

A comparison of scheduling algorithms in HSDPA

Stefan M. Scriba and Fambirai Takawira

School of Electrical, Electronic and Computer Engineering, University of KwaZulu-Natal

King George V Avenue, Durban, 4041, South Africa

E-mail: scribas@gmail.com or ftakaw@ukzn.ac.za

Tel: +27-31-260-2730

Abstract—This paper compares the behaviour of Earliest Deadline First (EDF) and Opportunistic-EDF (O-EDF) with that of Round Robin, Proportional Fair Throughput (PFT), and Maximum Carrier to Interference ratio (MaxC/I). Video, voice and web traffic are transmitted through an HSDPA air interface. The resultant queuing delay and fairness are measured and compared for each scheduler.

Index Terms—Earliest Deadline First (EDF), queuing delay.

I. INTRODUCTION

As part of the UMTS standard, the third generation partnership project (3GPP) included High Speed Downlink Packet Access (HSDPA) in Release 5. In HSDPA, all resources are usually made available to a single user on a Transmission Time Interval (TTI) basis [1]. HSDPA distinguishes itself from the previous WCDMA standard, by using an Adaptive Modulation and Coding (AMC) scheme, which enables it to respond rapidly to channel fluctuations, without the need for fast power control, which is disabled in HSDPA. Furthermore, Variable Spreading Factors have also been disabled, while the TTI has been decreased from 10ms to 2ms, allowing for much faster scheduling responses. Finally, the model makes use of a Fast Physical Layer Hybrid ARQ scheme.

This paper compares the behaviour of EDF and Opportunistic-EDF (O-EDF) with that of Round Robin, PFT, and MaxC/I. Video, voice and web traffic are transmitted through an HSDPA air interface. This is achieved by means of a custom built simulator, which shows how the Earliest Deadline First (EDF) scheduler fills an essential role in providing QoS for delay sensitive traffic.

II. TRAFFIC GENERATION MODEL

The simulation model assumes that multiple sources are producing voice, video, and web traffic that will arrive at the base station, enter the queuing system and is then transmitted via a channel to the mobile stations. Note that each physical mobile unit can simultaneously generate voice, video and web traffic, in other words, it can consist of several traffic sources. The same traffic generating models for video and web traffic will be used as [1], while the video generating model will be extended to give a realistic voice generating model. The scheduling intervals are defined in terms of the Transmission

Time Interval (TTI). In other words, a new packet is scheduled every TTI. In an HSDPA system, the TTI is fixed at 2ms.

A. Video traffic source

A video source can be modeled as M independent ON-OFF Markov mini-sources. As was the case in [1], this model assumes that $M = 10$. In the ON state, a mini-source produces a constant rate of V bps, while in the OFF state, a mini-source produces no traffic. Each mini-source spends a mean time of p TTIs in the ON state and q TTIs in the OFF state, where their respective random variables P and Q are geometrically distributed.

Parameters p , q and V are obtained as follows:

$$p = \frac{1}{a \cdot TTI} \left(1 + \frac{\mu^2}{M\sigma^2} \right) \quad [\text{TTIs}] \quad (1)$$

$$q = \frac{1}{a \cdot TTI} \left(1 + \frac{M\sigma^2}{\mu^2} \right) \quad [\text{TTIs}] \quad (2)$$

$$V = \frac{\mu}{M} + \frac{\sigma^2}{\mu} \quad [\text{bps}] \quad (3)$$

A video source has a mean bit rate of μ bps, a standard deviation of σ bps, and exponent of the auto-covariance with coefficient $a s^{-1}$. As was the case in [1], $\mu=128$ kbps, $\sigma=8$ kbps and $a=3.9 s^{-1}$. Video packets are chosen to have a fixed packet length of 1500 Bytes = 12kb. Video traffic is considered to have a deadline of $d=100$ ms, after which it will be dropped.

