Fair Class-Based Downlink Scheduling with Revenue Considerations in Next Generation Broadband Wireless Access Systems

Bader Al-Manthari, *Member*, *IEEE*, Hossam Hassanein, *Senior Member*, *IEEE*, Najah Abu Ali, *Member*, *IEEE*, and Nidal Nasser, *Member*, *IEEE*

Abstract—The success of emerging Broadband Wireless Access Systems (BWASs) will depend, among other factors, on their ability to manage their shared wireless resources in the most efficient way. This is a complex task due to the heterogeneous nature, and hence, diverse Quality of Service (QoS) requirements of different applications that these systems support. Therefore, QoS provisioning is crucial for the success of such wireless access systems. In this paper, we propose a novel downlink packet scheduling scheme for QoS provisioning in BWASs. The proposed scheme employs practical economic models through the use of novel utility and opportunity cost functions to simultaneously satisfy the diverse QoS requirements of mobile users and maximize the revenues of network operators. Unlike existing schemes, the proposed scheme is general and can support multiple QoS classes with users having different QoS and traffic demands. To demonstrate its generality, we show how the utility function can be used to support three different types of traffic, namely best-effort traffic, traffic with minimum data rate requirements, and traffic with maximum packet delay requirements. Extensive performance analysis is carried out to show the effectiveness and strengths of the proposed packet scheduling scheme.

Index Terms—BWASs, packet scheduling, QoS, utility, opportunity cost, fairness.

1 Introduction

Emerging Broadband Wireless Access Systems (BWASs), such as High-Speed Downlink Packet Access (HSDPA) [1] and 802.16 broadband wireless access system (WiMAX) [2], pose a myriad of new opportunities for leveraging the support of a wide range of multimedia applications with diverse Quality of Service (QoS) requirements. This is due to the high data rates that are supported by these systems, which were previously only available to wireline users. Despite the support for high data rates, satisfying the diverse QoS of users while maximizing the revenues of network operators is still one of the major design issues in these systems. Therefore, QoS provisioning is crucial for the success of BWASs. QoS provisioning in BWASs is a challenging problem due to the diverse QoS requirements of the applications that these systems support and the utilization of downlink-shared channels for data delivery instead of dedicated ones. QoS provisioning in BWASs can be done at three different levels, as shown in Fig. 1. These levels are admission level, class level, and packet level. Admission-level QoS provisioning is responsible for accepting or rejecting new users' connections. It aims at satisfying the long-term QoS of users by maximizing the number of admitted connections while maintaining the QoS of ongoing ones. Whereas, class-level QoS provisioning deals with the aggregate demand of admitted users. It determines the number of transmission time frames that each QoS class needs in order to maintain the QoS of its admitted users. Once the time frames are provisioned between different QoS classes, packet-level QoS provisioning is utilized in order to determine which of the users' packets are transmitted in a single frame. The functionality of packet-level QoS provisioning is, therefore, equivalent to the functionality of packet scheduling in BWASs. Throughout this paper, packet-level QoS provisioning will be refereed as packet scheduling.

Packet scheduling is one of the most important components of BWASs that affects system capacity, revenue, and potential QoS provided to users. A downlink packet scheduler is implemented at the base stations of BWASs to control the allocation of the downlink-shared channels to the mobile users by deciding which of them should be transmitted to during a given time frame, and thus to a large extent, the scheduler determines the overall behavior of these systems. Therefore, packet schedulers should be carefully designed in order to maximize the efficiency of the wireless access systems, satisfy the various QoS requirements of users, and maximize the obtained revenues. This paper focuses on downlink packet scheduling.

1.1 Design Issues of Packet Scheduling

There are four important design issues that should be considered in developing packet scheduling schemes. The first is the consideration of the mobile users channel quality conditions. Mobile users experience varying channel conditions that affect their supportable data rates (i.e., maximum

- *B. Al-Manthari and H. Hassanein are with the Telecommunications Research Laboratory, School of Computing, Queen's University, Kingston, ON K7L 3N6, Canada. E-mail: {manthari, hossam}@cs.queensu.ca.*
- *N.A. Ali is with the College of Information Technology, UAE University, Al-Ain, PO Box 17555, UAE. E-mail: najah@uaeu.ac.ae.*
- *N. Nasser is with the Department of Computing and Information Science, University of Guelph, Guelph, ON N1G 2W1, Canada. E-mail: nasser@cis.uoguelph.ca.*

Manuscript received 14 May 2008; revised 5 July 2008; accepted 29 Sept. 2008; published online 22 Jan. 2009.
For information on obtaining reprints of this article, please send e-mail to: tmc@computer.org, and reference IEEECS Log Number TMC-2008-03-0085. Digital Object Identifier no. 10.1109/TMC.2009.30.

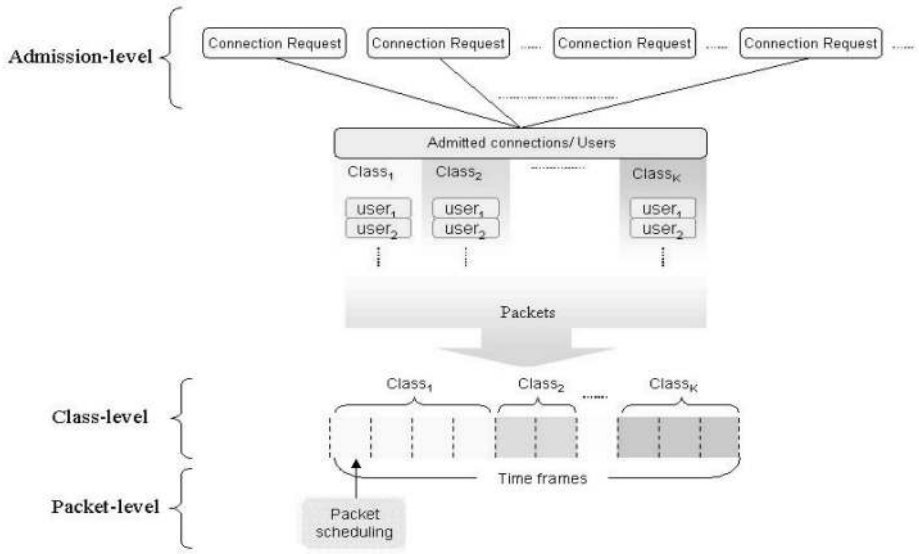


Fig. 1. Levels of QoS provisioning.

downlink data rates) from time to time due to their mobility, interference from other users, obstacles, etc. The packet scheduler should track the instantaneous channel quality conditions of the users and select for transmission those who are experiencing good channel quality conditions in order to maximize the system capacity. However, serving the users based on their favorable channel quality conditions raises the issue of fairness as those users with bad channel quality conditions may not get served, and hence, they may suffer from starvation. Therefore, fairness is another important issue that has to be taken into consideration while designing packet scheduling schemes. The third issue is the QoS requirements of different applications. Next generation BWASs will support a wide range of multimedia applications that have different QoS requirements, and therefore, these requirements should be taken into consideration to maximize satisfactions of users. The fourth issue is revenue loss due to scheduling low-revenue generating users. Network operators incur different revenue losses from serving users depending on their channel quality conditions, amount of the buffered data they have at the base station, and the amount of money they are willing to pay for different services. A good design of a scheduling scheme should consider such revenue losses and aim at minimizing them. Striking a proper balance between these design issues is the main focus of this paper.

1.2 Related Work

Existing packet scheduling schemes can be classified into two groups: nonreal-time and real-time scheduling. Nonreal-time scheduling schemes are designed for nonreal-time and best-effort traffic, where the users' average throughputs are the main QoS metric. Whereas real-time scheduling schemes are designed for multimedia traffic with various QoS requirements such as minimum data rate or maximum delay requirements. An overview of CDMA-related QoS provisioning techniques including state-of-the-the-art packet scheduling is presented in [3]. The most well-known nonreal-time packet scheduling schemes in next generation BWASs are the Maximum Carrier to Interface Ratio (Max CIR) [4] and Proportional Fairness (PF) [5]. Max CIR serves the users with the best channel quality conditions, hence, maximizing the system capacity at the expense of fairness. PF, on the other hand, tries to balance the capacity-fairness trade-off by serving the users with the best relative channel quality, where the relative channel quality is the user's channel quality condition divided by his average throughput. Therefore, the PF scheme gives more priority to users as their average throughputs decrease in order to prevent users with good channel quality conditions from monopolizing the wireless resources as is normally the case with Max CIR. It has been shown, however, that the PF scheme is fair only in ideal cases when users experience similar channel conditions. The PF scheme, therefore, becomes unfair and unable to exploit multiuser diversity in more realistic situations, where users usually experience different channel conditions [6]. To solve this problem, a Score-Based (SB) scheduling scheme is proposed in [6]. Unlike the PF scheme, the SB scheme selects the user whose current channel quality condition is high relative to his own rate statistics instead of selecting the one whose channel quality condition is high relative to his average throughput. Another proposal is Fast Fair Throughput (FFT) [7]. FFT modifies the PF scheme by multiplying the relative channel quality of the users by an equalizer term to ensure a fair long-run throughput distribution among them. In an earlier study [8], we attempted to maximize the system capacity by considering the instantaneous channel quality conditions of the users while at the same time prioritizing users who are receiving very low average throughputs. In another study [9], we represented each user by a utility function that measures his satisfaction of the perceived service. Then, we formulated our scheduling problem so that it maximizes the social welfare of the system in terms of the aggregate utility of system. However, in both studies, we considered only nonreal-time services.

In [10] and [11], a packet scheduling scheme known as the Modified Largest Weighted Delay First (M-LWDF) is proposed to accommodate real-time traffic. M-LWDF uses the relative channel quality condition to compute the user's priority similar to PF. However, to accommodate real-time traffic with delay requirements, M-LWDF multiplies the

user's relative channel quality condition by a term representing the user's packet delay. This term ranges from 0 to 1, where it approaches 1 as the user's head of queue packet delay approaches its delay threshold. It has been shown in [12] that M-LWDF may result in unfair distribution of wireless resources since if two users have the same head of queue packet delay, they will be assigned different priorities if their supportable data rates are different. Therefore, an enhancement of M-LWDF, referred as the Fair Modified Largest Weighted Delay First (FM-LWDF), is proposed in [12] to improve the fairness of M-LWDF. FM-LWDF borrows the equalizer term from the FFT scheme and adds it to M-LWDF in order to improve fairness among users. Another proposal for real-time traffic is the Max CIR with Early Delay Notification (EDN) [13]. This scheme tries to maximize the system capacity by scheduling users using the Max CIR scheme as long as their packet delays are below a certain threshold. If the packet delays of one or more users exceed a certain threshold, then the packets that have been queued the longest time are served first.

