Journal of Computer Science 7 (11): 1717-1723, 2011
ISSN 1549-3636

Dynamic Bandwidth Allocation for Multiple Traffic Classes in IEEE 802.16e WiMax Networks: A Petrinet Approach

S. Geetha and R. Jayaparvathy
Department of EEE, Coimbatore Institute of Technology,
Coimbatore 641 014, Tamilnadu, India

Abstract: Problem statement: WiMAX supports multiple types of traffic such as data, voice and video. Each flow requires a certain minimum bandwidth to achieve its QoS. Bandwidth allocation to traffic classes should be in such a way that fairness criteria is met with. Hence, we propose a dynamic bandwidth allocation mechanism to achieve fair and efficient allocation. **Approach:** We present a Generalized Stochastic Petri Net (GSPN) approach to model bandwidth allocation in Broadband Wireless Access (BWA) networks with multiple traffic classes. A dynamic weight assignment mechanism is proposed to enable fair bandwidth allocation among the competing traffic classes. Performance of the weight assignment mechanism is analytically evaluated using the GSPN model developed. **Results:** Results show performance improvement in terms of mean delay and normalized throughput of traffic classes compared to existing mechanisms. Simulation is carried out for different traffic rates. Analytical results are validated using simulations. **Conclusion:** Performance of the proposed system is evaluated in terms of mean delay and normalized system throughput. The model developed is generic and can be extended to any wireless network with multiple traffic classes.

Key words: Bandwidth allocation, dynamic weight assignment, model developed, multiple traffic, bandwidth requirement, resource allocation, assignment mechanism, traffic load, analytical approach, based simulator

INTRODUCTION

WiMAX provides low cost all IP solutions for scalable networks with voice, data and video services. The radio network of IEEE 802.16e BWA provides interoperable, flexible, low cost solutions to the 10-66 GHz (line of sight) and 2-11 GHz (non-line of sight) spectral bands (Anderson, 2003). Data rates of 32-130 Mbps can be achieved depending on the channel bandwidth and modulation techniques used. Multiple types of traffic flows (data, voice and video) are supported. Each flow requires certain minimum bandwidth to achieve its QoS. Bandwidth should be allocated so that all flows share the available capacity in compliance with the fairness criteria. Increased flow of traffic belonging to any QoS class increases its bandwidth requirement. Hence, it is essential to change the bandwidth allocation policy dynamically based on instantaneous traffic load. Several bandwidth allocation mechanisms have been proposed in literature.

The UPS (Uplink Packet Scheduling) (Wongthavarawat and Ganz, 2003) and Deficit Fair Priority Queue (DFPQ) (Chen *et al*., 2005) employ service classes to meet differentiation and fairness. A simple mathematical approach for delay analysis for WiMax networks has been presented in (Sharieh *et al*., 2008). Scheduling strategies for multimedia networks has been presented. Dynamic adjustment of DL (downlink) and UL (uplink) is performed in (Ma, 2009) to maximize bandwidth utilization. Lin *et al*. (2009) a bandwidth allocation algorithm, HUF (Highest Urgency First), is proposed which calculates slot allocation in two phases. The algorithm is validated through simulations. Sayenko *et al*. (2006) strict priority is applied which could result in starvation for low-level service class even with the implementation of admission control scheme. Petrinet approach to bandwidth allocation has been studied in (Raja and Kumanan, 2007). Liu *et al*. (2005); Chen *et al*. (2005) and Wongthavarawat and Ganz (2003) discuss complex schedulers such as Earliest Deadline First (EDF), Deficit Round Robin (DRR) (Shreedhar and Varghese, 1996), weighted fair queuing (WFQ) and worst case weighted fair queuing (W^2FQ). Using a hierarchy of schedulers is a challenging task because per connection QoS must be translated into scheduler configuration at each level. Performance evaluation of prioritized queues has been considered.

Corresponding Author: S. Geetha, Department of EEE, Coimbatore Institute of Technology, Coimbatore 641 014, Tamilnadu, India

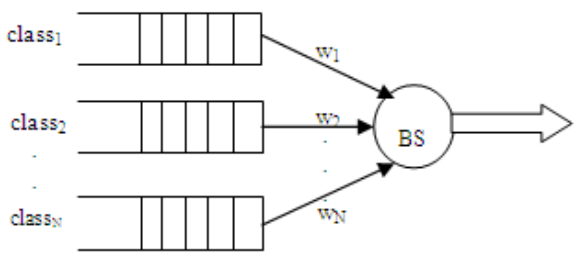

Fig. 1: System model

In this study, we present a GSPN approach to model bandwidth allocation in wireless systems with multiple traffic classes. We also present a dynamic weight adjustment mechanism for fair resource allocation in the system. According to this mechanism, weights assigned to traffic flows are varied dynamically, depending on priority of traffic class and traffic load conditions. We compute the average system throughput and mean delay suffered by the first packet (i.e., the packet in the Head Of Line (HOL) of each queue) through the proposed GSPN model. Mean delay of subsequent packets is determined by modeling each queue as M/G/1 queue (Jayaparvathy *et al*., 2006). The mean service time for the computation is obtained from the mean delay suffered by the HOL packet. Our analytical model is validated by comparing the results with simulations carried out using event based simulator.