The result is that the mean ON-time is

$$p = 3410.26 \text{ TTI} = 6,820,513 \mu s \quad (4)$$

Similarly, the mean OFF-time is

$$q = 133.21 \text{ TTI} = 266,420 \mu s \quad (5)$$

And finally, the transmission rate of each mini-source is

$$V = 13,300 \text{ bps} \quad (6)$$

B. Voice traffic source

For the voice traffic generation, exactly the same model is used as for video traffic generation, except that $M=1$. In other words, every source no longer consists of 10 mini-sources. For voice traffic, $\mu=64,000$ bps and $\sigma=4,000$ bps. Voice packets have a fixed packet length of 80 Bytes = 640b. Voice traffic has a deadline of $d=50$ ms, after which it will be dropped.

Once again, the mean ON-time is calculated to be

$$p = 32,948.7 \text{ TTI} = 65,897,436 \text{ } \mu\text{s} \quad (7)$$

Similarly, the mean OFF-time is given by

$$q = 128.71 \text{ TTI} = 257,412 \text{ } \mu\text{s} \quad (8)$$

And finally, the source transmission rate is

$$V = 64,250 \text{ bps} \quad (9)$$

C. Web traffic source

Although it is fairly difficult to find a statistical distribution that accurately describes web traffic generation patterns, creating simulated web-traffic is fairly straight forward. We use the same model as in [1], where an end-user can be seen to oscillate between two states while browsing. He is either requesting a new webpage or reading. The reading state has a geometric distribution with an average duration of $T_{OFF}=1000\text{TTIs} = 2\text{s}$. During this time no traffic is generated. Web traffic has a virtual deadline of $d=500\text{ms}$. It is virtual because the packet is still served and not dropped when the delay exceeds d and creates a deadline violation.

When in the request state, the addressed data source will produce a geometrically distributed number of packets, with a mean of $\bar{P}=300$ packets. The packet inter-arrival time too is geometrically distributed, with a mean of $\Delta T=200\text{TTIs} = 0.4\text{s}$. The packet length L can be found by finding the floor of a truncated Pareto pdf, in other words, $L = \lfloor x \rfloor$, where:

$$p(x) = \frac{\zeta \cdot l_{\min}^{\zeta}}{x^{\zeta+1}} [u(x - l_{\min}) - u(x - l_{\max})] + \delta(x - l_{\max}) \left(\frac{l_{\min}}{l_{\max}} \right)^{\zeta} \quad [\text{bits}] \quad (10)$$

Here $\zeta = 1.1$ is a constant, $u(\cdot)$ is the unitary step function, $\delta(\cdot)$ is the Dirac Delta function, while $l_{\min} = 80 \text{ Bytes} = 640\text{b}$ and $l_{\max} = 1500 \text{ Bytes} = 12\text{kb}$ are respectively the minimum and maximum message lengths. The values of l_{\min} and l_{\max} are different to those proposed in [1] and were chosen to lie within the bounds of the packet lengths of voice and video packets.

The problem with generating web traffic is that the IMSL library that was used to perform statistical tasks does not include a function for generating Pareto distributed random numbers. To solve this problem, it was noted that the Pareto density function is given by

$$p(x) = \frac{ab^a}{x^{a+1}} \quad [\text{bits}], \quad (11)$$

while the Pareto cumulative distribution is given by

$$F(x) = 1 - \left(\frac{l_{\min}}{x} \right)^{\zeta} \quad [\text{bits}]. \quad (12)$$

Pareto distributed random numbers can be obtained, by generating random numbers for a uniform distribution with limits $[0,1]$. For every uniformly distributed random u generated in this fashion, $F(x) = u$. One may then solve for x , as follows:

$$x = \frac{l_{\min}}{(1-u)^{1/\zeta}} \quad [\text{bits}]. \quad (13)$$

The result is that one is able to translate every uniformly distributed random number u into a Pareto distributed random number x .

