Adaptive Token Bank Fair Queuing Scheduling in the Downlink of 4G Wireless Multicarrier Networks

Feroz A. Bokhari*, William K. Wong†, and Halim Yanikomeroglu*

*Broadband Communications and Wireless Systems Centre, Department of System and Computer Engineering
Carleton University, Ottawa Ontario K1S 5B6, Canada
Email: fbokhari, halim@sce.carleton.ca

† Communication Research Centre Canada
3701 Carling Ave., P.O. Box 11490 Station H, Ottawa, Ontario, Canada, K2H 8S2
Email: william.wong@crc.ca

***Abstract*—In this paper we present an efficient cross-layer scheduling algorithm designed for resource allocation in downlink of 4G wireless multicarrier networks. This study focuses on the mechanisms of efficiency, fairness, as well as quality of service (QoS) provisioning and algorithm development for resource allocation in multiuser frequency-selective fading environments. The performance of the proposed scheme is compared to that of the Score Based (SB) scheduling, which is a variation of the Proportional Fair (PF) algorithm (the most widely adapted opportunistic scheduling technique), in the presence of interference. It is observed from simulation results that the proposed scheme provides better fairness in terms queuing delays, and dropped packets for various loading factors while the throughput remains comparable. A gain in the performance of cell edge users is also observed for the proposed scheme, this may result in substantial savings in the deployment cost since fewer base stations (BS) will be needed to cover regions.**

Keywords: Scheduling, Radio Resource Management, QoS, Cross layer.

I. Introduction

The last decade has witnessed a tremendous growth in the wireless market. To meet the growing demands in the number of subscribers, rates required for high speed data transfer and multimedia applications 3G (third generation) standards started evolving. Now, the approaching 4G (fourth generation) wireless communication networks are projected to solve the still-remaining problems of 3G networks to provide a wide variety of new services ranging from highquality voice to other high-data-rate wireless applications. One such example of a 4G wireless network approach is being developed in the WINNER (Wireless World Initiative New Radio) project. The key objective of the WINNER project is to develop an innovative concept in radio access in order to address high flexibility and scalability with respect to data rates and radio environments [1].

Compared to wireline networks, wireless resources are very scarce. While more wired network bandwidth is created when new physical resources (cable, fiber, router, etc.) are added to the network, wireless communications require the sharing of a finite natural resource: the radio frequency spectrum. The data-rate capacity that a radio frequency channel can support is limited by Shannon's capacity laws. Hence, the allocation, management and scheduling of radio resources are crucial for wireless networks.

Traditionally, the research on packet scheduling has focused mostly on QoS and fairness for different QoS classes or different applications, while opportunistic scheduling algorithms have focused on exploiting the timevarying nature of the wireless channels in order to maximize throughput. This segregation between packet scheduling and radio resource scheduling is not efficient since none of the two types of scheduling algorithms focus both on providing QoS for the applications and exploiting the time-varying characteristics of the wireless channel. For these reasons, it is necessary to merge the scheduling of packets and the allocation of radio resources to design cross-layer scheduling algorithms [2].

In order to improve the QoS experienced by the wireless users, cross-layer scheduling algorithms need to take both the time-varying characteristics of the wireless channels and the QoS demands of the applications into account. In addition, it is often necessary to consider the characteristics of the packet load of the queues at the mobile users or the BS containing packets waiting to be transmitted over the uplink or downlink, respectively [3]. There is considerable work done on scheduling which analyzes Physical and Medium Access Control (MAC) related design issues by assuming that all the users are back-logged, i.e., that all the users in the system have nonempty buffers that always contain packets to send or receive. Such scheduling algorithms are termed as non-queue aware algorithms. However in [4], it is shown that when analyzing the QoS performance of scheduling algorithms this assumption is not always correct since the number of packets in the buffers can vary significantly, and there is a relatively high probability that the buffers are empty. For example, in time-slotted networks, the packets in the queues are aggregated into time-slots. Consequently, empty queues and partially filled

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada under participation in the Wireless World Initiative New Radio (WINNER) project – www.ist-winner.org.

time-slots will affect the system performance.

