relations, including an approximate mapping from the delay constraint to the inter-service time for each queue. In addition, since the proposed scheduling is deployed only by comparing a simple and unified preference metric of each queue, the time complexity is at the order of $O(NlogN)$. Therefore, the proposed scheduling scheme has low computing complexity and can be deployed easily.

6 Simulation results

In the simulation, the number of SSs is 4, and the BS handles the downlink scheduling of each SS with four logic queues corresponding to the UGS, rtPS, nrtPS, and BE services, respectively. Therefore, the total number of

queues under consideration at the BS is 16, which are indexed from 1 to 16. Queues 1, 2, 3, and 4 correspond to the UGS, rtPS, nrtPS, and BE queue at SS1, respectively; queues 5, 6, 7 and 8 correspond to the UGS, rtPS, nrtPS, and BE queue at SS2, respectively; and so on. The buffer size of each queue is 30 (packets). Unless otherwise noted, the value of T_i for UGS, rtPS, nrtPS, and BE queue at each SS are 100, 120, 140, and 160, respectively. In order to obtain a better granularity, a time slot is defined as the duration it takes for a packet to transmit when Binary Phase Shift Keying (BPSK) modulation is used. Therefore, for the SSs with a better channel condition and higher modulation level such as 16-state Quadrate Amplitude Modulation (16QAM) or 64QAM, the maximum number of packets transmitted in a time slot is larger than one.

Four scenarios are examined: The first is to observe the capability of service differentiation and to verify the analytical model in terms of inter-service time; the second scenario is to test the capability of providing delay guarantee for each queue given traffic arrival rates and channel conditions; the third scenario is to evaluate the fairness; the fourth scenario is to illustrate the adaptability to variable network situations, where any change on the traffic arrival and channel weighting are taken as a network event that drives the reconfiguration of the system operation parameter T_i.

6.1 Scenarios I: service differentiation capability

Figures 7 and 9 show the inter-service time of each queue (denoted as N_i) given identical channel weighting $\bar{w}_i$ for queues 1–16 with 4 SSs and 20 SSs, respectively. It is observed that the inter-service time increases with the increase of T_i for the same SS. UGS queues have the smallest inter-service time due to the smallest value of T_i while BE queues have the largest inter-service time. For the queues of the same service type, the inter-service time with 20 SSs is larger than that with 4 SSs. With a larger number of SSs, the resource allocated to each queue decreases accordingly, leading to a larger inter-service time.

Figure 8 shows the relationship between the inter-service time and parameter T_i for the rtPS queue at SS1 given that the parameters of all the other queues are the same as that for Fig. 7. It is observed that the inter-service time N_2 increases with the increase of T_2. From Figs. 7–9, it is concluded that the inter-service time of each queue is relatively sensitive to the parameter T_i. This fact makes T_i a very effective operation parameter that can be manipulated to achieve service differentiation for each queue.

The comparison of the proposed scheme with PF in terms of throughput is shown in Fig. 10. It is observed that for the proposed scheme, the queues with different T_i achieve different throughputs. With a smaller value of T_i, queues 1, 5, 9, and 13 obtain more service chances. Consequently, they achieve higher throughputs. On the contrary, the PF can not provide the service differentiation. All queues achieve similar throughputs.

Figure 11 illustrates the effect of different $\bar{w}_i$ on the inter-service time. Let $\bar{w}_i$ for SS1 to SS4 be set as the average signal noise ratio of each SS, which are 20, 15, 10, and 5 for SS1–SS4, respectively. Figure 11 shows the inter-service time of each queue. It is observed that a queue taking a smaller T_i yields a smaller inter-service time. In other words, a queue with a smaller channel weighting has to take a smaller T_i in order to achieve the same inter-service time as the queues of larger channel weightings. The perfect match between the analytical and simulation results further validates the assumption made in (7).