III. QUEUEING MODEL

To be able to compare all schedulers, identical conditions must be created in each case. The only difference may be the way the data is scheduled, in other words, the order in which it is transmitted. As described earlier, traffic that has violated its delay deadline is either physically dropped or virtually dropped, depending on whether it is real-time traffic (voice and video) or best-effort traffic (web). Because the real-time traffic can be dropped, its queue-length is indirectly limited to the product of the average service rate and the delay deadline. Under high enough load conditions, the best-effort queue could, on the other hand, grow to a significant size, which in turn drastically affects its delay behaviour. Limiting the queue size in any way would also indirectly cap the possible maximum delay. The result is that infinitely long buffers were chosen for all queues.

IV. CHANNEL MODEL

HSDPA uses Direct-Sequence Code Division Multiple Access (DS-CDMA). The bit energy-to-interference spectral density $\frac{E_b}{I_0}$ that the User Equipment (UE) measures is fed back to Node-B using 5 bits, known as the Channel Quality Indicator (CQI) value. The CQI value is used to decide on the highest transmission rate that can be chosen.

The $\frac{E_b}{I_0}$ value that the UE measures, can be modeled as follows [1]:

$$\frac{E_b}{I_0} = \text{SIR} \cdot \text{PG}, \quad (14)$$

where SIR is the signal-to-interference power ratio, PG is the DSSS Processing Gain, which is defined by the ratio $\frac{W}{R_b}$, where W is the spreading bandwidth and R_b is the bit-rate that depends on the used modulation and coding scheme (MCS). The SIR value can be estimated as

$$\text{SIR} = \frac{P_{RX}}{I_{Inter} + I_{Extra} + N}, \quad (15)$$

where P_{RX} is the useful received power, I_{Inter} is the intra-cell interference, I_{Extra} is the extra-cell interference and N is the noise power. Note that P_{RX} , I_{Inter} , I_{Extra} and N can be characterized as follows [1]:

$$P_{RX} = P_{TX} G_c^{P_{RX}}, \quad I_{Inter} = \alpha G_c^{P_{Inter}} P_D, \quad (16)$$

$$I_{Extra} = \varepsilon G_c^{I_{Extra}} P_D, \quad N = N_0 W \quad (17)$$

where P_{TX} is the transmitted power to the user, $G_c^{P_{RX}}$, $G_c^{P_{Inter}}$, and $G_c^{I_{Extra}}$ are the user channel gains of the

transmission, the intra-cell interference, and the extra-cell interference, respectively. P_D is the Node B available power, N_0 is the noise spectral density, α is the intra-cell non-orthogonality coefficient and ε is the extra-cell interference power-to-the total received power ratio. Note that all powers must be converted to Watts and cannot be left as dBs.

P_{TX} can be found quite simply, by assuming that all of Node B's power will be used during transmissions. Node B's power is therefore divided amongst each of the transmission codes. If one assumes that 15 codes are used for transmission, then P_{TX} becomes:

$$P_{TX} = \frac{P_D}{15} \quad [\text{W}]. \quad (18)$$

Hence

$$\frac{E_b}{I_0} = \frac{P_{TX} G_c^{P_{RX}} W / R_b}{(\alpha G_c^{P_{Inter}} + \varepsilon G_c^{P_{Extra}}) P_D + N_0 W} \quad (19)$$

per CDMA code.

From [2], one can derive that the channel gain for the transmission, the intra-cell interference, and the extra-cell interference are given by:

$$G_{c(\text{dB})} = -L(t)_{(\text{dB})} + s(t)_{(\text{dB})}, \quad [\text{dBW}] \quad (20)$$

where $s(t)$ is the shadowing component, which will be discussed later. $L(t)$ is the path-loss component of a UHF signal over flat terrain (in a semi-urban environment), which is given by [2]:

$$L(t)_{(\text{dB})} = K + \gamma \log_{10} r \quad \text{dBW}. \quad (21)$$

Here K is the signal strength measured at 1km from the base-station, with r in km, while γ is the rate at which $L(t)_{(\text{dB})}$ changes as r changes. Both K and γ may vary as distance r between a mobile and the base-station changes.