The schemes in [14], [15], and [16] represent each user by a utility function depending on his traffic type and aim at maximizing the users' utilities. The scheme in [14] provides two utility functions, one for delay-constrained traffic and the other for best-effort traffic. The scheme, however, ignores users with data rate requirements. In addition, even though the scheme supports fairness among best-effort users, it ignores fairness among delay-sensitive users. Moreover, the scheme does not provide any efficient way to achieve inter- and intraclass prioritization (i.e., prioritization between different classes of traffic and prioritization between different users, respectively), which may limit its practicality. Similar to the scheme in [14], the scheme in [15] uses different utility functions depending on the data rate requirements of users (e.g., stringent, flexible, etc.). The scheme, however, ignores delay-sensitive users. In addition, the scheme does not take into account the instantaneous channel quality conditions of mobile users in the scheduling decisions, which is one of the most important features of packet scheduling in next generation BWASs. The scheme in [16] seems to provide acceptable QoS support as it considers best-effort traffic, traffic with data rate requirement, and traffic with delay requirements. The scheme, however, ignores inter- and intraclass fairness, which may result in unfair distribution of the wireless resources.

Therefore, packet-level QoS provisioning in BWASs is still an open issue due to the need for a packet scheduling scheme that is capable of simultaneously supporting various QoS requirements in addition to providing effective inter- and intraclass prioritization and fairness. We remark that none of the schemes discussed in this section considers the revenues of the network operator, which may limit their viability. Hence, network operator's revenues should be considered as an additional dimension to the scheduling problem.

1.3 Contributions

In this paper, we propose a novel packet scheduling scheme to be implemented at base stations of next generation BWASs. Our proposed scheme is designed to simultaneously achieve the following objectives:

1. Supporting multiple classes of service for users having different QoS and traffic demands.
2. Satisfying the conflicting requirements of the users and network operators.
3. Maximizing the throughput of the wireless system.
4. Ensuring a fair distribution of wireless resources.

Unlike most existing schemes, where different users within each class are assumed to have the same QoS requirements, we consider a more generalized problem, which is supporting multiple users with different QoS requirements within each class. This is more practical since each QoS class in next generation BWASs can include various applications with different QoS requirements (e.g., video and audio streaming in the streaming class). Another problem that is uniquely dealt with in our scheme is satisfying the conflicting requirements of the network operator (i.e., high revenues) and the users (i.e., guaranteed QoS). In practice, different users may have different preferences depending on many factors including the types of applications they are running, age, budgets, etc. These preferences are accounted for in our scheme by employing a utility function with certain realistic properties. To this end, we provide specific definitions for the utility function to support three different types of traffic, namely, best-effort traffic, traffic with minimum data rate requirements, and traffic with maximum delay requirements. The preferences of the network operator are represented by an opportunity cost function to bound revenue loss resulting from serving low-revenue-generating users. To maximize the system throughput, the proposed packet scheduling scheme utilizes the information about the channel quality conditions of the users in its scheduling decisions. Furthermore, we provide unique fairness measures for the traffic cases that are considered in this paper to provide a fair distribution of the wireless resources.

Organization of the paper. The rest of the paper is organized as follows: Section 2 outlines the system and packet scheduling models. In Section 3, we introduce our proposed packet scheduling scheme and show its effectiveness and unique properties. Simulation models, results, and comparisons are given in Section 4. Finally, conclusions drawn from the paper are discussed in Section 5.

2 System Model and Packet Scheduling Model

We consider a BWAS consisting of a downlink time-slotted channel, as shown in Fig. 2. Transmission is done in time frames of fixed or variable size duration, where each frame consists of a number of time slots. We assume that the base station serves N users. We also assume that there are K classes of traffic, where class i has higher priority than class $i+1$. Let N_i denote the number of class i users and $N = \sum_{i=1}^{K} N_i$. We allow users within the same class to have different QoS requirements depending on the type of applications they are running. Packet scheduling in next generation BWASs works as follows. Each user regularly informs the base station of his channel quality condition by sending a report in the uplink to the base station. The report contains information about the instantaneous channel quality condition of the user. The base station then would use this information to select the appropriate user(s)

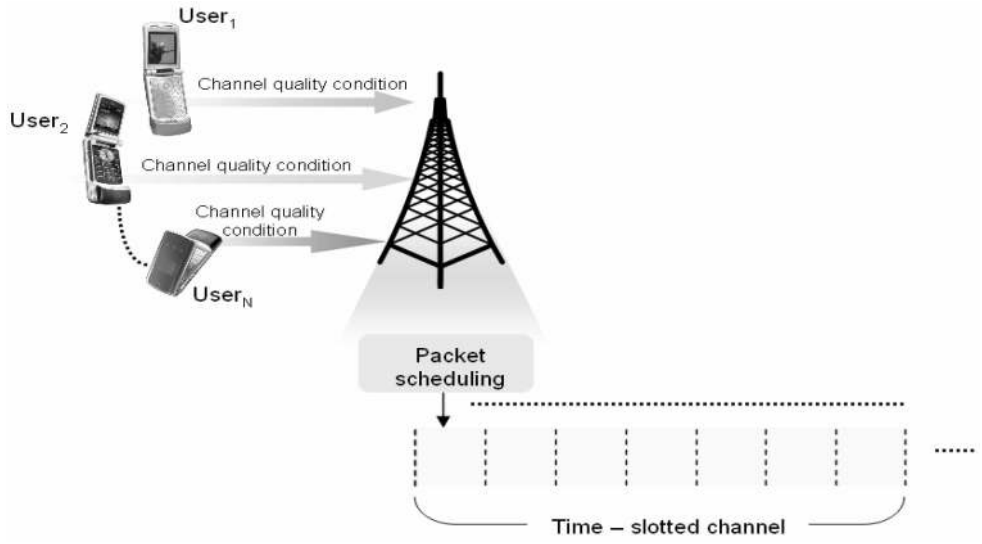


Fig. 2. Packet scheduling.

according to the adopted scheduling scheme. For example, in HSDPA, users are able to measure their current channel quality conditions by measuring the power of the received signal from the base station and then using a set of models described in [17], to determine their current supportable data rate.

3 Fair Class-Based Packet Scheduling Scheme

In this section, we present our proposed packet scheduling scheme,[1] which we refer to as Fair Class-Based Packet Scheduling (FCBPS). We first begin by outlining the general formulation of the scheduling problem, which includes a general utility function to represent the satisfactions of mobile users and an opportunity cost function to represent the cost of serving them (in terms of revenue loss). Then, we state the conditions that the utility function should satisfy and propose a possible utility function that meets the stated conditions.

The satisfaction of user j of class i at time t as perceived by the network operator can be expressed by a utility function $U_{ij}(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}), \{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}} = \{X_{ij}^{1}(t), X_{ij}^{2}(t), \ldots, X_{ij}^{m_{ij}}(t)\}$, where $X_{ij}^{1}(t), \ldots, X_{ij}^{m_{ij}-1}(t)$ are chosen QoS quantitative measures of the user's satisfactions of the wireless system such as the average throughput, current data rate, average delay, etc., $X_{ij}^{m_{ij}}(t)$ is a fairness measure that represents the level of fairness of the scheduling scheme to the traffic generated by user j, $z = 1, 2, \ldots, m_{ij}$ is an index that refers to any of the QoS measures, and m_{ij} is the maximum number of chosen quantitative measures for user j. The main objective of our packet scheduling scheme is to find a subset of users ($\mathbf{N}^*$) to transmit their packets to in order to maximize social welfare, which is the summation of user utilities [18]. Thus, the scheduling scheme can be formulated as the following optimization problem:

$$
\begin{aligned}
&\text{Objective:} \max_{(i,j)\in\mathbf{N}^*,\mathbf{N}^*\subseteq\mathbf{N}} \sum_{i=1}^{K}\sum_{j=1}^{N_i} U_{ij}\left(\{\mathbf{X}_{ij}^{z}(t)\}_{z=1}^{m_{ij}}\right)\\
&\text{Subject to:} \ \nu_{ij}^{z,\min} \le X_{ij}^{z}(t) \le \nu_{ij}^{z,\max}, \ \ \forall j \in \mathrm{N}, \ \ \forall z, \ 1 \le z \le m_{ij},\\
&\left(\sum_{(i,j)\in\mathbf{N}^*} R_{ij}(t)\right) \le C,\\
&OC_{\mathbf{N}^*}(t) \le H,
\end{aligned}
\tag{1}
$$

where $\mathbf{N}^* \subseteq \mathbf{N}$ is the set of users (represented by the tuple (i, j), where i is the class index and j is the user's index within the class) that are selected to transmit to, $\mathbf{N}$ is the set of the total number of users in the system, the first constraint is used to ensure lower and upper bounds on QoS provided to users (e.g., minimum and maximum data rate), $\nu_{ij}^{z,\min} \in \{\nu_{ij}^{z,\min}\}_{z=1}^{m_{ij}}$ and $\nu_{ij}^{z,\max} \in \{\nu_{ij}^{z,\max}\}_{z=1}^{m_{ij}}$ are predefined values for the lower and upper bounds corresponding to the zth QoS measure for user j (i.e., $X_{ij}^{z}(t)$), respectively, $R_{ij}(t)$ is the current supportable data rate of user j at time t, which depends on his channel quality condition,[2] C is the system capacity, $OC_{\mathbf{N}^*}(t)$ is a cost function representing the cost of serving the selected users at time t (i.e., the users in set $\mathbf{N}^*$), and H is a predefined value. We consider the opportunity cost[3] as our cost function. The concept of opportunity cost can be used to manage the trade-off between fairness and revenue. This is because fairness may force the scheduler to serve low-revenue-generating users resulting in revenue loss to the network operator. Therefore, $OC_{\mathbf{N}^*}(t)$ is used to bound this revenue loss. We define $OC_{\mathbf{N}^*}(t)$ as follows. Let

1. A simplified version of the proposed scheduling scheme appeared in the IEEE ICC 2007 [9].

2. Note that $R_{ij}(t)$ is computed based on the channel quality condition of the user as explained in Section 2. However, if the user requires less than $R_{ij}(t)$ to empty his buffer, then we set $R_{ij}(t)$ to the data rate that is just enough to empty the user's buffer in order to avoid giving more slots than the user needs.

3. The opportunity cost for a good is defined as the value of any other goods or services that a person must give up in order to produce or get that good [18].

- p_{ij}: price per bit for user j of class i.
- $\{\mathbf{Rv}^g\}_{g=1}^{N} = \{Rv_{ij}^1, Rv_{ij}^2, \ldots, Rv_{ij}^N | Rv_{ij}^g \geq Rv_{ij}^{g+1}\}$, where $Rv_{ij}^g = p_{ij} \cdot R_{ij}(t)$ is the revenue that the network operator will earn from user j given that this user is served in the current time frame. That is, the set $\{\mathbf{Rv}^g\}_{g=1}^{N}$ contains all users in descending order of the revenue that the network operator will earn from each one of them provided that they are served in the current time frame.
- $\mathrm{Re}\ v_{Max} = \sum_{g \in \{\mathbf{Rv}^g\}_{g=1}^{N}} Rv^g$, given that

$$\left(\sum_{(i,j) \in \{\mathbf{Rv}^g\}_{g=1}^{N}} R_{ij}(t) \right) \leq C.$$

  $\mathrm{Re}\ v_{Max}$ is the maximum obtainable revenue in the current time frame (i.e., the maximum revenue the network operator can generate in the current time frame). $\mathrm{Re}\ v_{Max}$ is obtained by calculating the revenues of all users that could send in the current time frame (i.e., without exceeding the system capacity) and that if served, they will generate the maximum revenue to the network operator.