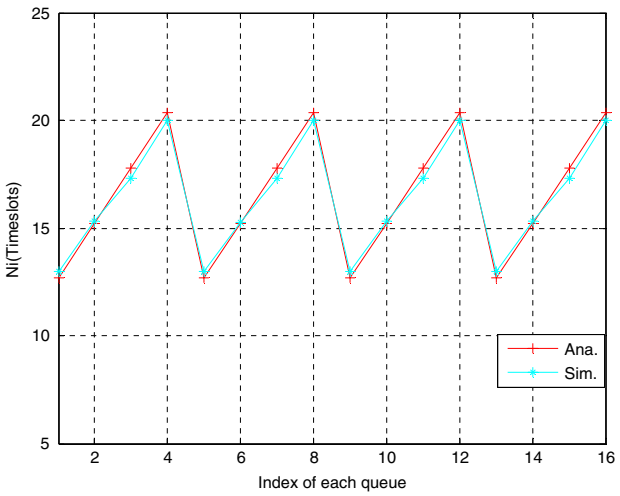


Fig. 7 The inter-service time of each queue with 4 SSs

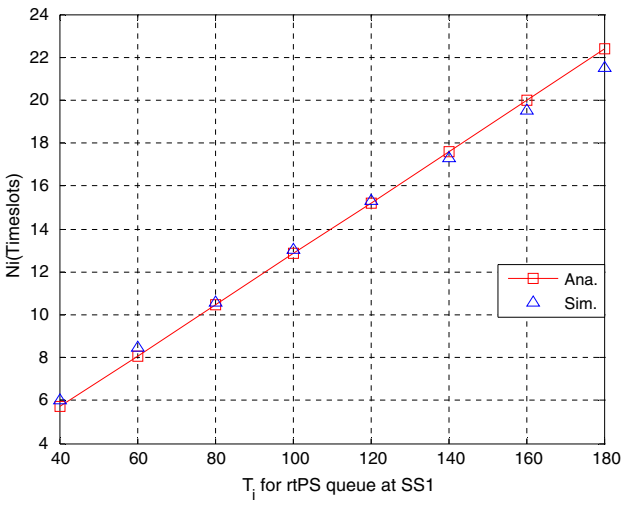


Fig. 8 The inter-service time versus T_i for rtPS queue at SS1

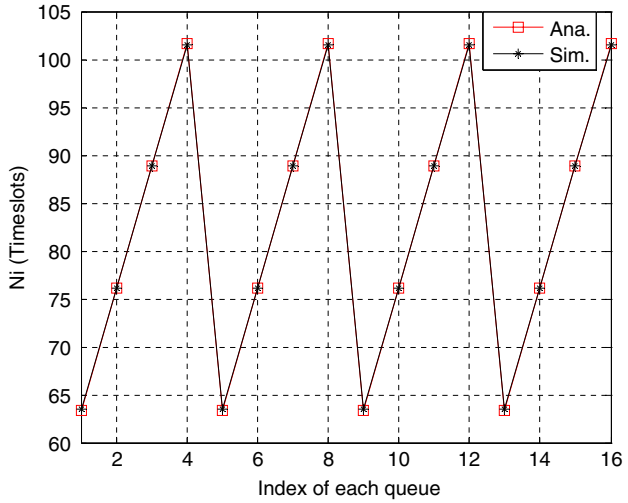


Fig. 9 The inter-service time for queues 1–20 with 20 SSs

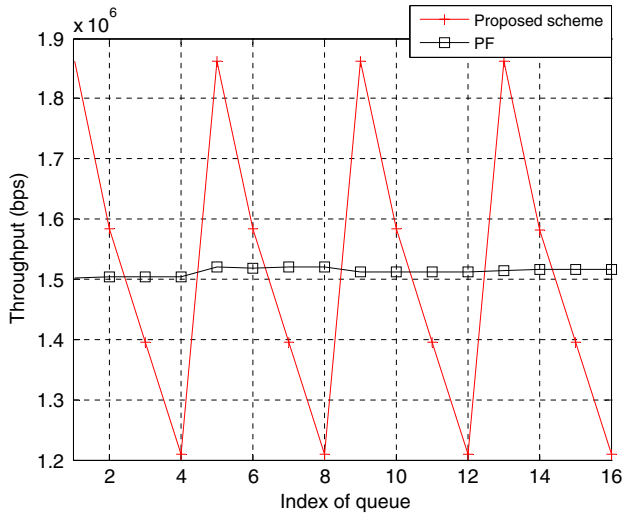


Fig. 10 Throughput of each queue for two schemes

6.2 Scenario II: delay constraint versus T_i

When the delay constraints for each queue, D_i, is given, the mapping from D_i to the corresponding inter-service time requirement N_i can be derived using the analytical results discussed in Sect. 4. The following case studies simply assume that such a mapping $D_i \rightarrow N_i$ is available.