In the recent years some publications have considered how to integrate packet scheduling and radio resource scheduling into queue-aware, channel-aware scheduling algorithms [4][5][6]. For example, one such publication handles how to implement Weighted Fair Queuing (WFQ) when the largest share of the radio resources is given to the users with the instantaneously best channel conditions in a (Code Division Multiplexing)CDM-based network [6]. Another popular approach of channel-aware scheduling algorithm is the Modified Largest Weighted Delay First (M-LWDF) algorithm where priorities are given to the users with maximum queuing delays weighted by their instantaneous and average rates [5].

The objective of this study is to develop a queue-aware scheduling scheme which takes QoS parameters into account for all users. For this purpose the Token Bank Fair Queuing (TBFQ) algorithm, which was originally proposed for single carrier Time Division Multiple Access (TDMA) systems [7], is modified. TBFQ is a credit based scheme which selects users by assigning them priorities based on interflow fairness and delay constraints. By selecting the wireless users in a certain prioritised manner, it is made sure that flows belonging to users suffering from bad interference conditions and shadowing in particular will have a higher priority. The TBFQ algorithm was designed to accommodate the bursty nature of traffic. This is done by the graceful acceptance of traffic profile violation when exceeding bandwidth is available, provided that the session does not exceed its bandwidth allocation in the long term. This prevents sudden degradation of QoS experienced by the end user as a result of traffic profile violations. The modified algorithm is called Adaptive Token Bank Fair Queuing (ATBFQ).

The performance of the ATBFQ algorithm is compared to that of the SB algorithm which is the current baseline scheduler in WINNER. The SB algorithm was proposed in [8]. It is a variation of the PF algorithm which is the most widely adopted opportunistic scheduling algorithm [9]. The SB scheduler selects at slot *k* the user *i* with the best score where the score is calculated based on the current rank of the user's signal to interference noise ratio (SINR) among its past values of SINR in the current window $\{\gamma_i(t_k), \gamma_i(t_{k-1}), ..., \gamma_i(t_{k-W+1})\}$,where $\gamma_i(t_k)$ is the SINR value of a user at time instant *k* and W is the window size. The corresponding score for the user *i* will be given by

$$s_i(t_k) = 1 + \sum_{l=1}^{W-1} 1_{\{r_i(t_k) < r_i(t_{k-l})\}} + \sum_{l=1}^{W-1} 1_{\{r_i(t_k) = r_i(t_{k-l})\}} X_l, \quad (1)$$

where X_l are i.i.d random variables on $\{0, 1\}$ with P_r(X=0)= P_r(X=1)=0.5 The performance of this modified scheme is studied in the context of the 4G WINNER system. A simulation model for the downlink is built adherent to specifications of this system:

1) *A traffic model which realistically models the burstiness of the video streaming service class.*
2) *An inter-cell interference model which takes the interference from dominant interferes into consideration.*
3) *A channel model which accurately depicts the large scale path loss, shadowing and fading for a micro-cell urban environment.*
4) *A modified version of the TBFQ algorithm for the multicarrier WINNER system.*
5) *An adaptive coding and modulation (AMC) technique.*

II. ATBFQ Scheduling Algorithm

A. Generic TBFQ Algorithm

The Token Bank Fair Queuing (TBFQ) algorithm was initially developed for wireless packet scheduling in the downlink channel [7][10] and was later modified for wireless multimedia services using uplink as well. Its concept was based on the leaky bucket mechanism which polices flows and conforms them to a certain traffic profile.