System model: A system consisting of single Base Station (BS) and n Subscriber Stations (SS) is considered as shown in Fig. 1. Each SS is associated with multiple queues, each corresponding to the different traffic class for which resources have to be allocated dynamically. The BS assigns bandwidth to each SS which in turn re-allocates the bandwidth to the traffic flows incident on it. Traffic classes are prioritized based on the QoS requirements. Hence, it is required to allocate the available bandwidth appropriately considering the fairness as well as QoS requirements.

The following are the assumptions made in the model:

- There are N different traffic classes in the system denoted as $class_i$, $j \in (1,N)$
- $class_i$ has a higher priority compared to $class_j$ for I < j
- Every traffic class is assigned a dynamic weight Wi
- We consider data-only traffic with on-off traffic model. Data bursts consist of active and idle periods. (Practically, a data burst represents a data packet of variable length, for example an IP packet with zero idle time between finite set of consecutive packets (Jayaparvathy *et al*., 2007)
- Data bursts arrival at a queue follows a Poisson process with mean arrival rate λ_i
- Service times of data bursts are exponentially distributed with mean $1/\mu_i$ seconds

MATERIALS AND METHODS

Dynamic bandwidth assignment mechanism: Different traffic classes have varying bandwidth requirement depending on the traffic load. Based on the stringent nature of QoS requirements, traffic classes are classified into higher and lower priority traffic classes. Each traffic class is assigned a dynamic weight Wi, which depends on (i) QoS requirement (ii) queue length (which depends on the load conditions) of the traffic class. Assignment of static weights could result in starvation problem for lower priority traffic class, particularly at higher loads. Hence, weight assigned should vary depending on instantaneous system load conditions.

Let ρ_i represent the traffic load of traffic $class_i$ given by $\rho_i = \frac{\lambda_i}{\mu_i}$ where, λ_i is the mean arrival rate and $\frac{1}{\mu_i}$ is the mean service time for traffic $class_i$. The following conditions hold good. (i) $0 < \rho_i < 1, \forall i = 1,2,...,N$ (ii) $\sum_{i=1}^{N} \rho_i < 1$ (iii) $\sum_{i=1}^{N} w_i = 1$. The first two conditions ensure stability of the queues and the last condition is a normalization condition. In order to account for the relationship between weight and traffic load, we introduce the term, sensitivity, which represents the change in weight of a given traffic class with respect to change in load of other traffic classes. Let α be the sensitivity of $class_i$ to the variations in the traffic load of $class_1$ to $class_{(i-1)}$. Note that $\alpha \in (1, \infty)$. Also, $\alpha \to 1$ indicates no sensitivity and $\alpha \to \infty$ indicates maximum sensitivity. Hence, $\alpha \to 1$ when $\sum_{j=1}^{i-1} \rho_j \to 0$ and $\alpha \to \infty$ when $\sum_{j=1}^{i-1} \rho_j \to 1$. An expression that satisfies the above condition is Eq. 1:

$$\alpha_i = \frac{1}{1 - \sum_{j=1}^{i-1} \rho_j} \quad (1)$$

Weights assigned to traffic classes need to satisfy the following properties:

- Weight has to be an increasing function of the corresponding traffic load
- The weight of lower priority traffic class has to decrease with increase in higher priority traffic load
- Under equal traffic load conditions, the weight of higher priority class has to be greater than that of the lower priority class
- When lower priority traffic load is greater than higher priority traffic load, higher weight is assigned to the lower priority traffic class. This avoids starvation for the lower priority traffic class and hence ensures fairness

Based on the properties discussed above, we formulate the weight of a traffic class as Eq. 2:

$$w_i = \left(1 - \sum_{j=1}^{i-1} w_j\right)\rho_i^{\alpha i} \quad (2)$$

Further, we normalize the weight assigned by assigning Eq. 3:

$$w_{iNorm} = \frac{w_i}{\sum_{i=1}^{N} w_i} \quad (3)$$

such that the relation $\sum_{i=1}^{N} w_{iNorm} = 1$ is satisfied. In the following sections of paper we represent w_{iNorm} as w_i.

The following theorems discuss the behaviour of the weight allocation mechanism under different load conditions.