The shadowing term $s(t)$ (in dB) is usually modeled as a zero-mean stationary Gaussian process, which when expressed in Watts has a log-normal pdf:

$$f_{s(t)} = \frac{1}{\sigma \sqrt{2\pi t}} \cdot \exp \left[-(\ln t - \mu)^2 / (2\sigma^2) \right] \quad (22)$$

$$= \frac{1}{\sigma \sqrt{2\pi t}} \cdot \exp \left[-\frac{\ln^2 t}{2\sigma^2} \right], \quad (23)$$

since $\mu = 0$.

V. HSDPA TRANSMISSION RATES

When a packet is selected for transmission to the j -th user, the bit-error-rate (BER) plays an important role. The applications using the various classes of service have different sensitivities to the BER.

The BER thresholds that were used in this paper are $9.6 \cdot 10^{-6}$ for video and voice traffic, and $8.4 \cdot 10^{-7}$ for web traffic, as listed in Table II in Section VII. These values were chosen to correspond to a packet-error-rate (PER) proposed in [1] of 10^{-1} for video traffic and a mean PER of 10^{-2} for web traffic, which varies as the packet length of the web packets vary. The BER of the voice traffic is kept the same as video,

TABLE I
HSDPA MCS MODES

MCS mode m	Data rate (15 codes) R_b
1. QPSK, rate $\frac{1}{4}$	1.8Mbps
2. QPSK, rate $\frac{1}{2}$	3.6Mbps
3. QPSK, rate $\frac{3}{4}$	5.3Mbps
4. 16QAM, rate $\frac{1}{2}$	7.2Mbps
5. 16QAM, rate $\frac{3}{4}$	10.7Mbps

but corresponds to a much lower PER, as can be calculated using the following expression:

$$PER = 1 - (1 - BER)^L, \quad (24)$$

where L in this context is the packet length, measured in bits.

The BER can be controlled directly by varying the transmission power. In the 3GPP's UMTS standard, power control therefore plays an important role. One of the improvements that HSDPA offers, is to track a BER by varying the modulation and coding scheme (MCS) as the channel conditions change. This enables the base station to keep its transmission power constant, simplifying the design.

In order to obtain high transmission rates, 15 codes were used with the MCS modes listed in Table I. As the channel quality varies, a different MCS will have to be used, which effectively varies the transmission rate.

To determine which of these MCS modes should be used at any time, the indirect relationship between BER and transmission rate needs to be considered. Note that the BER is related to the bit-energy-to-interference spectral density $\frac{E_b}{I_0}$ value, which is communicated from the User Equipment (UE) to the transmitting Node-B using the Channel Quality Indicator (CQI) value [3]. This 5-bit CQI value, ranging from 0 to 30, is calculated as follows:

$$CQI = \begin{cases} 0 & \text{if } \frac{E_b}{I_0} \leq -16, \\ \left\lfloor \frac{E_b}{I_0} / 1.02 + 16.62 \right\rfloor & \text{if } -16 < \frac{E_b}{I_0} \leq 14, \\ 30 & \text{if } 14 < \frac{E_b}{I_0}. \end{cases} \quad (25)$$

When the CQI value is received by Node-B, it can be converted back to an estimate of $\frac{E_b}{I_0}$.

In the previous section, the following expression was developed that explains the relationship between the transmission rate R_b and the resulting $\frac{E_b}{I_0}$ that can be expected:

$$\frac{E_b}{I_0} = \frac{P_{TX} G_c^{P_{RX}} W / R_b}{(\alpha G_c^{P_{Inter}} + \varepsilon G_c^{P_{Extra}}) P_D + N_0 W} \quad (26)$$

By choosing the appropriate MCS value, the transmission rate can be increased while the resulting $\frac{E_b}{I_0}$ still meets the following condition:

$$\frac{E_b}{I_0} > \left(\frac{E_b}{I_0} \right)_{\text{Threshold}}. \quad (27)$$

The $\left(\frac{E_b}{I_0} \right)_{\text{Threshold}}$ is a threshold that is directly related to the maximum PER threshold that each class of traffic can endure.

Reference [4] contains the required SNR-to-PER curves in a flat fading Rayleigh model for QPSK rate 1/3, rate 1/2, rate 3/4, 16QAM rate 1/2, and rate 3/4.