Each traffic flow *i* is characterized by the following parameters:

λ_i : Packet arrival rate

r_i : Token generation rate

P_i : Token pool size

E_i: Counter that keeps track of the number of tokens borrowed from or given to the token bank by flow *i*

Each *L*-byte packet consumes *L* tokens. For each flow *i*, E_i is a counter that keeps track of the number of tokens borrowed from or given to the token bank. As tokens are generated at rate r_i, the tokens overflowing from the token pool are added to the token bank, and E_i is incremented by the same amount. When the token pool is depleted and there are still packets to be served, tokens are withdrawn from the bank by flow *i*, and E_i is decreased by the same amount.

Thus, during periods when the incoming traffic rate of flow *i* is less than its token generation rate, the token pool always has enough tokens to serve arriving packets, and E_i becomes positive and increasing. On the other hand, during periods when the incoming traffic rate of flow *i* is greater than its token generation rate, the token pool is emptied at a faster rate than it can be refilled with tokens. In this case, the connection may borrow tokens from the bank. The priority of a connection in borrowing tokens from the bank is determined by the priority index (P_i) given by

$$P_i = \frac{E_i}{r_i}. \quad (2)$$

By assigning the priority in this manner, we can make sure that flows belonging to user terminals (UTs) suffering from bad interference conditions and shadowing in particular will have a higher priority index as they will be contributing to the bank more often.

B. ATBFQ Algorithm

The original TBFQ algorithm was proposed for TDMA single carrier systems. In this study, this has been modified according to WINNER which is a multicarrier system where the multiple access technique used is Orthogonal Frequency

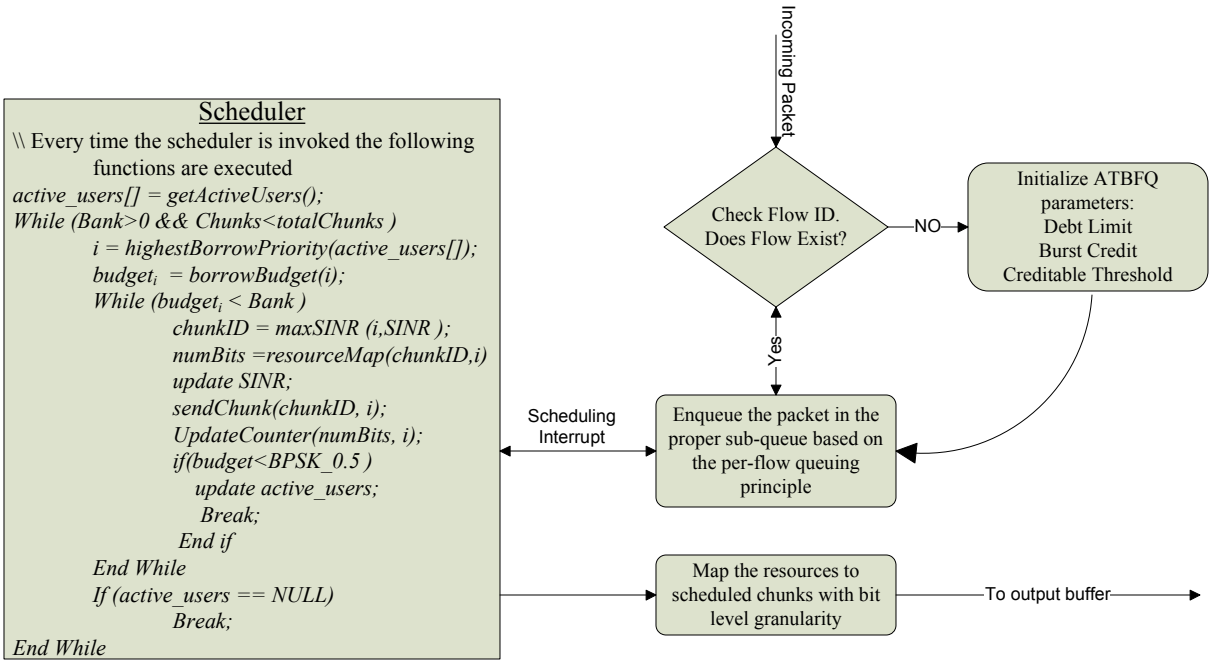

Fig. 1. Overview of scheduling operation

Division Multiple Access (OFDMA) [11]. The basic time-frequency resource unit in OFDM links is denoted as a chunk. It consists of a rectangular time-frequency area that comprises a number of subsequent OFDM symbols and a number of adjacent subcarriers. Packets from the traffic flows are exclusively mapped on to these chunks.