Therefore, $OC_{\mathbf{N}^*}(t)$ is defined as follows:

$$OC_{\mathbf{N}^*}(t) = \mathrm{Re}\ v_{Max} - \sum_{(i,j) \in \mathbf{N}^*} p_{ij} \cdot R_{ij}(t). \tag{2}$$

That is, the opportunity cost is a measure of how much revenue the network operator would forego if the users in set $\mathbf{N}^*$ are selected for transmission given that there are higher revenue-generating users (i.e., the users that generate $\mathrm{Re}\ v_{Max}$). The network operator can determine the appropriate level of opportunity cost of fairness by choosing the value of H, and hence, the appropriate level of fairness revenue. For example, the network operator could restrict the revenue loss to be no more than 20 percent of the maximum obtainable revenue (i.e., $H = \zeta \cdot \mathrm{Re}\ v_{Max}$, where $\zeta = 0.2$). Note that if $H = 0$, then this implies that the network operator cannot tolerate any revenue loss, and therefore, only the highest revenue-generating users are scheduled to transmit. On the other hand, if $H = \mathrm{Re}\ v_{Max}$, then the opportunity cost is ignored. In this case, all users are considered for transmission.

3.1 The Utility Function

To ensure the practicality of the scheduling scheme, we require $U_{ij}(\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}})$ to meet the following conditions:

1. $\frac{\partial U_{ij}(t)}{\partial X_{ij}^z(t)} \geq 0, \forall z, z \in \{1,2,\ldots,m_{ij}\}; X_{ij}^z(t) \in \{X_{ij}^1(t), X_{ij}^2(t), \ldots, X_{ij}^{m_{ij}}(t)\}$, the utility should be a nondecreasing function of $X_{ij}^z(t)$ to ensure that the user is satisfied with more allocated network resources (i.e., more $X_{ij}^z(t)$).
2. $$\begin{aligned} U_{ij}(\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}}) &= U_{\min}, \quad \text{if } \mathbf{X}_{ij}^z(t) = \mathbf{X}_{ij}^{z,\min}(t), \\ &\qquad \forall z,\ 1 \leq z \leq m_{ij}, \\ U_{ij}(\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}}) &\geq U_{\min}, \quad \text{otherwise}, \end{aligned}$$

   where $X_{ij}^{z,\min}(t)$ is the minimum value of the zth QoS measure. That is, if all QoS measures are at their minimum values, then the user's utility is at its minimum value (i.e., $U_{\min}$) reflecting that the user is dissatisfied with receiving low QoS. If only some QoS measures are at their minimum values, then the user's utility is larger than or equal to the minimum value.
3. $$\begin{aligned} U_{ij}(\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}}) &= U_{\max}, \quad \text{if } \mathbf{X}_{ij}^z(t) = \mathbf{X}_{ij}^{z,\max}(t), \\ &\qquad \forall z,\ 1 \leq z \leq m_{ij}, \\ U_{ij}(\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}}) &\leq U_{\max}, \quad \text{otherwise}, \end{aligned}$$

   where $X_{ij}^{z,\max}(t)$ is the maximum value of the zth QoS measure. That is, if all QoS measures are at their maximum values, then the user's utility is at its maximum value (i.e., $U_{\max}$). If only some QoS measures are at their maximum values, then the user's utility is less than or equal to the maximum utility.
4. $$\lim_{\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}} \to \{\mathbf{X}_{ij}^{z,\max}(t)\}_{z=1}^{m_{ij}}} U_{ij}(\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}}) = U_{\max},$$

   the higher the network resources the user receives, the higher the user's utility up to a certain maximum value $U_{\max}$, then the utility stays at that level reflecting that any additional allocated network resources will not increase the user's utility.

In addition to the above conditions, we require the utility function to support interclass prioritization. Solving for the above conditions will not produce a unique solution. We, hence, introduce a plausible utility function in (3), with constants $a_i > 0$, $\forall i,\ 1 \leq i \leq K$, to capture the feasible area of the solution

$$U_{ij}\left(\{\mathbf{X}_{ij}^z(t)\}_{z=1}^{m_{ij}}\right) = 1 - e^{-a_i \cdot \sum_{z=1}^{m_{ij}} \left(X_{ij}^z(t)\right)}, \tag{3}$$

where a_i serves as an interclass distinguishing parameter in order to prioritize different classes of traffic, and larger values of a_i result in higher class prioritization. This is because larger values of a_i make the utility function more sensitive to any increase or decrease in the QoS measures (i.e., larger values of a_i increase the slope of the utility function). As explained later, users with steep utility function result in the highest rate of change in it, and hence, they maximize the social welfare of the system.

It is imperative to point out that at every scheduling decision, the variations in the users' QoS measures can be computed whether the users are served or not. Therefore, a solution to (1) can be found by computing the aggregate utility of the system if user j is scheduled and all other users are not, and then finding the set of users (i.e., $\mathbf{N}^*$) with the highest aggregate utility (in descending order[4]) provided that they satisfy the constraints of (1). In other words, a solution to (1) can be found by choosing the set $\mathbf{N}^*$ of users for transmission such that

4. That is, the user with the highest aggregate utility is scheduled to transmit. If this user does not have enough data in his queue to fill the frame, then the user with the next highest aggregate utility is added to the set of selected users and so forth until the frame is filled. The base station in BWASs can send to multiple users simultaneously using code multiplexing as in HSDPA [1] or frequency multiplexing as in WiMAX [2].

$$\begin{aligned}\text{Objective: } \underset{(i,j)\in\mathbf{N}^*,\mathbf{N}^*\subseteq\mathbf{N}}{\arg\max} & \left(\sum_{i\in\mathbf{N}^*}^{K} \sum_{j\in\mathbf{N}^*}^{N_i} 1 - e^{-a_i \cdot \sum_{z=1}^{m_{ij}} \left(X_{ij}^z(t)\right)} \right. \\ & \left. + \sum_{i=1}^{K} \sum_{y=1, y\notin\mathbf{N}^*}^{N_i} 1 - e^{-a_i \cdot \sum_{z=1}^{m_{iy}} \left(X_{iy}^z(t)\right)} \right), \\ \text{Subject to: } & \nu_{ij}^{z,\min} \le X_{ij}^z(t) \le \nu_{ij}^{z,\max}, \ \forall j \in \mathbf{N}, \forall z, \ 1 \le z \le m_{ij}, \\ & \left(\sum_{(i,j)\in\mathbf{N}^*} R_{ij}(t) \right) \le C, \\ & OC_{\mathbf{N}^*}(t) \le H, \end{aligned} \tag{4}$$

where all users (j) in set $\mathbf{N}^*$ are selected to transmit and all other users ($y \notin \mathbf{N}^*$) are not. Since (4) involves the summation of user utilities given that each user is selected to transmit and all others are not, then clearly the users, who result in the highest rate of change in the utility function, are actually the ones that are going to maximize the social welfare of the system. This implies that the steeper the slope of the user's utility, the greater his chance of getting scheduled to transmit. The slope of the utility function in (3) is steeper at low values for the QoS measures. This implies that the users with low QoS measures result in the highest rate of change in the utility function. Hence, these users are given more priority for transmission in order to improve their QoS. This property, which is known in economics as diminishing marginal utility [18], is very important because it can be used to ensure fairness of the scheduling scheme. More discussions about this property are in Section 3.3.

3.2 Dynamic Computation of Opportunity Cost

It is imperative to point out that in some cases, the optimization problem in (4) may not have a feasible solution. This is because the scheduling scheme may have to serve certain users to guarantee certain levels of QoS (e.g., minimum data rate or maximum delay), even though these users do not satisfy the opportunity cost constraint. Therefore, to satisfy both constraints, the bound on opportunity cost (i.e., H) has to be dynamically computed in order to ensure the existence of a feasible solution of (4) as follows. Let

- $\mathrm{Re}\ v_{\mathbf{n}^*} = \sum_{(i,j)\in\mathbf{n}^*} p_{ij} \cdot R_{ij}(t)$, where $\mathbf{n}^* \in \mathbf{N}^*$ is the set of users that must be served at time t (i.e., current time frame) in order to guarantee their QoS requirements. That is, $\mathrm{Re}\ v_{\mathbf{n}^*}$ is the obtainable revenue from users that require QoS guarantees.

In this case, the opportunity cost of serving the users in $\mathbf{n}^*$ is given by $OC_{\mathbf{n}^*}(t) = \mathrm{Re}\ v_{Max} - \mathrm{Re}\ v_{\mathbf{n}^*}$, where $\mathrm{Re}\ v_{Max}$ is defined in Section 3.3. Therefore, to avoid infeasibility in (4), we must have $H \ge OC_{\mathbf{n}^*}(t)$. The network operator could, for example, set a predefined value for H, say ϑ, and use it only when $H \ge OC_{\mathbf{n}^*}(t)$ is satisfied as follows:

$$H = \begin{cases} OC_{\mathbf{n}*}(t), & \text{if} \quad \vartheta \le OC_{\mathbf{n}^*}(t), \\ \vartheta, & \text{otherwise.} \end{cases} \tag{5}$$

3.3 Scheduling Different Types of Traffic

In this section, we define the QoS measures that are used in the utility function to support best-effort traffic, where the user's average throughput is the main QoS, traffic with minimum data rate requirements, and traffic with maximum delay requirements. The QoS measures are chosen so that the scheduling scheme achieves the objectives outlined in Section 1.3. We make the following definitions:

- $\overline{S_{ij}(t)} =$ average throughput for user j of class i up to time t.
- $\max_{ij} \overline{S_{ij}(t)} =$ maximum average throughput achieved among all users up to time t.
- $S_{ij}^{\min} =$ minimum required average data rate of user j of class i.
- $S_{ij}^{\max} =$ maximum required average data rate of user j of class i.
- $D_{ij}^{\max} =$ maximum tolerable average packet delay of user j of class i at time t.
- $\overline{D_{ij}(t)} =$ actual average packet delay of user j of class i at time t.