VI. SCHEDULERS TO COMPARE

The behaviour of the following schedulers is explored in this paper:

A. Round Robin [3]

The Round Robin (RR) scheduler is one of the simplest. At the end of each TTI, it moves onto the next queue and tries to fill up the scheduling interval with as much data as possible from this queue. If a queue is empty, the scheduler simply moves to the next queue. No attempt at optimisation is made. If all queues are full, then over a prolonged period of time, each one would be given an equal number of TTIs for transmission.

B. Proportional Fair Throughput [3]

The aim of the Proportional Fair Throughput (PF-T) scheduler is to maximise throughput, but in a fair manner. The scheduling rule is given by:

$$\text{Next packet} = \max_i \frac{r_i}{\bar{r}_i}, \quad (28)$$

where r_i is the instantaneous transmission rate that HSDPA could assign to class i traffic, if chosen by the scheduler. \bar{r}_i is the average transmission rate that was assigned to class i traffic. It can be found as follows:

$$\bar{r}_i(k+1) = \begin{cases} (1-\alpha)\bar{r}_i(k) + \alpha r_i(k) & \text{if } i \text{ is served in slot } k, \\ (1-\alpha)\bar{r}_i(k) & \text{otherwise,} \end{cases} \quad (29)$$

where $0 < \alpha < 1$ is a weighting factor, whose value was chosen to be 0.001.

The numerator in the scheduling rule ensures that the scheduler will take advantage of temporary throughput improvements. On the other hand, the denominator ensures that over a long-term period, the scheduler will attempt to assign equal resources to all classes.

C. Maximum Carrier to Interference ratio [3]

The Maximum Carrier to Interference ratio (Max C/I) scheduler selects packets based purely on the best channel conditions. The relative instantaneous channel quality indicator η is defined by

$$\eta_i = \frac{\zeta_i}{\bar{\zeta}}, \quad (30)$$

where $\zeta_i = \frac{E_b}{I_0}$ is the SNR of the channel that the head-of-queue packet of the class i queue will be transmitted through, while $\bar{\zeta} = \frac{1}{n} \cdot \sum_{j=1}^n \zeta_j$ is the average SNR of all channels. Here n is the total number of traffic classes.

The scheduling rule is simply given by:

$$\text{Next packet} = \max_i \eta_i. \quad (31)$$

D. Earliest Deadline First

The Earliest Deadline First (EDF) scheduler attempts to meet the required deadlines of traffic classes. Distributing bandwidth and achieving fair throughput among the traffic classes is not a sufficient criterion to be able to sustain real-time traffic on a packetised network. Strict deadlines need to be adhered to. The scheduling rule for EDF is given by:

$$\text{Next packet} = \min_i (d_i - D_i), \quad (32)$$

where D_i is the queueing delay that the head-of-queue packet of the class i queue has experienced, while d_i is the delay deadline of the packet.

E. O-EDF Scheduler

Finally, Opportunistic EDF (O-EDF) uses the rules of EDF, but prioritises traffic classes which will be transmitted through good channel conditions.

$$\text{Next packet} = \min_i d_i \cdot \frac{\bar{G}_c}{G_c} - D_i \quad (33)$$

VII. SIMULATION PARAMETERS

Table II lists the parameters for the simulation, which are similar to those proposed in [1]. A variable number of traffic sources were modeled, depending on the required load. Each traffic source generated the video, voice, and web traffic according to the model discussed in Section II. Traffic was generated independently of other traffic sources.

For the HSDPA links between Node B and the UEs, 20 mobile channels were created. Each packet that was generated was assigned a flat random number ranging from 0 to 19, which implied the target channel that it would be transmitted through. Each of the 20 channels was modeled independently of all other channels. No spacial correlation among transmission channels was taken into consideration.

VIII. RESULTS

In this paper, the behaviour of five schedulers was simulated.