Each time a packet is generated, the scheme check to see whether if it belongs to an already existing flow. If it belongs to a new flow, then ATBFQ parameters are initialized. Based on the service class type, a debt limit, burst credit, and the creditable threshold are set. The value for these parameters varies from one service class to another. The packets are then queued in subqueues in a manner such that each subqueue belongs to a particular flow.

The operation of the ATBFQ scheduler is shown by the flowchart shown in Fig. 1. This can be summarized by the following functions which are executed each time the scheduler is invoked at the beginning of the frame.

1) At the scheduler, information is retrieved from the higher layer about all active users using the *getActiveUsers()* function. An active user is defined as a backlogged queue which has packets waiting to be served.
2) Based on this list of active users, a priority is calculated given by the priority index in 2. The *highestBorrowPriority()* function is called to calculate this for all active users N_{act}. This function then returns the user *i* with the highest priority given by:

$$i^*(t_k) = \underset{1 \le i \le N_{act}}{\arg\max} \left(P_i\right). \tag{3}$$

3) Using the *borrowBudget()* function, a certain budget is calculated for user *i* based upon the amount of tokens it has contributed to the bank and the debt limit it has incurred from the previous rounds of scheduling.
4) If the calculated budget is less than the bank size, resources are allocated to the user *i* using the *maxSINR()* function. This is the second level of scheduling and this deals with allocation of chunk resources to the selected user *i*. This allocation is based on the Maximum SINR principle where the chunk *j* with the best SINR is given to the selected user [12]:

$$j^*(t_k) = \underset{1 \le j \le N_{chunks}}{\arg\max} \left(\gamma_{ij}(t_k)\right), \tag{4}$$

where γ_{ij} is the SINR of the selected user *i* in chunk *j*. This is the most opportunistic of all scheduling algorithms for time-slotted networks. This means that the AMC policy maximally exploits the multiuser diversity in a time-slotted cell where only one user is scheduled at a time.
5) The *resourceMap()* function determines the amount of bits that can be mapped to the chunk depending on the type of modulation and coding used.
6) Each time a chunk resource is allocated, the *updateCounter()* function is called. This function updates the bank, the counter *Ei* and the allocated budget.

The selected user *i* gets to transmit as long as its queue is backlogged and the allocated budget is less than the total bank size and more than the number of bits that can be supported for the smallest modulation and coding scheme (for this study, it is BPSK rate as shown in Table 1).

If either of these conditions is not satisfied, then the user

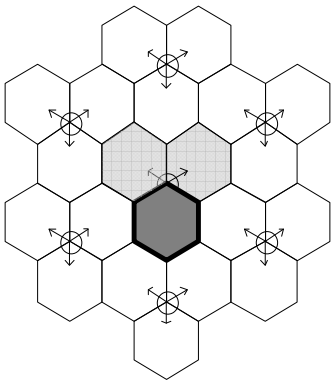

Fig. 2. Network layout

is classified as non-active. A new priority is calculated on the updated active users and Steps 1-6 are repeated. This procedure is carried till there are no chunk resources available or there are no active users.

III. System Model

This scheme is studied in the wide area down link scenario. To reduce the simulation complexity, the bandwidth is reduced to 15 MHz form the original 45 MHz. The chunk dimensions are given as 8 subcarriers by 12 OFDM symbols or 312.5 kHz $\times$ 345.6μs. The frame duration is defined as 0.6912 μs, i.e., there are a total of 96 chunks per frame.