To achieve our design objectives, we let $m_{ij} = 2$ in (3) and let

- $X_{ij}^1(t) = \mu_{ij}(t) = (\mathrm{P}_{ij}^1 - \frac{R_{ij}(t)}{C})$, where $0 \le \mathrm{P}_{ij}^1 \le 1$. We define this measure in order to exploit the user channel quality conditions in the scheduling decision, and hence, maximize the users' individual data rates and the system throughput. This is because the higher the instantaneous data rate of the user (normalized by the system capacity C), the lower $\mu_{ij}(t)$, which results in a higher rate of change in the utility function in (3) due to its diminishing marginal property as mentioned earlier. Therefore, users with good channel quality conditions will have higher priority to transmit. In addition, when $\frac{R_{ij}(t)}{C} > \mathrm{P}_{ij}^1$, $\mu_{ij}(t)$ becomes negative, and consequently, the utility function in (3) sharply decreases (i.e., its slope becomes steeper). This is shown in Fig. 3, which plots the utility as a function of $X_{ij}^1(t)$ for $a_i = 4$ and $\mathrm{P}_{ij}^1 = 0.3, 0.5$, and 0.7 (the graphs with $\mathrm{P}_{ij}^1 = 0.7$ and 0.3 are shifted on the X-axis by 0.2 and -0.2, respectively, to better show the differences between them). Therefore, P_{ij}^1 can be interpreted as a "penalty" incurred from not serving users with

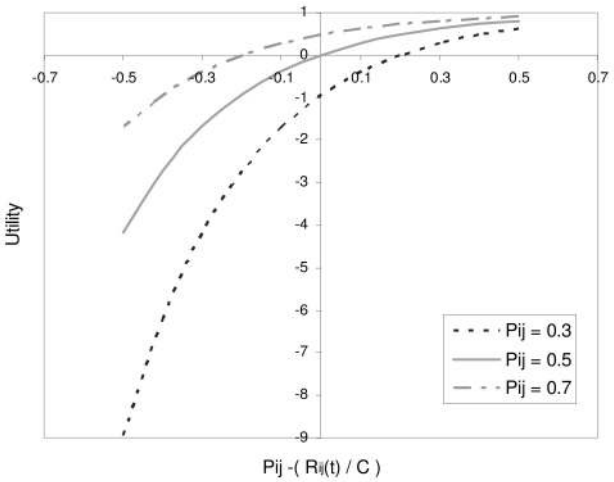


Fig. 3. Effect of P_{ij}^1 on the shape of the utility function.

good channel quality conditions, where smaller values of P_{ij}^1 increase the penalty, and hence give more weight to the users' channel quality conditions in the scheduling decisions.

In our utility function, we use the same $X_{ij}^1(t)$ for all traffic types to increase the system throughput. However, we provide different definitions for $X_{ij}^2(t)$ for the different traffic types. For presentation purpose, let the class index i in $X_{ij}^2(t)$ be e, r, and d for best-effort traffic, traffic with maximum data rate requirements, and traffic with delay requirements, respectively.

For best-effort traffic, we define $X_{ej}^2(t)$ as follows:

- $X_{ej}^2(t) = \alpha_{ej}(t) = \left(\frac{\overline{S_{ej}(t)}}{\max_{ej} \overline{S_{ej}(t)}} - P_{ej}^2\right),\ 0 \le P_{ej}^2 \le 1.$ We define this measure to provide fairness for best-effort traffic. Using this measure, if the user is receiving significantly lower average throughput compared to the one with the maximum average throughput, his fairness will be low indicating his dissatisfaction for the unfairness of the system. In this case, the scheduler will be forced to serve this user to increase his fairness measure. This is because, if a user with high average throughput is served, though his utility will increase, the social welfare of the system will not be maximized because of the rapid decrease of the utilities of those users with low average throughputs as a result of the diminishing marginal property of our proposed utility function. The role of P_{ej}^2 in determining the weight of this measure is similar to the role of P_{ij}^1 in $X_{ij}^1(t)$. However, in this case, larger values of P_{ej}^2 give more weight to $X_{ej}^2(t)$.

For traffic with minimum data rate requirements, we define $X_{rj}^2(t)$ as follows:

- $X_{rj}^2(t) = \sigma_{rj}(t) = \left(\frac{\overline{S_{rj}(t)}}{S_{rj}^{\max}} - P_{rj}^2\right), 0 \le P_{rj}^2 \le 1.$ We define this measure in order to satisfy the users by granting them their required data rates. $\sigma_{rj}(t)$ also represents a fairness measure. This is because if the user is receiving a low average throughput compared to other users who request the same data rate, the rate of decrease in his utility function will be higher than the other users. The scheduler, in this case, will be forced to serve the user to increase his utility, and hence, maximize the social welfare of the system. Larger values of P_{rj}^2 can be used to give more weight to $X_{rj}^2(t)$.

Finally, for traffic with maximum delay requirements, we define $X_{dj}^2(t)$ as follows:

- $X_{dj}^2(t) = \varphi_{dj}(t) = \left(P_{dj}^2 - \frac{\overline{D_{dj}(t)}}{D_{dj}^{\max}}\right), 0 \le P_{dj}^2 \le 1.$ We include this measure in order to satisfy the users' required average packet delays. $\varphi_{dj}(t)$ also represents a fairness measure similar to the case of traffic with data rate requirement. In this case, however, small values of P_{dj}^2 can be used to provide higher weight on $X_{dj}^2(t)$.

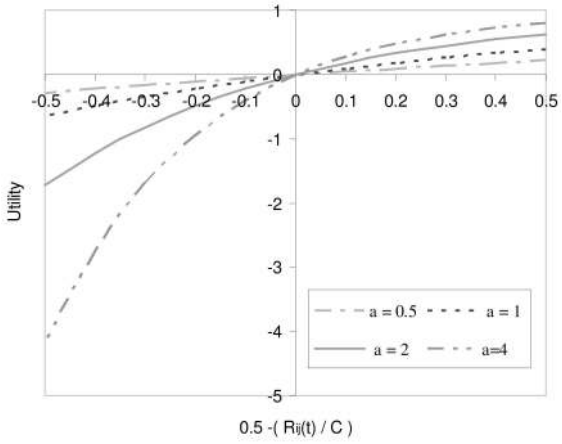


Fig. 4. Effect of a_i on the shape of the utility function.

Using the above definitions, the scheduling problem becomes

$$\begin{aligned}
&\text{Objective:}\ \max_{(i,j)\in \mathbf{N}^*, \mathbf{N}^* \subseteq \mathbf{N}} \sum_{i=1}^{K} \sum_{j=1}^{N_i} 1 - e^{-a_i \cdot \left(\left(X_{ij}^1(t)\right) + \left(X_{ij}^2(t)\right)\right)},\\
&\text{Subject to:}\ S_{rj}^{\min} \le \overline{S_{rj}(t)} \le S_{rj}^{\max}, \quad \forall j \in \mathbf{N}, \forall z,\ 1 \le z \le m_{rj},\\
&\qquad \overline{D_{dj}(t)} \le D_{dj}^{\max}, \quad \forall j \in \mathbf{N}, \quad \forall z, \quad 1 \le z \le m_{dj},\\
&\qquad \left(\sum_{(i,j)\in \mathbf{N}^*} R_{ij}(t)\right) \le C,\\
&\qquad OC_{\mathbf{N}^*}(t) \le H.
\end{aligned} \tag{6}$$

The first constraint ensures that the users' average throughputs lie between their minimum and maximum requirements. The second constraint ensures that the users' average packet delays do not exceed their maximum delay.

While P_{ij}^1, P_{ej}^2, P_{rj}^2, and P_{dj}^2 are used to determine the weights of the QoS measures, a_i plays an important role in determining the shape of the utility function, and hence, the level of interclass prioritization. Larger values of a_i increase the slope of the utility function, and thus result in higher class prioritization. This is shown in Fig. 4, which plots the utility function in (3) for different values of a_i (and penalty, i.e., P_{ij} of 0.5). a_i, along with other parameters ($P_{ij}^1, P_{ij}^2 =$, P_{ej}^2, P_{rj}^2, and P_{dj}^2), therefore, should be set appropriately by the network operator as to achieve its desired level of inter- and intraclass prioritization, and hence its desired level of fairness. In the following section, we show the effect of some of these parameters on the system performance.

4 Performance Evaluation

In this section, we evaluate the performance of our proposed packet scheduling scheme by means of dynamic discrete event simulation. We tested our scheme on HSDPA. More information about HSDPA can be found in [1] and [8].

TABLE 1
Utility Function Parameters

Traffic Type	a_i	P_{ij}^1	P_{ij}^2
VoIP	4	0.5	0.4
Audio Streaming	3.5	0.5	0.4
Video Streaming	3	0.5	0.7
FTP	2.5	0.5	0.7

4.1 Simulation and Traffic Models

We consider a single-cell scenario (though we considered inter and intracell interference as discussed in Section 4.2). The base station is located at the center of the cell. The cell radius is 1 km and the base station's transmission power is 38 dBm. To demonstrate the ability of our scheme to support different QoS with users having different QoS requirements, we assume three different QoS classes with four different types of traffic namely VoIP (class 1), audio streaming (class 2), video streaming (class 2), and FTP (class 3). In addition, to demonstrate the ability of our scheme to prioritize different QoS classes (i.e., interclass prioritization), we assume that class 1 has the highest priority and class 3 has the lowest priority. Moreover, we assume that audio streaming has a higher priority than video streaming in order to demonstrate the ability of our scheme to prioritize traffic with different QoS within the same class (i.e., intraclass prioritization). To achieve such prioritizations, we choose appropriate values for a_i, P_{ij}^1, P_{ej}^2, P_{rj}^2, and P_{dj}^2 according to their role in the utility function, as explained in Section 3.3.3. These values are shown in Table 1. Furthermore, for demonstration purpose, we assume that $p_{ij} = 6, 4, 2$, and 1 units of money for VoIP, audio streaming, video streaming, and FTP users, respectively.

For VoIP traffic, we adopt the model in [19], which assumes Adaptive MultiRate (AMR) codec. In this model, packets are generated using a negative exponentially distributed ON-OFF traffic source to simulate the talk and silence spurts, where the mean duration of both ON and OFF periods is 3 s. During the ON periods, a voice packet of 244 bits is generated every 20 ms, corresponding to a source bit rate of 12.2 Kbps, which is comparable to one of the AMR bit rates [20]. The compressed IP/UDP/RTP header increases the bit rate to 13.6 Kbps [21]. The ITU E-model [22] states that when the one-way mouth-to-ear delay exceeds 250 ms, the voice quality rating rapidly deteriorates. About 80-150 ms remain for the base station processing and connection reception when the delay induced by the voice encoder/decoder and other components in the system is subtracted [23]. Therefore, we set the maximum delay threshold for VoIP traffic to a value between 80 and 150 ms, specifically 100 ms.