A. Average delay and unfairness

Figs. 1(a), 1(c), and 1(e) contain the average queueing delay that video, voice and web traffic respectively experience. At first glance, EDF seems to perform the least favourable, as of all three classes it has the highest queueing delay for large portions of the load conditions. One needs to remember though that EDF schedules traffic based on the delay deadlines of the classes, which are 100ms for video traffic, 50ms for voice traffic, and 500ms for web traffic. In the case of voice traffic, EDF traffic does not reach the deadline, but is still much higher than the other schedulers under high load conditions. The reason for voice delays not reaching the deadline is due to multiple transmissions of voice packets per TTI. The delay of video and web traffic, on the other hand, approaches the deadline at relatively low loads.

The advantage of the EDF scheduler is the clear queueing delay differentiation of traffic classes. The delay deadlines

of the various traffic classes create behavioral expectations with customers, which they are prepared to pay for. If no discernable difference is noticeable, the requirement for an advanced delay-targeting QoS design becomes questionable. To further illustrate the point, a form of delay unfairness is proposed that is defined as follows:

$$\text{Unfairness} = 100\% \cdot \frac{d_i - D_i}{d_i}, \quad (34)$$

where d_i is the delay deadline of class i and D_i is the queuing delay that the front-of-queue- i packet has experienced by the time it is served. In other words, with this expression an attempt is made to measure how close the delay was to its deadline, by the time the packet completed transmission.

Figs. 1(b), 1(d), and 1(f) contain the average unfairness, measured according to (34). This highlights, more clearly, how EDF is able to differentiate the delay of the 3 classes, based on their respective deadlines. For all 3 classes, as the load increases, EDF approaches the deadline, marked as 0 on the unfairness graphs.

In the case of video traffic, all schedulers are able to approach the delay deadline quite well. For most of the schedulers this is due to their preference of voice over video traffic. As the video delay approaches the delay deadline, it is dropped. This creates an artificial ceiling around the delay deadline. In the case of voice and web traffic, all schedulers, except for EDF, completely ignore the delay deadline.

IX. CONCLUSION

This paper presented an HSDPA simulation environment. Video, voice, and web traffic were realistically produced by a varying number of sources. The number of sources was varied from 1 to 20, thereby achieving load-conditions varying from 0% to 100%. The model focussed on the downlink transmissions from the base-station to the mobile units, migrating around the cell. It measured the average queuing delay and fairness. The results show how EDF is able to differentiate between traffic classes.

TABLE II
SIMULATION PARAMETERS [1]

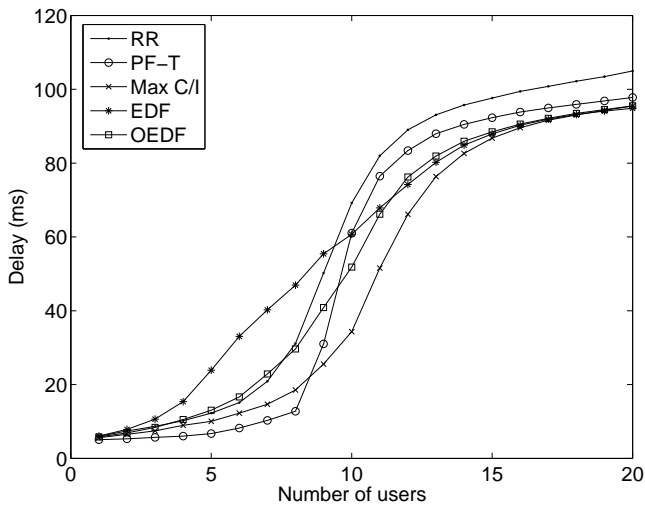
PARAMETER	VALUE
α	0.025
P_D	30 Watt
Cell radius R	500 meters
$N = N_0 W$	$2.00245 \cdot 10^{-14}$ Watt
W	5MHz
ε (close to Node B: $R \leq 10\text{m}$)	0
ε (intermediate: $10\text{m} < R < 450\text{m}$)	0.25
ε (cell border: $450\text{m} \leq R \leq 500\text{m}$)	0.5
PATHLOSS ATTENUATION $L(t)$ (in dB)	
Near zone ($R < 300\text{m}$)	$92.92 + 10.96 \log(R)$
Far zone ($R \geq 300\text{m}$)	$106.48 + 43.85 \log(R)$
None-line of sight (probability = 0.2) (R = distance from Node B)	$151.32 + \log(R)$
SHADOWING ATTENUATION	
$s(t)$ (Gaussian variable)	(in dB)
Mean	0
Standard deviation	8
Bit Error Rate Threshold	
Video	$9 \cdot 10^{-6}$
Voice	$9 \cdot 10^{-6}$
Web	$8.4 \cdot 10^{-7}$