The network layout is shown in Fig. 2. Each cell in the network has three sectors. In the simulation we are only consider the effect of interference on the central cell. For this purpose the interference form the first tier will be taken into account. We also assume a frequency re-use factor of one in each sector (all resources are used in each sector). The UTs are uniformly placed in the central sector.

A. Channel Model

Time and frequency correlated Rayleigh channel samples obtained from power delay profile for WINNER wide area scenario are used to generate the channel fading. The user speed is defined to be 70 km/hr and the inter site distance is 1 km. The following exponential path-loss (*PL*) model has been used [13]:

$$PL = 38.4 + 35.0 \log_{10}(d)[dB], \tag{5}$$

where d is transmitter-receiver separation in meters.

B. Background Noise and Shadowing

The average thermal noise power is calculated with a noise figure of 5 dB. We have considered independent lognormal random variables with a standard deviation of 8 dB for shadowing. Sector transmit power is assumed to be 39.81 Watts and chunks are assigned fixed equal powers.

C. Interference Model

The interference model used is obtained by considering the inter-cell interference from the surrounding BS, and the intra-cell interference from the remaining two sectors of the central cell. Different levels of interference can be defined by a certain activity factor (AF). An AF is defined as the percentage of active interferers. For example, AF=1 denotes a high level of interference where all the links are being interfered at (100% interference). For a link of interest in the concerned sector in the central cell, the interference will comprise of 18 inter-cell (6 BS x 3 sectors) and 2 intra-cell links. In this study AF=0.7 case is considered.

D. Adaptive Coding and Modulation

Adaptive modulation with block low-density parity-check (B-LDPC) code is used. Thresholds for transmission schemes are determined assuming a block length of 1704 bits and 10% block error rate (BLER) as shown in Table I [14]. A chunk using QPSK rate 1/2 can carry 96 information bits. This is based on the initial transmissions, i.e. HARQ retransmissions are not considered.

E. Traffic Model

For this study, real-time video streaming traffic is used. Two *Interrupted Renewal Process* (2IRP) sources are superimposed to model user's video traffic in the downlink transmission as indicated in [15]. In each of these sources, modeled traffic is bursty in nature as both have separate ON and OFF distributions. During the OFF state the IRP process does not generate packets. During the ON state, packets are generated with exponentially distributed inter-arrival time.

The average packet rate of one 2IRP generator is 1263.8 packets per second. The resulting downlink data rate for each user is 1.92 Mbps.

TABLE I
Lookup Table for AMC Modes and the Corresponding Chunk Sizes

AMC Mode	SINR (dB)	Chunk Size (bits)
BPSK 1/2	0.2311 ≥ SINR > -1.7	48
BPSK 2/3	1.231 ≥ SINR > 0.231	72
QPSK 1/2	3.245≥ SINR > 1.231	96
QPSK 2/3	4.242≥ SINR > 3.245	128
QPSK 3/4	6.686 ≥ SINR > 4.242	144
16QAM 1/2	9.079 ≥ SINR > 6.686	192
16QAM 2/3	10.33 ≥ SINR > 9.079	256
16QAM 3/4	14.08≥ SINR > 10.33	288
64QAM 2/3	15.6≥ SINR > 14.08	384
64QAM 3/4	SINR > 15.6	432

Packets are scheduled in a frame by frame basis at the start of every frame. Any packet arrives at current frame time will have to wait at least until the start of the next frame. Since the traffic generated is of video streaming type, packets are dropped if they face a delay in excess of 190 ms. The ATBFQ parameters are summarized in the Table III [14].