Audio streaming is modeled using AMR codec with a minimum rate of 12 Kbps, maximum rate of 64 Kbps, maximum packet delay of 150 ms, and a packet size uniformly distributed between 244 and 488 bits. These values are chosen from within the range of specific QoS requirements defined by WCDMA in order to provide adequate service to mobile users [24], [25], [26]. Video streaming is modeled with a minimum data rate of 64 Kbps, a maximum data rate of 384 Kbps, and a packet size uniformly distributed between 1,200 and 2,400 bits [24], [25], [26]. FTP traffic is simulated by a maximum rate of 128 Kbps and a fixed packet size of 1,200 bits. Call durations of VoIP and video streaming users are modeled by an exponential distribution with a mean value of 30 s. Whereas in case of FTP users, it is assumed that each user requests one FTP file of size 50 MB and terminates his call after the download is complete. At initialization, N users are uniformly distributed in the cell. Pedestrian A environment is used in our experiments, where every mobile user moves inside the cell with a constant speed of 3 km/hr. This speed is the recommended value for Pedestrian A environment by the 3 GPP [27]. A total of 10 channel codes are used, which correspond to a total capacity of 7.2 Mbps [17]. In addition, a feedback delay of three time frames is considered in reporting the instantaneous channel quality conditions of users. Call arrivals are modeled as a Poisson process. The simulation time step is one time frame, which is 2 ms in HSDPA [1], and the simulation time is 400 s.

4.2 Channel Model

The channel model describes the attenuation of the radio signal on its way from the base station to the user, and therefore, it describes how the channel condition of the user changes with time depending on the user's environment and speed. In our simulation, we adopt the channel model used in [27], which consists of five parts: distance loss, shadowing, multipath fading, intracell interference, and intercell interference.

4.3 Simulation Results

To provide QoS guarantees (e.g., minimum data rates or maximum packet delays), the scheduling scheme must be supported by a call admission control in order to block users when there is not enough capacity to support such guarantees. In this paper, we focus on packet scheduling in order to show its performance independently from call admission control. We, therefore, do not consider the case of guaranteed QoS in our experiments. In addition, since existing packet scheduling schemes cannot effectively support different types of traffic with different QoS requirements at the same time, we, therefore, distinguish between two cases. In the first case, all users in the system belong to one traffic type only (i.e., VoIP, audio streaming, video streaming, or FTP). For VoIP and audio streaming, we compare the performance of our scheduling scheme (denoted by FCBPS) to that of the M-LWDF, FM-LWDF, and the Maximum CIR with EDN schemes since these schemes are designed for real-time traffic with delay requirements. For video streaming and FTP, we compare the performance of our scheme with that of the CIR, PF, and the FFT schemes since these schemes are designed for nonreal-time traffic with throughput requirements only.

In the second case, we evaluate the performance of our scheme under a multiplexed scenario in which users could request any type of the four traffics considered in our simulation. Such a case is designed to show the ability of scheme to simultaneously serve different users with

different QoS requirements in addition to show its ability to provide inter and intraclass prioritization. In this case, the total arrival rate to the system is equally divided between the three QoS classes.

The following performance metrics are used:

- Average packet delay: average amount of time the packet spends in the queue at the base station in addition to the transmission time (delay of discarded packets is not counted).
- Average throughput: average number of successfully delivered bits over the lifetime of the user's connection.
- Percentage of channel utilization: percentage of the number of transmitted bits to the maximum number of bits that could be transmitted depending on the channel quality conditions of the users.
- Cell throughput: average number of transmitted bits by the base station. It equals to the total number of transmitted bits over the number of servings (i.e., number of transmissions).
- Service coverage: ratio of users who achieve their required QoS with certain outage level. For VoIP and audio streaming, a user's call is considered an outage and is therefore dropped if his packet loss (due to packet discarding, transmission errors, and/or buffer overflow) exceeds 2 percent [19], [28]. For video streaming, a user's call is considered an outage if his achieved average throughput is less than his minimum required rate. Finally, for FTP traffic, a user's call is considered an outage if his achieved average throughput is less than 9.6 Kbps [12].
- Percentage of revenue loss: percentage of revenue loss to the maximum amount of revenue that could be earned, where the maximum revenue is equal to $\mathrm{Re}\, v_{Max}$, as defined in Section 3, and revenue loss is calculated from (2).
- Jain Fairness Index (JFI): a fairness index used to calculate fairness among users that belong to the same class (i.e., intraclass fairness). Let ψ_{ij} be the performance metric for user j, where ψ_{ij} is set to the user's average packet delay for VoIP and it is set to the user's average throughput for video streaming and FTP, then the JFI is calculated as follows [29]:

$$JFI = \frac{\left(\sum_{z=1}^{N_{ij}} \psi_{ij}\right)^2}{N_{ij} \sum_{z=1}^{N_{ij}} \left(\psi_{ij}\right)^2}, \quad \psi_{ij} \geq 0 \quad \forall j, \qquad (7)$$

where N_{ij} is the number of users of class i who request the same QoS. Note that if all users that request the same QoS get the same ψ_{ij}, then JFI $= 1$. Lower JFI values indicate that users have high variances in their achieved QoS, which reveals unfairness in distributing the wireless resources among them according to this scheme.

4.3.1 Case 1: Single Traffic Class

In this section, we discuss the performance results of the evaluated schemes for VoIP and video streaming traffic only. The performance results of audio streaming and FTP are similar to those of VoIP and video streaming, respectively, and hence, they are omitted.

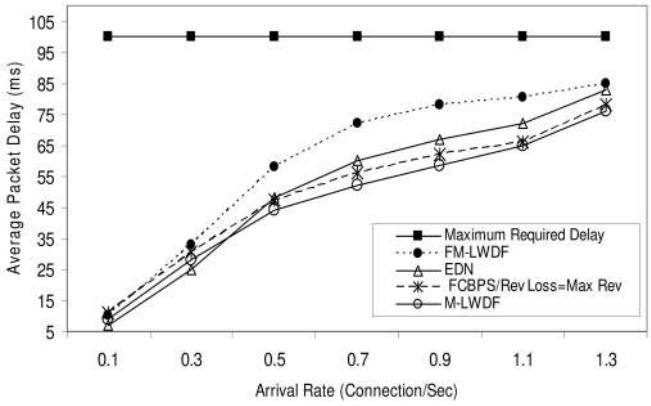


Fig. 5. Average packet delay (VoIP).

VoIP. Fig. 5 depicts the average packet delay for VoIP traffic as a function of the arrival rate to the system. The figure shows that M-LWDF achieves the best packet delay under most network loads, whereas FM-LWDF has the worst packet delay. FM-LWDF performs poorly compared to the other schemes because of its fairness measure (i.e., the equalizer term), which is in terms of throughput and not in terms of delay. Hence, more resources are given to users with bad average throughput at the expense of those users with high packet delays. FCBPS (with maximum tolerable revenue loss of $\mathrm{Re}\, v_{Max}$, i.e., opportunity cost is ignored) achieves reasonably low packet delays at different network loads (within 5 percent of the performance of M-LWDF). This is due to the fact that as the user's average packet delay increases, the sharp decrease in his utility forces the scheduler to serve him, and hence improve his packet delay. The average packet delay achieved by EDN is worse than our scheme and M-LWDF because as the network load increases (i.e., arrival rate ≥ 0.5), the packet delays of users exceed the threshold in EDN, and hence, users are only served based on their packet delays without exploiting their channel quality conditions. Such users require more resources to transmit, causing more packet delays to users with good channel quality conditions. The average packet delays achieved by FCBPS with three different maximum tolerable revenue losses, namely $\mathrm{Re}\, v_{Max}$, $0.5 \cdot \mathrm{Re}\, v_{Max}$ and 0, are shown in Fig. 6. As the maximum tolerable revenue loss decreases, the average packet delay increases. This is because when the maximum tolerable revenue loss is low, only high-revenue-generating users are served by FCBPS, and hence, the packet delays of other users in the system increase causing an increase in the overall average packet delay.

FCBPS also achieves the best performance in terms of service coverage and revenue loss compared to other schemes, as shown in Figs. 7, 8, 9, and 10. The exponential decrease in our proposed utility function when a user experiences high average packet delays causes the scheduler to serve him, and hence, more users are covered by FCBPS. As can be seen, when the maximum tolerable revenue loss decreases, the revenue loss of FCBPS decreases, however, at the expense of service coverage. Therefore, using our scheme, the network operator can determine the level of revenue loss and the corresponding level of service coverage as to maximize its short and long-run revenues.

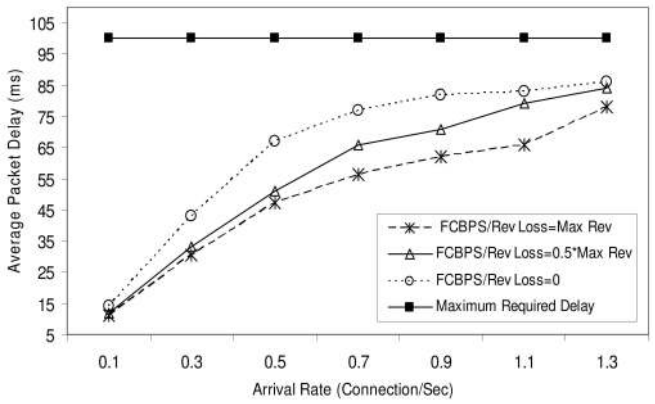
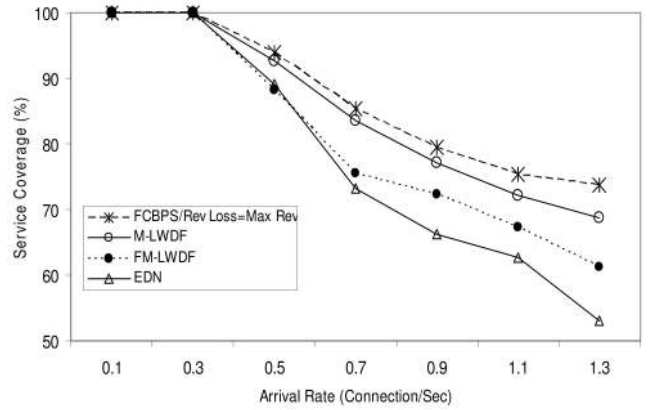


Fig. 6. Average packet delay of FCBPS with different revenue losses.



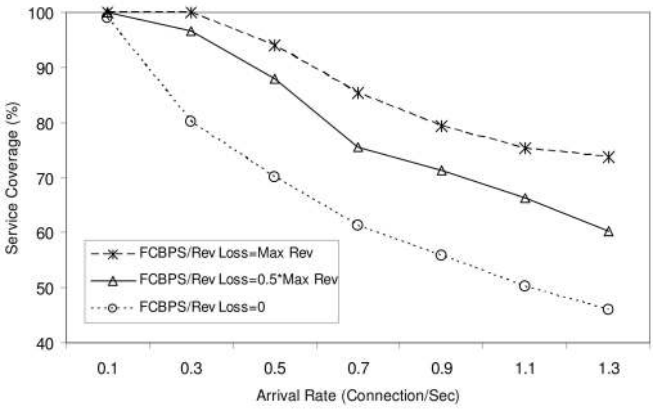


Fig. 8. Percentage of service coverage of FCBPS with different revenue losses.