REFERENCES

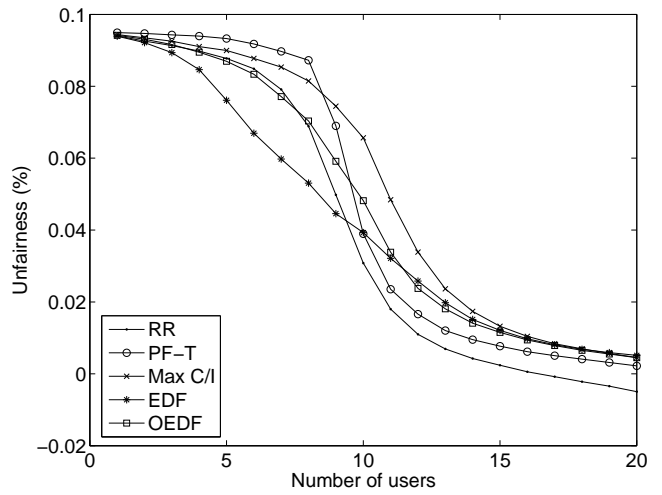
- [1] F. D. Angelis, I. Habib, G. Giambene, and S. Giannetti, "Scheduling for differentiated traffic types in HSDPA cellular systems," in *IEEE Globecom '05*, 2005.
- [2] A. Farrokhi, F. Blomer, and V. Krishnamurthy, "A comparison of opportunistic scheduling algorithms for streaming media in high-speed downlink packet access (HSDPA)," in *Proc. of MIPS 2004, Grenoble, France*, November 16–19, 2004.
- [3] A. Haider and R. Harris, "A novel proportional scheduling algorithm for HSDPA in UMTS networks," in *The 2nd International Conference on Wireless Broadband and Ultra Wideband Communications* (I. C. Society, ed.), 2007.
- [4] A. Harada, S. Abeta, and M. Sawahashi, "Adaptive radio parameter control considering QoS for forward link OFCDM wireless access," in *IEEE VTC*, pp. 1175–1179, 2002.

Stefan M. Scriba completed his BScEng degree in December 2000 in the School of Electrical, Electronic and Computer Engineering at the University of Natal, Durban, South Africa. He completed his MScEng degree at the beginning of 2003 and is currently working on his PhD, researching scheduling theory for both wired and wireless networks in the Radio Access Technology Research Centre at the University of KwaZulu-Natal, South Africa. Since 2006 he is working at Telkom SA Ltd.

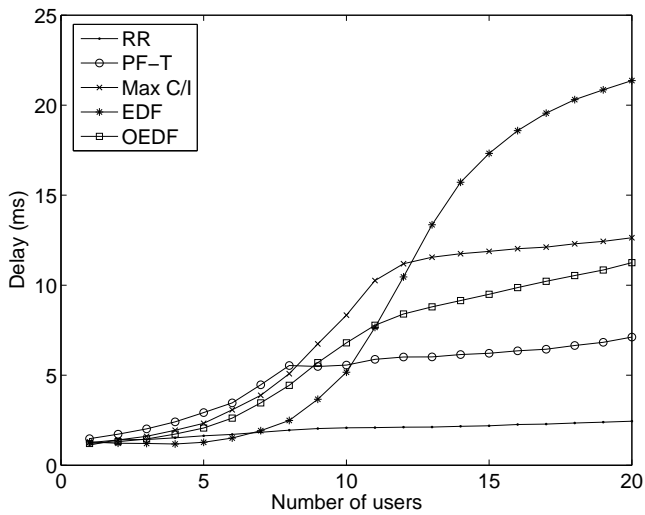
Professor Fambirai Takawira is the Telkom Professor of Digital Communications at the School of Electrical, Electronic and Computer Engineering at the University of KwaZulu-Natal. He completed his BScElecEng in Manchester and was awarded a PhD at Cambridge. His research interests are in the general areas of adaptive signal processing, digital communications and data networks.