6.2.1 With a uniform channel weighting

Let all SSs be with a uniform channel weighting, and the delay constraints for the UGS, rtPS, and nrtPS queues are 4, 6, and 12 timeslots, respectively. The arrival rate is 0.4 packets per timeslot. By using the mapping from D_i to N_i, the delay constraints are transferred to the inter-service time requirement, which is 8.5339, 11.7337, and 16.0127 timeslots for the UGS, rtPS, and nrtPS queue of each SS, respectively. The BE queues are allocated with the left-over resources.

Based on (11), the inter-service time for the BE queues is 127.2094 timeslots. When we have the inter-service time of each queue and let T_1 be 50, we can obtain the value of operation parameter T_i for the rtPS, nrtPS, and BE queues as 68.5590, 93.3777, and 738.3397, respectively, using (10). With the analytically derived parameters T_i, Table 1 gives the comparison of inter-service time requirements and simulation results for the UGS, rtPS, and nrtPS queue at each SS. It is observed that the requirement of inter-service time of each queue is very well satisfied by employing the analytically derived operation parameter T_i. Figure 12 shows the average delay. It is observed that the three types of queues are differentiated in terms of average delay, where the UGS queues are the most prioritized while the nrtPS queues are the least. The results coincide with the objectives defined in the IEEE 802.16 Standard.

6.2.2 With a non-uniform channel weighting

The channel weightings of SS1 to SS4 are set as their average SNR, which are 20, 15, 10, and 5, respectively. With different modulation schemes, the maximum numbers of packets delivered within a timeslot for queues belonging to SS1–SS4 are taken as 8, 4, 2, and 1, respectively. The inter-service time requirements for the UGS, rtPS, and nrtPS queues at each SS are taken as 10, 14, and 18 timeslots, respectively. The remaining resources are allocated to the BE queues. Let T_1 be taken as 50. Note that the above parameters can be arbitrarily assigned without influencing the result. With the given inter-service

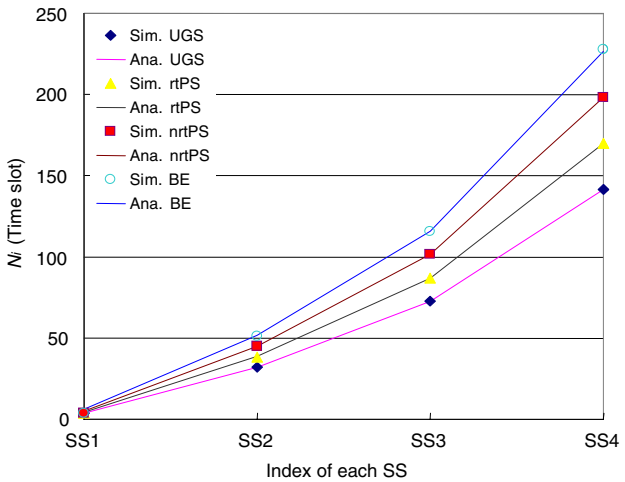


Fig. 11 The effect of different weightings of SSs on the inter-service time at each queue

Table 1 The comparison of inter-service time requirements N_i and simulation results for UGS, rtPS, and nrtPS queue of each SS based on the analytically derived T_i of each queue

Service type	UGS	rtPS	nrtPS
N_i	8.5339	11.7337	16.0127
Simulation Results (at SS1)	8.6789	11.5704	15.9762
Simulation Results (at SS2)	8.6789	11.5704	15.9762
Simulation Results (at SS3)	8.6802	11.5668	15.9797
Simulation Results (at SS4)	8.6809	11.5704	15.9797

requirements and channel weighting of each queue, the value of operation parameter T_i for all the queues are obtained analytically and shown in Table 2.

With the obtained parameters T_i, simulation is conducted to further examine the ability of satisfying the inter-service time requirements, which is shown in Table 3. It is observed that the simulation results match very well with the analytical results at each queue for meeting the inter-service time constraint. Figure 13 shows the average delay for the UGS, rtPS, and nrtPS queue, respectively.

6.2.3 *Comparison to other schemes*

To evaluate the efficiency of the proposed scheme, two counterpart schemes are adopted for comparison. One is the strict priority (denoted as SP) where the order of priority is UGS, followed by rtPS and nrtPS, and the other is Largest Weighted Delay First (denoted as LWDF), where the parameter α is 0.023. Figure 14 shows the delay performance of three service types for the proposed scheme and two counterpart schemes. It is observed that the proposed scheme has a better performance in terms of satisfying the delay constraint and providing delay differentiation among three service types. SP provides low delay for UGS, but nrtPS is starved for a long time. LWDF provides almost same delay for three service types.