TABLE II
ATBFQ PARAMETERS

ATBFQ Parameter	Used Value
Debt limit	80,000,000 bits = 9.54 Mb
Burst Credit	50000 bits
Token Generation Rate (r_i)	5.43 Mbps (3 **x** video source)

TABLE III
SUMMARY OF SIMULATION PARAMETERS

Parameter	Used Value/Model
Scenario	Wide Area DL
Channel model	WINNER C2 channel
Sector Tx antenna	120^0 directional with baseline antenna pattern
UT receive antenna	Omni-directional
Signal bandwidth	15 MHz (i.e., 48 chunks which is 1/3rd of the WINNER baseline assumptions)
Frequency re-use	1
Mobility	70 km/hr
Sector Tx power	46 dBm
Antenna configuration	SISO
Coding	B-LDPCC
Frame duration	0.6912 ms (This is also the scheduling interval)
Traffic model	1.9Mbps 2IRP model for MPEG video
Packet size	188 Bytes

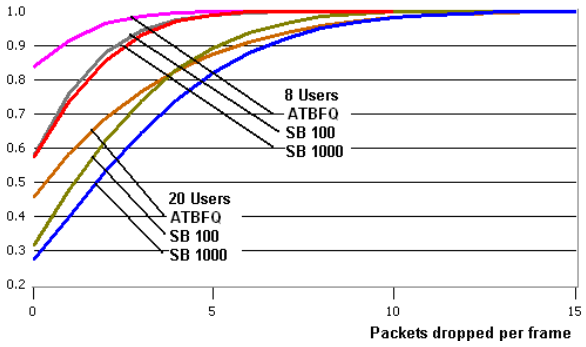

Fig. 3. CDF of packets dropped per user per frame

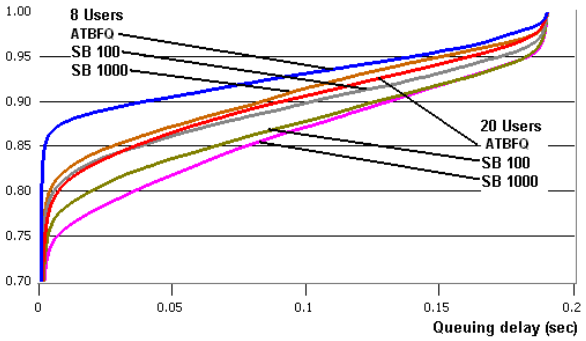

Fig. 4. CDF of user queuing delay

IV. RESULTS

In this section, performance measures such as average packets dropped, average user throughput and average user queuing delays are shown and compared to those of the SB algorithm. As shown in [8], window size plays an important role in the performance of the SB algorithm (larger window size leads to a higher complexity). Therefore, two different window sizes of 100 and 1000 are considered in the results.

Fig. 3 shows the cumulative density function (CDF) of the packets dropped per frame for low and high loading, respectively. These curves indicate the opportunistic nature of SB as it only tends to favor the users with the better channel conditions. Thats why we observe a higher drop rate even at low loading.

Fig. 4 shows the CDF of the average user queuing delay for low (8 users) and high (20 users) loading. The performance of ATBFQ is compared to that of SB at AF=0.7. For low loading we observe that ATBFQ faces minimal queuing delay till the 90th percentile mark, whereas for SB, this mark is around 80th percentile. We observe for high loading that although the performance degrades for both ATBFQ and SB as expected, ATBFQ still outperforms SB. For SB, we observe at the 95th percentile point that users constantly face a delay of 190 msec; this indicates that the cell edge users suffer continuous packet losses.

The CDF of average user throughput (measured in bytes per frame) is shown in Fig. 5 for 8 and 20 users. ATBFQ performs better for the lower loading case whereas SB achieves higher throughput at higher loading. For the high loading case we also observe that the curve for ATBFQ has a steeper slope. This indicates fairness as more users are serviced with similar throughput. Note that this is not the case for SB. As ATBFQ attempts to maintain fairness, it tries to service cell-edge users who are receiving poorer channel conditions as compared to those users which are located closer to the BS. Therefore it also utilizes more chunks. On the other hand SB tries to maximize the throughput.