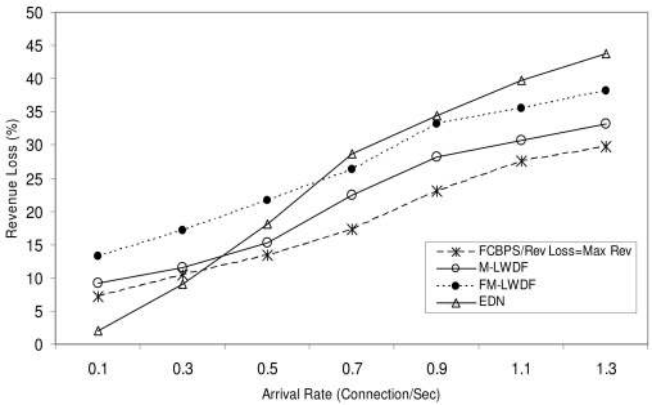


Fig. 9. Percentage of revenue loss (VoIP).

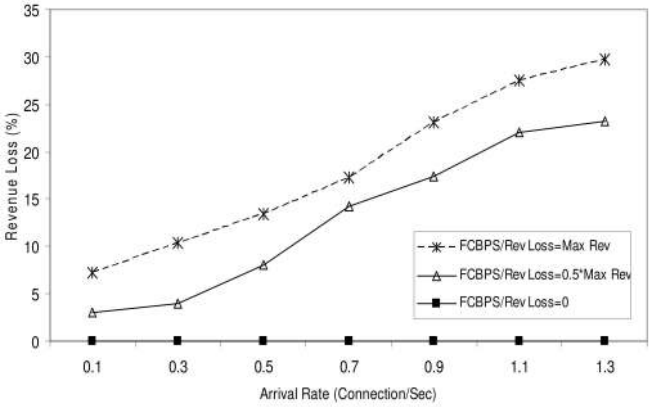


Fig. 10. Percentage of revenue loss of FCBPS with different revenue losses.

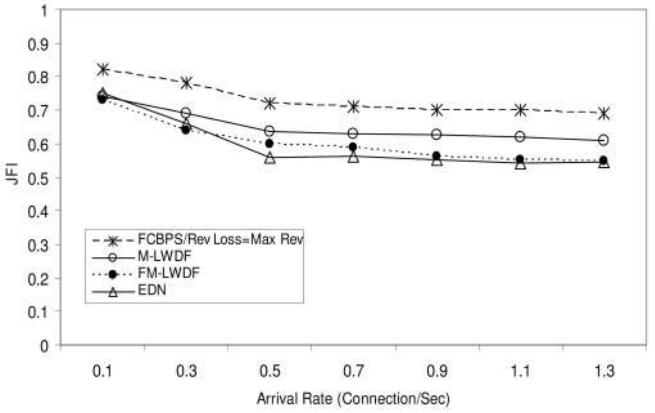


Fig. 11. The Jain fairness index (VoIP).

Moreover, Fig. 11 depicts the JFI of the evaluated schemes, which shows that FCBPS achieves the best fairness performance. This is due to the use of fairness measures in our proposed utility function, which allow the scheme to distribute the wireless resources fairly among users while exploiting the variations in their channel quality conditions. However, when the maximum tolerable revenue loss is decreased, the fairness of FCBPS deteriorates as shown in Fig. 12. This behavior is expected as users are selectively chosen for transmission based on the revenue they generate to the network operator.

Video Streaming. The average through put for video streaming users is shown in Fig. 13. Max CIR achieves the best performance since it schedules the users based on their best channel quality conditions. FCBPS (with maximum tolerable revenue loss of $\mathrm{Re}\ v_{Max}$) outperforms PF. FFT, on the other hand, has the lowest average throughput because of the equalizer term in FFT, which forces it to achieve long-term fairness at the expense of exploiting the channel quality conditions of different users. In addition, the average throughputs of users increase as the maximum tolerable revenue loss is decreased in FCBPS, as shown in Fig. 14. This is because high-revenue-generating users (from the network operator's perspective) are those with good channel quality conditions since more bits could be transmitted in this case. Therefore, as the maximum tolerable revenue loss is

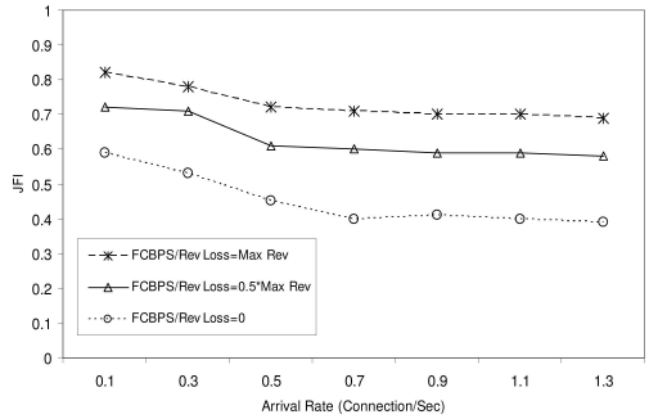


Fig. 12. The Jain fairness index of FCBPS with different revenue losses.

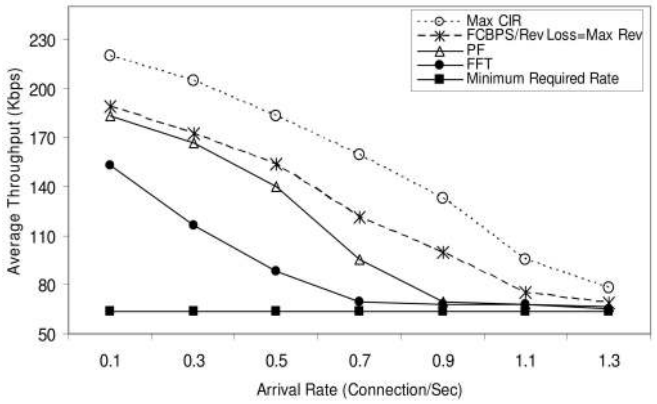


Fig. 13. Average throughput (video streaming).

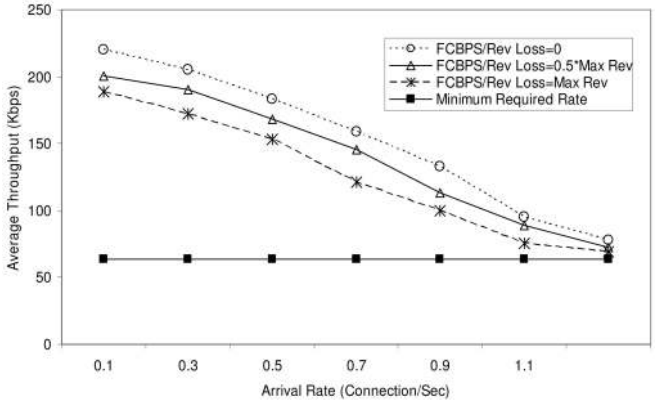


Fig. 14. Average throughput of FCBPS with different revenue losses.

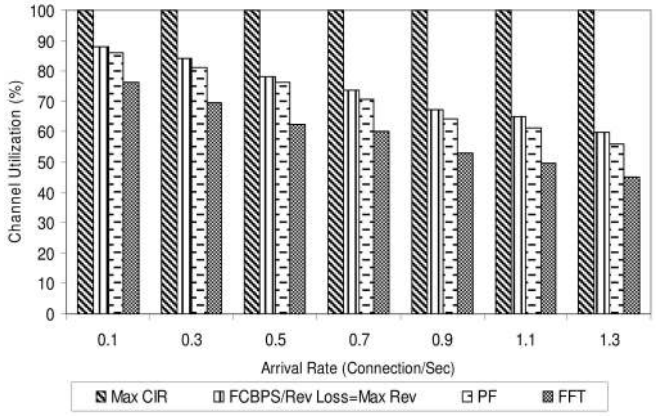


Fig. 15. Percentage of channel utilization (video streaming).

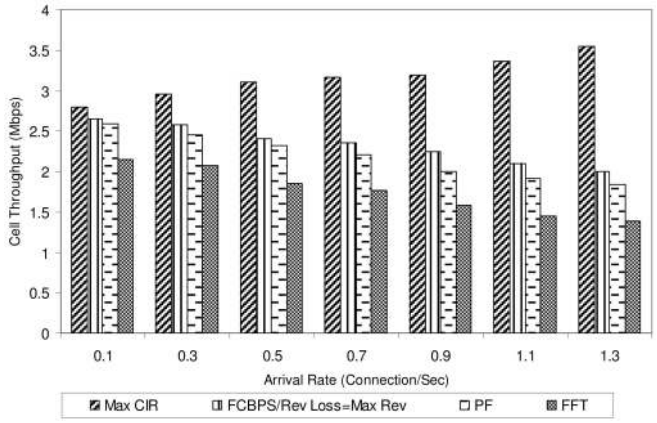


Fig. 16. Cell throughput (video streaming).

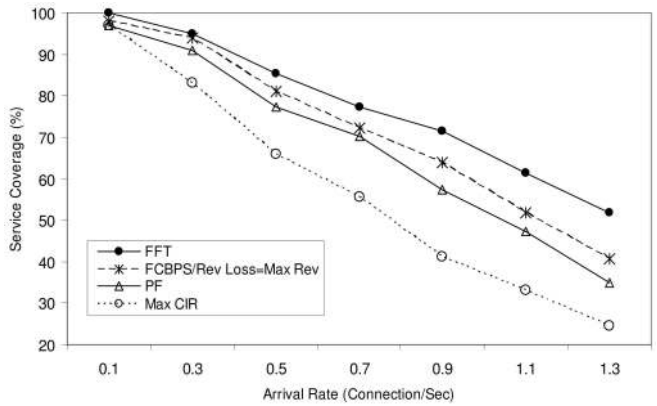


Fig. 17. Percentage of service coverage (video streaming).

decreased, the performance of FCBPS approaches that of Max CIR. Fig. 15 depicts the percentage of channel utilization. FCBPS achieves good utilization levels compared to PF and FFT. Max CIR, however, achieves the best channel utilization (100 percent under all arrival rates) because it serves only the users with the best channel quality conditions, and hence, the channel is fully utilized.

The good channel utilization levels of FCBPS result in good cell throughputs compared to PF and FFT, as shown in Fig. 16. FCBPS also achieves good levels of service coverage compared to PF and Max CIR, as shown in Fig. 17. The best service coverage, nevertheless, is achieved by FFT, as expected due to the equalizer term. This happens, however, at the expense of low channel utilization and low cell throughput as mentioned earlier. Fig. 18 shows the percentage of service coverage of FCBPS for different revenue losses. It can be seen that as the maximum tolerable revenue loss decreases, the service coverage decreases until it reaches that of Max CIR. This confirms our argument that with low maximum tolerable revenue loss, users with good channel quality conditions are favored for transmission over those with bad channel quality conditions since more bits can be transmitted, and hence, greater revenues can be earned. Therefore, packet scheduling schemes that better exploit the channel quality conditions of users result in lowest revenue losses as confirmed in Fig. 19. This, however, comes at the expense of fairness as is clearly shown in Fig. 20.