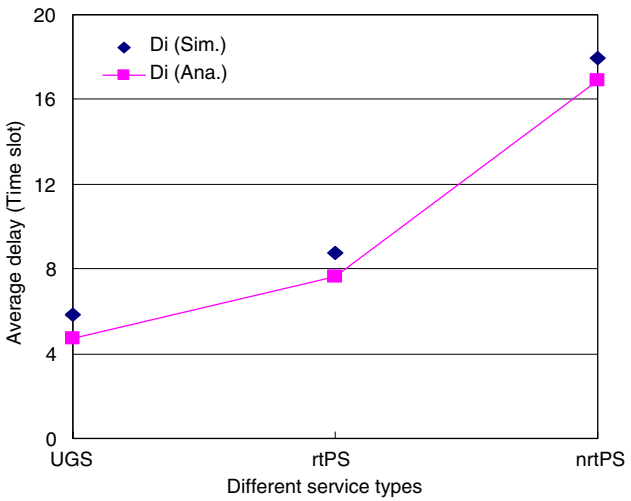


Fig. 12 Average delay for different service types with a uniform channel weighting

In summary, we have demonstrated that the proposed scheduling scheme can function well in both cases with a uniform and non-uniform channel weighting for each SS. It can correctly configure the operation parameter T_i such that the inter-service time requirements and delay constraints of all queues can be satisfied. In addition, the comparison of the proposed scheme to two counterparts demonstrates the good performance of the proposed scheme in terms of providing service differentiation and satisfying different delay constraints.

6.3 Scenario III: fairness

Based on the Jain fairness index, simulation is conducted to evaluate the fairness of the proposed scheme, in which the inter-service time requirements for rtPS and nrtPS queues are 15 and 30 timeslots, respectively, while the inter-service time for UGS queues is changing from 6 to 15 timeslots. The BE queues take the remaining resource. Figure 15 shows the result in comparison with that for the round robin and strict priority schemes. It is observed that the proposed scheme has a perfect fairness performance while the fairness of strict priority scheme is the worst.

6.4 Scenario IV: impact by network dynamics

Figure 16(a)–(d) shows the relation between the inter-service time and average delay with different arrival rates. In the scenario of simulation for Fig. 16(a)–(d), the maximum number of packets delivered in a timeslot is 8, 4, 2, and 1, respectively, which reflects a different modulation scheme possibly taken at each SS.

It is observed that the more stringent the delay requirement is, the smaller value of inter-service time is needed for the queue. Based on the relationship of average delay and inter-service time, the mapping from the given

Table 2 The derived operation parameters T_i of each queue

SS	SS1			SS2		
Queue	UGS	rtPS	nrtPS	UGS	RtPS	nrtPS
T_i	50	69.8	89.6	20.9	29.1	37.3
SS	SS3			SS4		
Queue	UGS	rtPS	nrtPS	UGS	RtPS	nrtPS
T_i	11.7	16.1	20.6	6.8	9.3	11.8

Table 3 The comparison of inter-service time requirements N_i and simulation results for GUS, rtPS, and nrtPS of each SS based on analytically derived T_i of each queue

Service type	UGS	rtPS	nrtPS
N_i	10	14	18
Simulation Results (at SS1)	9.9344	13.8220	18.9390
Simulation Results (at SS2)	10.3568	14.1976	18.3837
Simulation Results (at SS3)	10.3545	14.1976	18.3798
Simulation Results (at SS4)	10.3560	14.2226	18.3876

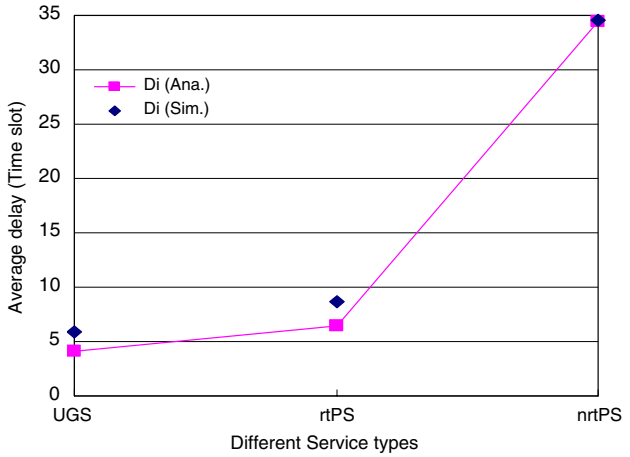


Fig. 13 Average delay for different service types with different channel weightings