Fig. 6 shows the packet transmit ratio (defined as the packets transmitted/total packets) vs. distance from BS for 20 users. It can be observed that as the distance increases, the packet transmit ratio for SB decreases ,i.e., the number of dropped packets increases. This can be further visualized by the quadratic-fitted curves for both algorithms which show their respective trends with the varying distance. As SB tries to maximize the throughput, the cell edge users are affected and suffer packet losses. ATBFQ, on the other hand, is fair in nature and tries to also look after the cell edge users. If a cell edge user is suffering from bad channel conditions, ATBFQ gives it priority to transmit in the next scheduling interval. By assigning priorities in such a manner, ATBFQ also keeps track of the queue levels and tries to maintain constant queuing delay for the cell edge users as shown in Fig. 7.

V. CONCLUSION

In this study, the ATBFQ algorithm was modified to satisfy the requirements for the WINNER 4G concept. It is a

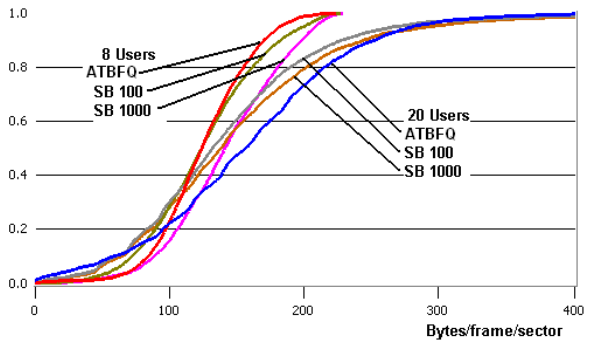

Fig. 5. CDF of user throughput

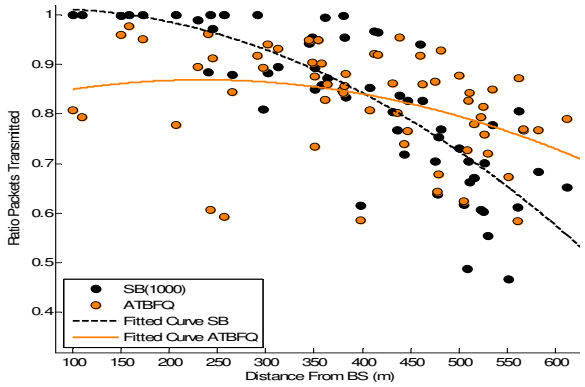

Fig. 6. Ratio of packets dropped vs. distance form BS

queue-aware scheduling algorithm which attempts to maintain fairness among all users. Performance of the modified ATBFQ is shown with reference to the SB scheduler for two different window sizes. SB is an opportunistic scheduler belonging to the proportional fair class. It tries to maximize throughput, making use of multiuser diversity while trying to maintain fairness. But this comes at a certain cost as the cell edge users in this scheme suffering from bad channel conditions are more severely affected. Also due to the bursty nature of the modeled traffic, such users face higher queuing delays which result in higher packet drops.