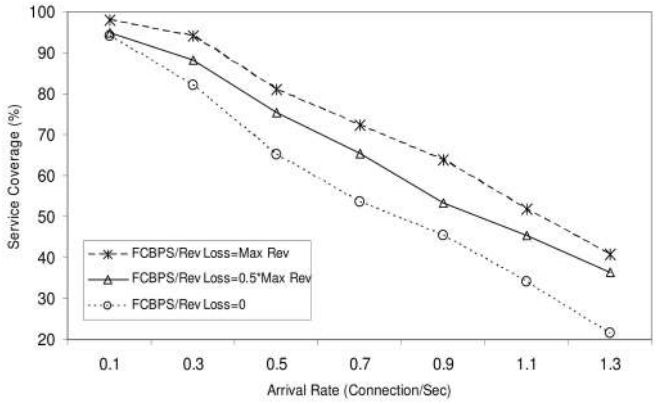


Fig. 18. Percentage of service coverage of FCBPS with different revenue losses.

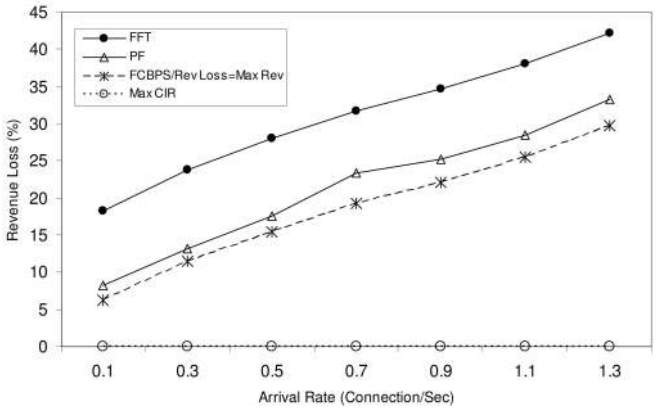


Fig. 19. Percentage of revenue loss (video streaming).

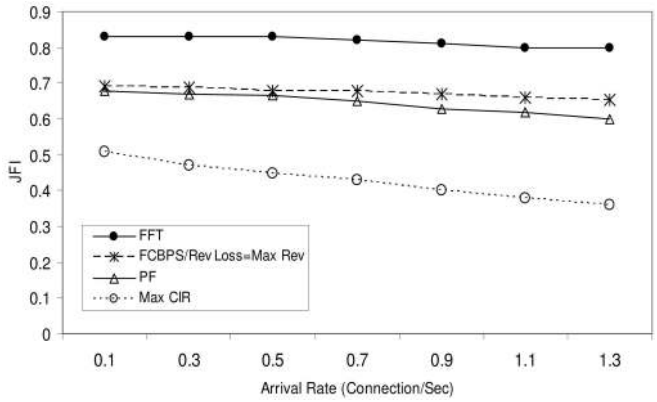


Fig. 20. The Jain fairness index (video streaming).

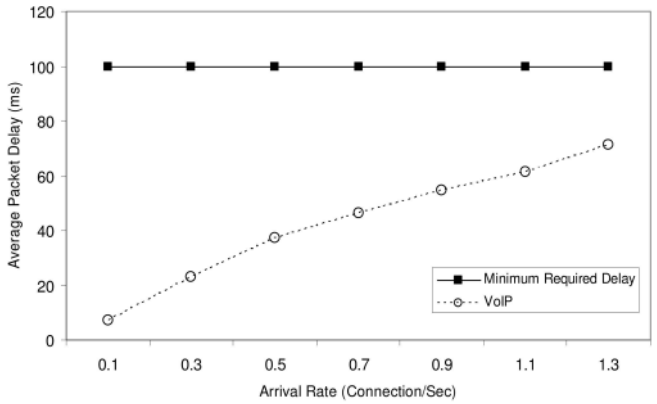


Fig. 21. Average packet delay for VoIP (multiplexed traffic).

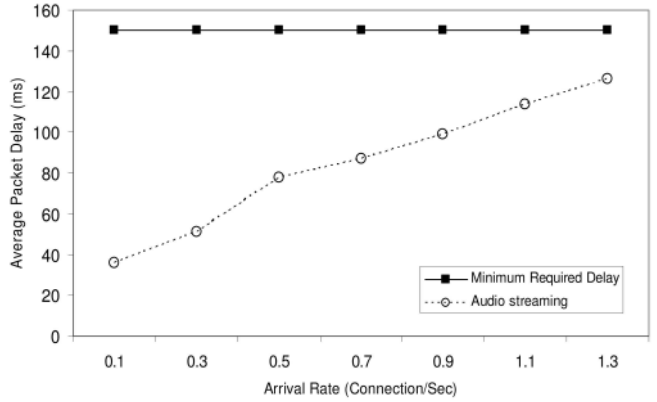


Fig. 22. Average packet delay for audio streaming (multiplexed traffic).

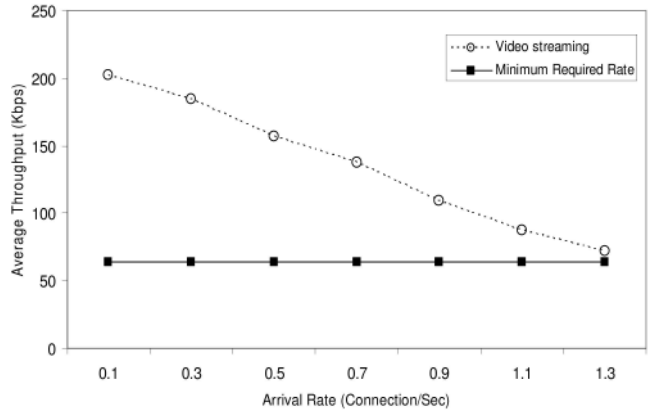


Fig. 23. Average throughput for video streaming (multiplexed traffic).

4.3.2 Case 2: Multiplexed Traffic

In this section, we discuss the performance results of our scheme with maximum tolerable revenue loss of $\mathrm{Re}\, v_{Max}$ under a multiplexed traffic case in order to show its effectiveness in supporting multiple traffic types simultaneously. Figs. 21 and 22 show the average packet delay for VoIP and audio streaming users, respectively. In general, both types of users achieve acceptable average packet delays under different network loads. It should be noted that the VoIP traffic outperforms audio streaming since the latter has lower priority. Moreover, the performance results of VoIP traffic are better than the single traffic case since the total arrival rate in multiplexed traffic is equally divided between the three classes of traffic, and hence, there are fewer VoIP users in this case than the single traffic case (the arrival rate for class 2 is also equally divided between audio and video traffic).

Figs. 23 and 24 depict the average user throughput for video streaming and FTP users, respectively. The figures show that video streaming users achieve higher average throughputs because they have higher priority than FTP users. The percentage of service coverage is shown in Fig. 25. In general, our scheduler achieves acceptable service coverage for different types of traffic at different network loads, where traffic of higher priorities receives higher coverage. Fig. 26 shows the JIF for each traffic type. The JIF of lower priority traffic (i.e., video streaming and

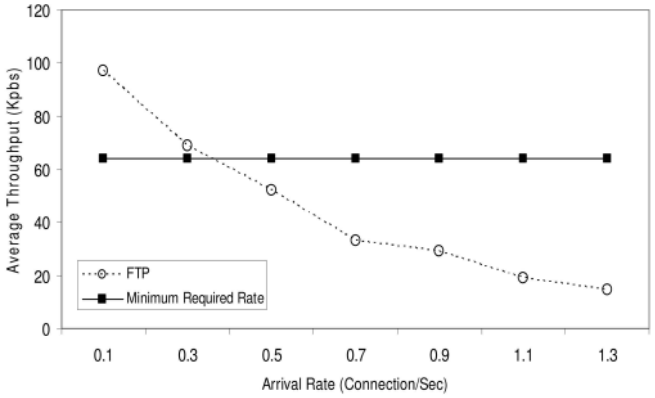


Fig. 24. Average throughput for FTP (multiplexed traffic).

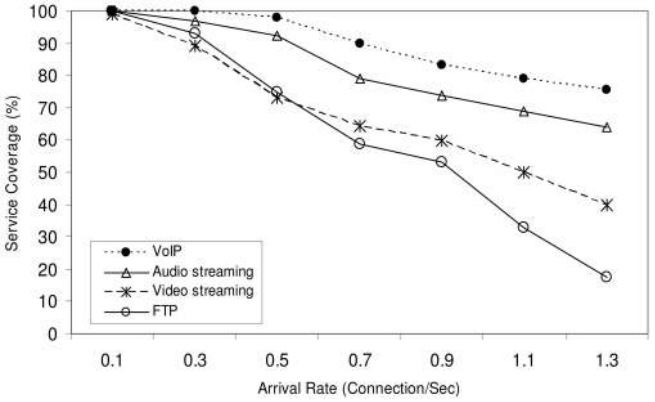


Fig. 25. Percentage of service coverage for all traffic types.

FTP) is less than that of higher priority traffic (i.e., VoIP and video streaming). This is because lower priority traffic is assigned fewer time frames than higher priority traffic, hence, not allowing enough time for our defined fairness measures to make an impact.

These results confirm that using our proposed scheme, the network operator can simultaneously support different types of applications with different QoS requirements, prioritize different types of traffic within the same class (i.e., audio and video streaming), prioritize different classes of traffic, and bound the revenue loss of serving users, hence determining the appropriate level of fairness in the system.