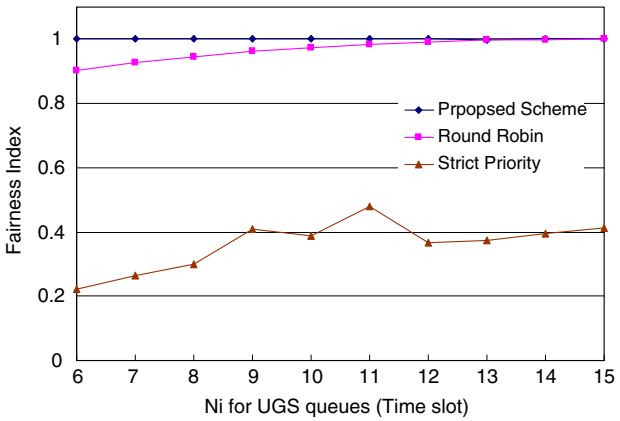


Fig. 15 The Jain fairness index with different inter-service time requirements on UGS queues

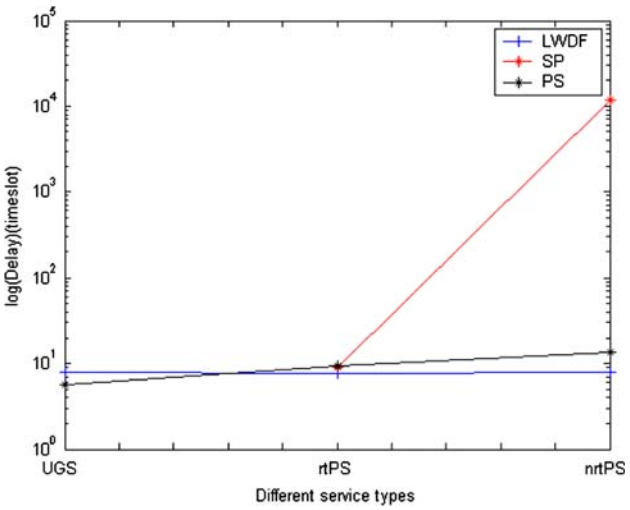


Fig. 14 The comparison of delay for different schemes

delay constraint to the inter-service time requirement can be derived under a specific channel condition.

Figure 17 shows the impact of inter-service time requirement N_i on operation parameter T_i. The values of N_i for the UGS, rtPS, and nrtPS queues are 8, 20, and 30 timeslots, respectively. The values of $\bar{w}_i$ for SS1 to SS4 are 20, 15, 10, and 5, respectively. N_i for rtPS queue at SS1 varies from 10 to 25 timeslots. From Fig. 17, it is observed that T_i for rtPS queue increases when the requirement of N_i for rtPS queue increases. The reason it that the increase of N_i reduces the chance of obtaining services, which can be achieved by adopting a larger T_i. On the other hand, when the chance of serving the rtPS queues decreases, more resources will be released to the BE services. Therefore, it is also observed that smaller T_i is adopted by the BE queue corresponding to this change.