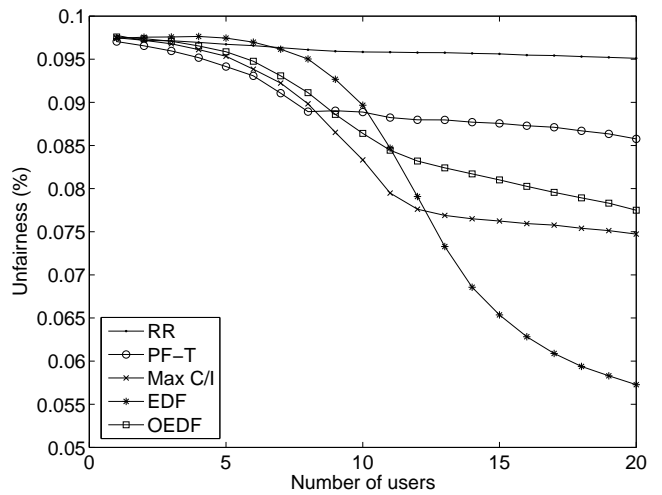
(a) Average delay of Video traffic



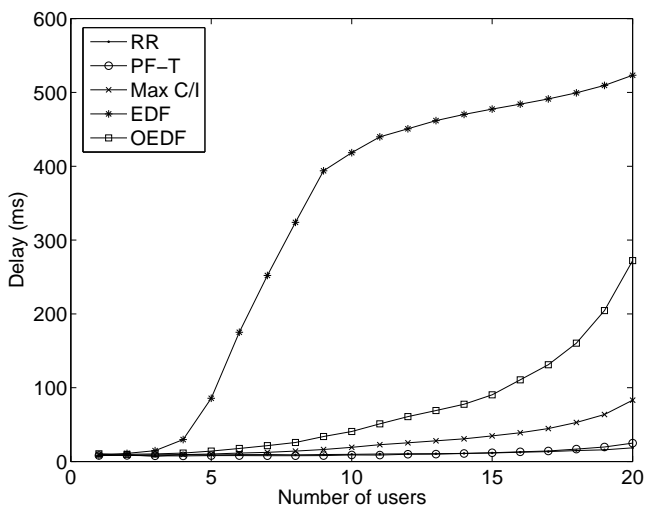
(b) Unfairness of Video traffic



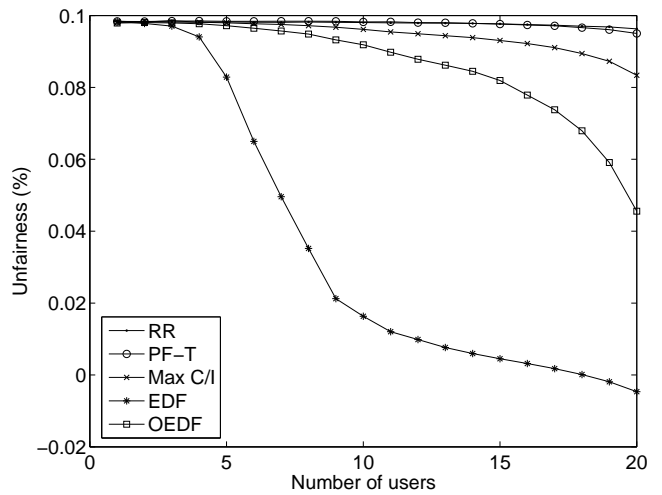
(c) Average delay of Voice traffic



(d) Unfairness of Voice traffic



(e) Average delay of Web traffic



(f) Unfairness of Web traffic

Fig. 1. Average delay and unfairness