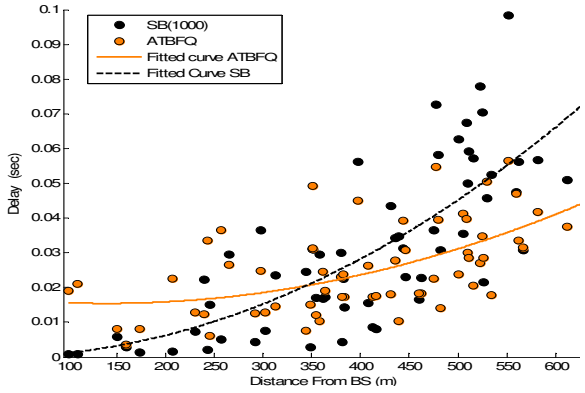

Fig. 7. Average user queuing delay vs. distance form BS

Compared to SB, ATBFQ is a credit based scheme which tries to accommodate the burstiness of the users by assigning them more resources in the short term provided that long term fairness is maintained. For lower to medium loading, ATBFQ performs better than SB in terms of throughput, queuing delay and packet dropping. At high loading ATBFQ still performs better than SB with regards to the queuing delay and packet dropping while the sector throughput slightly drops as ATBFQ attempts to satisfy its users by attempting to assign more resources at a lower spectral efficiency. An overall increase in the performance of cell edge users is observed in terms of queuing delay and packet drop ratio for ATBFQ as compared to SB.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Mr. Mahmudur Rahman and Mr. Jiangxin Hu for their technical support. They also thank OPNET Technologies Inc. for providing software license to carry out the simulations of this research.

REFERENCES

[1] *Project Presentation*, WINNER Deliverable D8.1, Mar. 2004. [Online]. Available: https://www.ist-winner.org

[2] Q. Liu, X. Wang, and G. B. Giannakis, "A cross-layer scheduling algorithm with qos support in wireless networks," *IEEE Trans. on Vehicular Technology*, vol. 55, pp. 839–847, May 2006.

[3] D. Wu and R. Negi, "Effective capacity: A wireless link model for support of quality of service," *IEEE Trans. on Wireless Communications*, vol. 2, pp. 630–643, July 2003.

[4] S. Borst, "User-level performance of channel-aware scheduling algorithms in wireless data networks," *in Proc. IEEE Joint Conference of the Computer and Communications Societies (INFOCOM03)*, vol. 1, pp. 321–331, March-April 2003.

[5] M. Andrews, K. Kumaran, K. Ramanan, A. Stolyar, P. Whiting, and R. Vijayakumar, "Providing quality of service over a shared wireless link," *IEEE Communications Magazine*, p. 150154, Febuary 2001.

[6] A. Stamoulis, N. D. Sidiropoulos, and G. B. Giannakis, "Time varying fair queueing scheduling for multicode cdma based on dynamic programming," *IEEE Trans. on Wireless Communications*, vol. 3, pp. 512–523, March 2004.

[7] W. K. Wong and V. C. M. Leung, "Scheduling for integrated services in next generation broadcast networks," *in Proc. IEEE Wireless Communications and Networking Conference (WCNC99)*, September 1999.

[8] T. Bonald, "A score-based opportunistic scheduler for fading radio channels," *European Wireless Conference*, Febuary 2004.

[9] E. F. Chaponniere, P. J. Black, J. M. Holtzman, and D. N. C. Tse, "Transmitter directed code division multiple access system using path diversity to equitably maximize throughput," *U.S. Patent 6449490*, September 2002.

[10] W. K. Wong, H. Y. Tang, and V. C. M. Leung, "Token bank fair queuing: a new scheduling algorithm for wireless multimedia services," *International Journal of Communication Systems*, vol. 17, pp. 591–614, 2004.

[11] *Final report on identified RI key technologies, system concept, and their assessment*, WINNER I Deliverable D2.10, Nov. 2005. [Online]. Available: https://www.ist-winner.org

[12] R. Knopp and P. Humblet, "Information capacity and power control in single cell multiuser communications," *in Proc. IEEE Int. Conference on Communications (ICC'95)*, pp. 331–335, June 1995.

[13] *Final report on link level and system level channel models*, WINNER I Deliverable D5.4, Nov. 2005. [Online]. Available: https://www.ist-winner.org

[14] *Test Scenarios and Calibration Cases Issue 2*, WINNER II Deliverable D6.13.7, Dec. 2006. [Online]. Available: https://www.ist-winner.org

[15] *Traffic model for 802.16 TG3 MAC/PHY simulations*, IEEE 802.16 Work-in-progress document 802.16.3c-01/30r1, Mar. 2001. [Online]. Available: http://ieee802.org/16