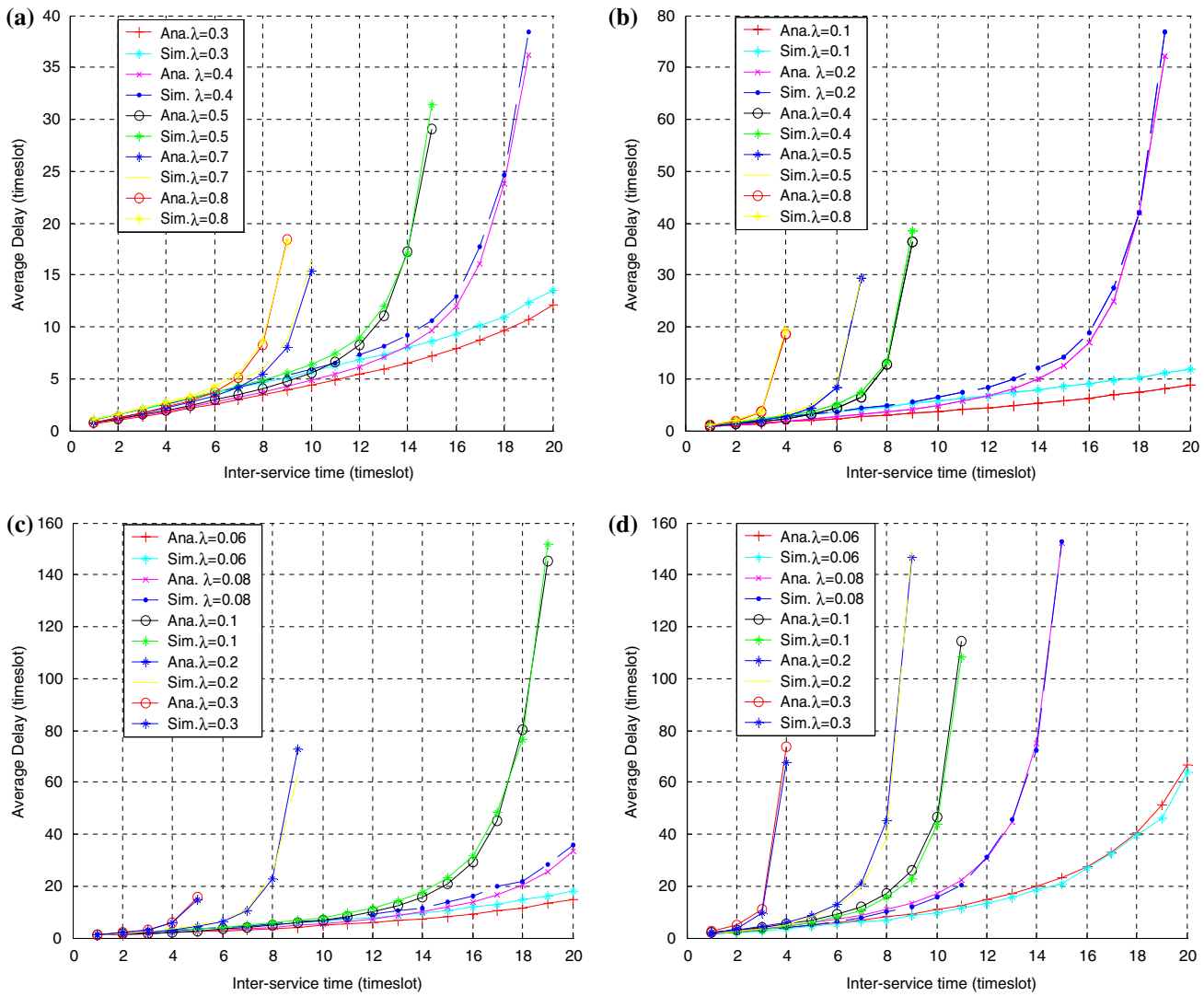


Fig. 16 The average delay versus inter-service time under different channel conditions: for each time slot, maximally (**a**) 8 packets, (**b**) 4 packets, (**c**) 2 packets, and (**d**) 1 packet, can be delivered

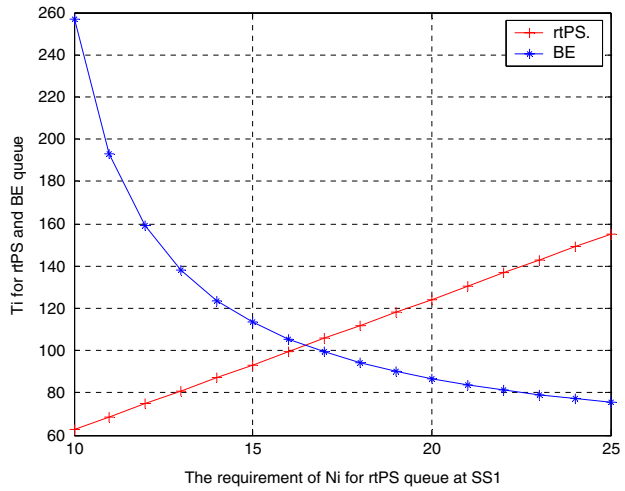


Fig. 17 T_i versus inter-service time requirement N_i

Based on the obtained operation parameter T_i, Table 4 illustrates the inter-service time requirement N_i for rtPS queue at SS1 and the simulation results. It is observed that the simulation results match N_i very well. Therefore, it can be easily concluded that the proposed scheduling scheme can maintain the required N_i by adapting the operation parameter T_i to the fluctuation of the network state, which equivalently meets the delay constraint of each queue.