5 Conclusions

The emergence of High-Speed Downlink Packet Access System (HSDPA) and 802.16 broadband wireless access system (WiMAX) will enhance the support of existing applications and will enable the development of a wide range of heterogeneous "content rich" multimedia applications that have different QoS requirements. However, to accommodate as many users as possible while maintaining the quality of their service, these systems require very robust QoS provisioning techniques. A key component of QoS provisioning is packet scheduling. Packet scheduling will play a very important role in broadband wireless access systems since these systems are characterized by high-speed downlink-shared channels to support the increasing number of mobile data users. In this paper, we propose a novel fair class-based downlink packet scheduling scheme for

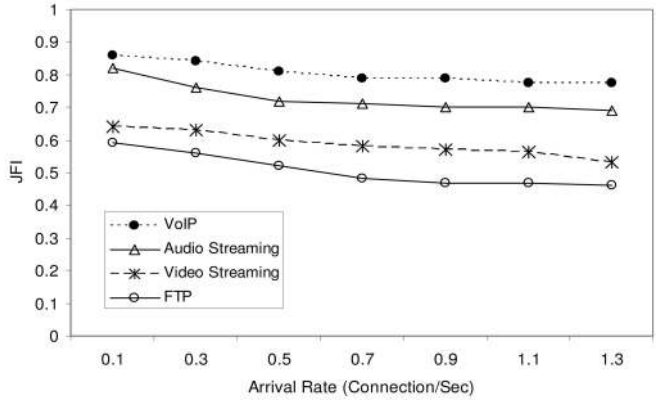


Fig. 26. The Jain fairness index for all traffic types.

broadband wireless access systems. The proposed scheduling scheme utilizes utility and opportunity cost functions to simultaneously satisfy the QoS requirements of users and minimize the revenue loss of network operators. Unlike existing schemes, the proposed scheduling scheme is designed to support multiple QoS classes, where users within the same QoS class can have different QoS requirements. To demonstrate its effectiveness, the proposed scheme is evaluated with three different types of traffic with different QoS requirements, namely, best-effort traffic, traffic with data rate requirements, and traffic with delay requirements. Simulation results show that the proposed scheme can enhance the performance of the wireless system by satisfying the QoS of users, bounding the revenue loss of serving them, and ensuring fairness among them. The scheme, however, is optimized in time domain only. Therefore, to further enhance its performance, we are currently working on optimizing the scheduling decisions in code/frequency domains as well. In addition, we are currently working on extending our scheme to support class level as well as call-level QoS provisioning in order to provide long-term in addition to short-term users' satisfactions.

References

[1] 3GPP TS 25.308, *High Speed Downlink Packet Access (HSDPA); Overall Description,* Release 5, Mar. 2003.

[2] *IEEE 802.16 Working Group, IEEE 802.16-2005e Standard for Local and Metropolitan Area Networks: Air Interface for Fixed Broadband Wireless Access Systems—Amendment for Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands,* IEEE, Dec. 2005.

[3] H. Jiang, W. Zhuang, X. Shen, and Q. Bi, "Quality-of-Service Provisioning and Efficient Resource Utilization in CDMA Cellular Communications," *IEEE J. Selected Areas in Comm.,* vol. 24, no. 1, pp. 4-15, Jan. 2006.

[4] S. Borst, "Connection-Level Performance of Channel-Aware Scheduling Algorithms in Wireless Data Networks," *Proc. IEEE INFOCOM,* vol. 1, pp. 321-331, Mar. 2003.

[5] A. Jalali, R. Padovani, and R. Pankaj, "Data Throughput of CDMA-HDR a High Efficiency-High Date Rate Personal Communication Wireless System," *Proc. IEEE Vehicular Technology Conf. (VTC '00),* pp. 1854-1858, May 2000.

[6] T. Bonald, "A Score-Based Opportunistic Scheduler for Fading Radio Channels," *Proc. European Wireless Conf. (EW '02),* pp. 2244-2248, Sept. 2002.

[7] G. Barriac and J. Holtzman, "Introducing Delay Sensitivity into the Proportional Fair Algorithm for CDMA Downlink Scheduling," *Proc. IEEE Seventh Int'l Symp. Spread Spectrum Techniques and Applications,* vol. 3, pp. 652-656, 2002.

[8] B. Al-Manthari, N. Nasser, and H. Hassanein, "Packet Scheduling in 3.5 G High Speed Downlink Packet Access Networks: Breadth and Depth," *IEEE Networks Magazine,* vol. 21, no. 1, pp. 41-46, Feb. 2007.
[9] B. Al-Manthari, N.A. Ali, N. Nasser, and H. Hassanein, "A Generic Centralized Downlink Scheduler for Next Generation Wireless Cellular Networks," *Proc. IEEE Int'l Conf. Comm. (ICC '07),* pp. 4566-4572, June 2007.
[10] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "Providing Quality of Service over a Shared Wireless Link," *IEEE Comm. Magazine,* vol. 39, no. 2, pp. 150-154, Feb. 2001.
[11] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, R. Vijayakumar, and P. Whiting, "CDMA Data QoS Scheduling on the Forward Link with Variable Channel Conditions," technical report, Bell Labs, Apr. 2000.
[12] P. Jose, "Packet Scheduling and Quality of Service in HSDPA," PhD dissertation, Aalborg Univ., Oct. 2003.
[13] A. Golaup, O. Holland, and A. Aghvami, "A Packet Scheduling Algorithm Supporting Multimedia Traffic over the HSDPA Link Based on Early Delay Notification," *Proc. Int'l Conf. Multimedia Services Access Networks (MSAN '05),* pp. 78-82, June 2005.
[14] G. Song and Y. Li, "Utility-Based Resource Allocation and Scheduling in OFDM-Based Wireless Broadband Networks," *IEEE Comm. Magazine,* vol. 43, no. 12, pp. 127-134, Dec. 2005.
[15] K.-H. Liu, L. Cai, and X. Shen, "Multiclass Utility-Based Scheduling for UWB Networks," *IEEE Trans. Vehicular Technology,* vol. 57, no. 2, pp. 1176-1187, Mar. 2008.
[16] X. Wang, G.B. Giannakis, and A.G. Maraques, "A Unified Approach to QoS-Guaranteed Scheduling for Channel-Adaptive Wireless Networks," *Proc. IEEE,* vol. 95, no. 12, pp. 2410-2431, Dec. 2007.
[17] 3GPP TS25.214 V5.5.0, *Physical Layer Procedures,* Release 5, June 2003.
[18] H. Varian, *Intermediate Microeconomics: A Modern Approach,* sixth ed. W.W. Norton & Company, 2003.
[19] Y.S. Kim, "Capacity of VoIP over HSDPA with Frame Bundling," *IEICE Trans. Comm.,* vol. E89-B, no. 12, pp. 3450-3453, Dec. 2006.
[20] 3GPP TS 26.071 V6, *AMR Speech Codec; General Description,* Dec. 2004.
[21] C. Bormann, C. Burmeister, M. Degermark, H. Fukushima, H. Hannu, L. Jonsson, R. Hakenberg, T. Koren, K. Le, Z. Liu, A. Martensson, A. Miyazaki, K. Svanbro, T. Wiebke, T. Yoshimura, and H. Zheng, *RObust Header Compression (ROHC): Framework and Four Profiles: RTP, UDP, ESP, and Uncompressed,* IETF RFC Standards Track 3095, July 2001.
[22] ITU-T, *One-Way Transmission Time,* G.114, May 2003.
[23] B. Wang, K.I. Pedersen, T.E. Kolding, and P.E. Mogensen, "Performance of VoIP over HSDPA," *Proc. IEEE Vehicular Technology Conf. (VTC '05),* vol. 4, pp. 2335-2339, May 2005.
[24] R. Lloyd-Evan, *QoS in Integrated 3G Networks,* first ed. Artech House, 2002.
[25] 3GPP TS 23.107 V5.12.0, *Quality of Service (QoS) Concept and Architectures,* Release 5, Mar. 2004.
[26] 3GPP TS 22.105 V 6.4.0, *Services and Service Capabilities,* Release 6, Sept. 2005.
[27] Deliverable D3. 2v2, *End-to-End Network Model for Enhanced UMTS,* http://www.ti-wmc.nl/eurane/, 2009.
[28] G. Rittenhouse and H. Zheng, "Providing VoIP Service in UMTS-HSDPA with Frame Aggregation," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05),* vol. 2, pp. 1157-1160, Mar. 2005.
[29] R. Jain, D. Chiu, and W. Hawe, "A Quantitative Measure of Fairness and Discrimination for Recourse Allocation in Shared Computer Systems," DEC Research Report TR-301, Sept. 1984.

Bader Al-Manthari received the BSc, MSc (with honors), and PhD degrees from Queen's University, Kingston, Canada, in 2004, 2005, and 2008, respectively. His research interests include economic-based radio resource management in next generation wireless cellular networks, wireless ad hoc and sensor networks, and performance evaluation of communication protocols and schemes. He is a member of the IEEE.

Hossam Hassanein received the PhD degree in computing science from the University of Alberta, Canada, in 1990. He is a leading researcher in the School of Computing at Queen's University in the areas of broadband, wireless and variable topology networks architecture, protocols, and control and performance evaluation. Before joining Queen's University in 1999, he worked in the Department of Mathematics and Computer Science at Kuwait University (1993-1999) and the Department of Electrical and Computer Engineering at the University of Waterloo (1991-1993). He is the founder and the director of the Telecommunication Research (TR) Lab (http://www.cs.queensu.ca/~trl) in the School of Computing at Queen's. He has publications more than 300 papers in reputable journals, conferences, and workshops in the areas of computer networks and performance evaluation. He has organized and served on the program committee of a number of international conferences and workshops. He also serves on the editorial board of a number of international journals and is currently the vice chair of the IEEE Communication Society Technical Committee on Ad Hoc and Sensor Networks (TC AHSN). He is the recipient of Communications and Information Technology Ontario (CITO) Champions of Innovation Research Award in 2003. In 2007, he received Best Paper Awards at the IEEE Wireless Communications and Networks and the IEEE Global Communication Conferences (both flagship IEEE Communications Society Conferences). He is a senior member of the IEEE.

Najah Abu Ali received the BS and MS degrees in electrical engineering from the University of Jordan, Amman, in 1989 and 1995, respectively, and the PhD degree in 2006 in computer networks from the Electrical Engineering Department, Queen's University, Kingston, Canada. She joined the College of Information Technology, United Arab Emirates University, as an assistant professor with the computer networks engineering track. From January 2006 to August 2006, she had a postdoctoral fellowship at the School of Computing, Queen's University, where she is currently a collaborator member within the research team of Telecommunications Research. From 1995 to 2003, she worked as an instructor and the head of the Engineering Department at Queen Noor College in Jordan. Her research interests include wired and wireless communication networks, specifically analytical and measurement-based network performance management and Quality of Service and resource management of single and multihop wireless networks. She is an expert on broadband wireless networks architecture, design, QoS provisioning, and performance, and has published extensively in that area. She delivered tutorials on WiMax Networks at ICC 2008 and CCNC 2009. She is a member of the IEEE.

Nidal Nasser received the BSc and MSc degrees (with honors) in computer engineering from Kuwait University, State of Kuwait, in 1996 and 1999, respectively, and the PhD degree from the School of Computing at Queen's University, Kingston, Ontario, Canada, in 2004. He is currently an associate professor in the Department of Computing and Information Science, University of Guelph, Ontario, Canada. He has authored several journal publications, refereed conference publications, and seven book chapters. He has also given tutorials in major international conferences. He is a technical editor of Wiley's *International Journal of Wireless Communications and Mobile Computing and Security and Communication Networks Journal.* He has been a member of the technical program and organizing committees of several international IEEE conferences and workshops. He is a member of several IEEE technical committees. He received a Scholarly and Professional Development Award in 2004 from Queen's University and the Best Research Paper Award at the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA '08). He is a member of the IEEE.