Figure 18 illustrates the effect of fluctuating channel weighting $\bar{w}_i$ on the operation parameter T_i, where the fixed delay constraints for UGS, rtPS, and nrtPS queues are set as 5, 6, and 10 timeslots, respectively. The arrival rate of each queue keeps unchanged while the weighting of SS2 varies from 4 to 15. In order to maintain the

Table 4 The comparison of the inter-service time requirement N_i and simulation results

N_i	10	12	14	16	18	20
Simulation results	9.5	11.8	13	15.8	17.3	20

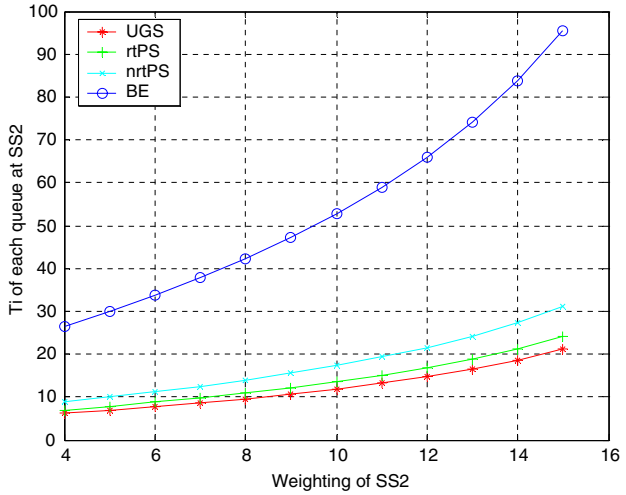


Fig. 18 T_i versus the weighting for each queue at SS2

delay constraint of each queue at SS2, T_i for these queues should be adjusted dynamically, which is shown in Fig. 17. It is observed that T_i increases when the channel weighting of SS2 increases such that the delay constraint of each service type is maintained.

7 Conclusions

An efficient scheduling scheme has been proposed for achieving service differentiation and delay-constraint satisfaction among different service types in IEEE 802.16 networks. Instead of applying strict priority on each type of service, a uniform preference metric is employed for queues with different service types. By jointly considering the effect of average channel conditions and the waiting time experienced by each queue from its previous transmission, the proposed scheme can not only realize opportunistic scheduling to some extent, but also provide the delay-constraint satisfaction for different service types by selecting a specific T_i. Both analytical and simulation studies have been conducted to verify the accuracy of the proposed analysis model and illustrate the efficiency of the proposed scheme.

Acknowledgement This work is partially supported by a Strategic Research Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC).

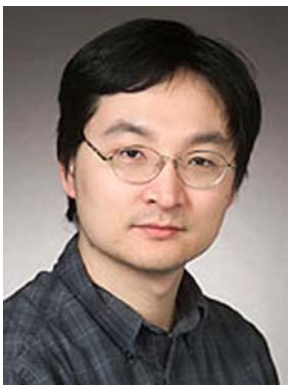
References

1. IEEE 802.16-2001 (April 2002). *IEEE Standard for local and metropolitan access networks part 16: Air Interface for Fixed Broadband Wireless Access Systems.*
2. IEEE 802.16a^{M}-2003 (April 2003). *IEEE Standard for local and metropolitan access network part 16: Air Interface for Fixed Broadband Wireless Access Systems—Amendment 2: Medium Access Control Modifications and Additional Physical Layer Specifications for 2–11 GHz.*
3. Lee, H., Kwon, T., & Cho, D. H. (2005). An enhanced uplink scheduling algorithm based on voice activity for VoIP services in IEEE 802.16d/e System. *IEEE Communications Letters, 9*, 691–693.
4. Chen, J., Jiao, W., & Guo, Q. (2005). An integrated QoS control architecture of IEEE 802.16 broadband wireless access systems. *Proc. Globecom'05* (Vol. 6, pp. 3330–3335).
5. Kitti, W., & Aura, G. (2003). Packet scheduling for QoS support in IEEE 802.16 broadband wireless access systems. *International Journal of Communication Systems 16*, 81–96.
6. Liu, Q., Zhou, S., & Giannakis, G. B. (2005). Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks. *IEEE JSAC, 23*, 1056–1066.
7. Liu, X., Chong, E. K. P., & Shroff, N. B. (2003). A framework for opportunistic scheduling in wireless networks. *International Journal of Computer and Telecommunication Networking, 41*, 451–474.
8. Mehrjoo, M., Dianati, M., Shen, X., & Naik, K. (2006). Opportunistic fair scheduling for the downlink of IEEE 802.16 wireless metropolitan area networks. *Proc. QShine'06.*
9. Xu, L., Shen, X., & Mark, J. W. (2004). Dynamic fair scheduling with QoS constraints in multimedia wideband CDMA cellular networks. *IEEE Transaction on Wireless Communications, 3*, 60–73.
10. Park, D., Seo, H., Kwon, H., & Lee, B. G. (2005). Wireless packet scheduling based on the cumulative distribution function of user transmission rates. *IEEE Transaction on Communications, 53*, 1919–1929.
11. Jalali, A., Padovani, R., & Pankaj, R. (2000). Data throughput of CDMA-HDR a high efficiency-high data rate personal communication wireless system. *Proc. IEEE Veh. Technol. Conf.* (pp. 1854–1858).
12. Hushner, H. J. (2002). Asymptotic properties of proportional-Fair Sharing Algorithm. *Proc. Allerton Conf 2002*. Champaign-Urbana, IL: University of Illinois Press.
13. Choi, J.-G., & Bahk, S. (2004). Cell throughput analysis of the proportional fair scheduling policy. *Proc. Networking 2004* (pp. 247–258).
14. Liu, Q., Wang, X., & Giannakis, B. B. (2006). A cross-layer scheduling algorithm with QoS support in wireless networks. *IEEE Transaction on Vehicular Technology, 55*, 839–847.
15. Stolyar, A. L., & Ramanan, K. (2001). Largest weighted delay first scheduling: Large deviations and optimality. *The Annals of Applied Probability, 11*(1), 1–48.
16. Andrews, M., Kumaran, K., Ramanan, K., Stolyar, A., Whiting, P., & Vijayakumar, R. (2001). Providing quality of service over a shared wireless link. *IEEE Communication Magazine, 39*, 150–154.

17. Shakkottai, S., & Stolyar, A. L. (2001). Scheduling algorithms for a mixture of real-time and non-real-time data in HDR. *Proc. International Teletraffic Congress (ITC)* (pp. 793–804).
18. Rhee, J. H., Holtzman, J. M., & Kim, D. K. (2004). Performance analysis of the adaptive EXP/PF channel scheduler in an AMC/TDM system. *IEEE Communications Letters, 8*, 497–499.
19. Bolch, G., Greiner, S., de Meer, H., & Trivehi, K. S. (2005). *Queueing networks and markov chain: Modeling and performance evaluation with computer science applications*. USA: John Wiley & Sons.
20. Jain, R., Chiu, D., & Hawe, W. (1984). A quantitative measure of fairness and discrimination for resource allocation in shared computer system. *DEC Technical Report 301*.
21. Dianati, M., Shen, X., & Naik, S. (2005). A new fairness index for radio resource allocation in wireless networks. *Proc. WCNC* (pp. 712–717).

Author Biographies

Fen Hou is currently a doctoral student in the Department of Electrical and Computer Engineering, University of Waterloo, Canada. Her research interests include resource allocation, connection admission control, and scheduling scheme in IEEE 802.16 networks.



Pin-Han Ho received his B.Sc. and M.Sc. degree from the Electrical and Computer Engineering department in the National Taiwan University in 1993 and 1995. He started his Ph.D. study in the year 2000 at Queen's University, Kingston, Canada, focusing on optical communications systems, survivable networking, and QoS routing problems. He finished his Ph.D. in 2002, and joined the Electrical and Computer Engineering department in the University of Waterloo, Waterloo, Canada, as an assistant professor at the same year. Professor Pin-Han Ho is the first author of more than 60 refereed technical papers and book chapters, and the co-author of a book on optical networking and survivability. He is the recipient of the Best Paper Award and the Outstanding Paper Award from SPECTS'02 and HPSR'02, respectively.

Xuemin (Sherman) Shen received the B.Sc. (1982) degree from Dalian Maritime University (China) and the M.Sc. (1987) and Ph.D. degrees (1990) from Rutgers University, New Jersey (USA), all in Electrical Engineering. From September 1990 to September 1993, he was first with the Howard University, Washington D.C., and then the University of Alberta, Edmonton (Canada). Since October 1993, he has been with the Department of Electrical and Computer Engineering, University of Waterloo, Canada, where he is a Professor. Dr. Shen's research focuses on mobility and resource management in interconnected wireless/wired networks, UWB wireless communications systems, wireless security, and ad hoc and sensor networks. He is a co-author of three books, and has published more than 200 papers in wireless communications and networks, control and filtering.