Theorem 1: Under equal traffic load conditions weight of higher priority traffic class is greater than lower priority traffic class. i.e., when $\rho_1 = \rho_2 = ... = \rho_N$. $w_1 > w_2 > ... > w_N$

Proof: Let $\rho_1 = \rho_2 = ... = \rho_N = \rho$. From (2), we have, $w_{N-1} = \left(1 - \sum_{j=1}^{N-2} w_j\right)\rho^{\alpha_{N-1}}$; $w_N = \left(1 - \sum_{j=1}^{N-1} w_j\right)\rho^{\alpha_N}$. From (1), $\alpha_{N-1} < \alpha_N$. Hence, for $\rho < 1$, we have $w_{N-1} > w_N$.

Remarks: The above condition enables the mechanism to maintain QoS requirements of the system.

Theorem 2: For a higher load of lower priority class, corresponding higher weight is assigned to the traffic class.
i.e., when $\rho_N > \ldots > \rho_2 > \rho_1$, $w_N > \ldots > w_2 > w_1$.

Proof: Let $\rho_N = k\rho_{N-1}$. From (2) we have:

$$w_{N-1} = \left(1 - \sum_{j=1}^{N-2} w_j\right)\rho_{N-1}^{\alpha_{N-1}}$$

$$w_N = \left(1 - \sum_{j=1}^{N-1} w_j\right)\rho_{N-1}^{\alpha_N}$$

For $k > 1$ and $\alpha_N \geq 1$ we have $w_N > w_{N-1}$, since $\alpha_{N-1} < \alpha_N$.

Remarks: When lower priority traffic class has higher load compared to higher priority traffic class, correspondingly higher weight is assigned to it. Though a lower weight is assigned to higher priority class, it does not degrade the overall system performance since the bandwidth requirement is comparatively less. This property brings fairness in the proposed weight allocation mechanism.

Performance analysis using GSPN model: Figure 2 shows the GSPN model, we consider one of the traffic classes as reference. The behavior of other traffic classes is aggregated and represented separately. It is observed from the Figure that the model consists of two parts , A and B. Part A, represents the events associated with the reference traffic class and Part B, represents the events associated with other traffic classes.

The model incorporates priority, pre-emption and time-out characteristics of traffic classes. Note that we use the subscript, i, to represent the reference traffic class and λi to represent combined events of other traffic classes. Transition usr_i generates packets at the given rate λi. and deposits them in the place qi. An inhibitor arc with cardinality buf_i is needed to ensure that the number λi of packets waiting to enter the current queue is finite. If all channels are busy, the data packets are buffered in q_i with finite buffer size bufi. Transition usr_{i*} represents the arrival of other, (N-1), traffic classes with arrival rate λ_{i*}. Packets are buffered in qi with capacity bufi. Access to channel by the reference class is controlled by transition chchki which is modeled as timed transition. Rate of firing of chchki is controlled by user-defined function chchki given by (2). Similarly, access to channel by other traffic classes is controlled by chchki*. The rate of firing of chchki* is given by:

$$w_{i*} = (1 - w) \text{ since}, \sum_{i=1}^{N} w_i = 1$$

A higher value of firing rate implies a higher probability to access channel resources. Thus, channel allocation can be varied dynamically based on traffic load conditions. Firing $chchk_i$ transfers a packet from q_i to usg_i indicating the packet is being served.

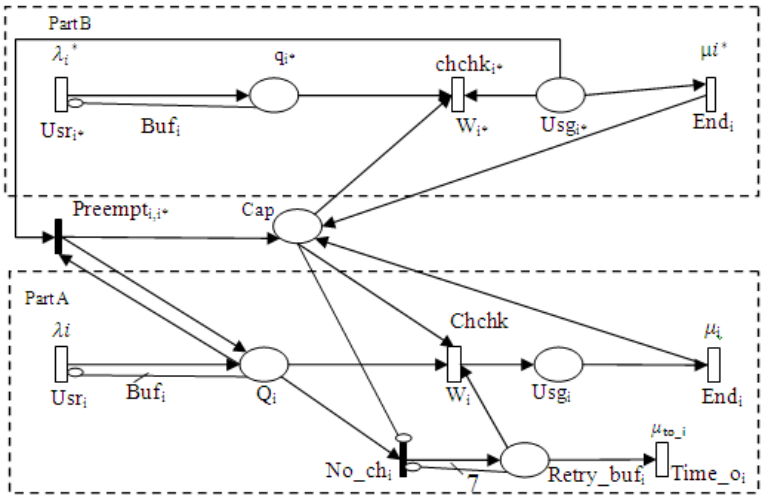

Fig. 2: GSPN model

After completion of service time, transition end_i is fired and the channel is returned to the central pool. Transition $preempt_{i,i*}$ is an immediate transition used to model pre-emption. $preempt_{i,i*}$ is enabled when packets are available in places qi and $usgi_*$ simultaneously. This indicates the presence of packets belonging to $class_i$ and $class_{i*}$ simultaneously , where the reference class, $class_i$ has higher priority compared to other classes, $class_{i*}$. Arc connecting $preempt_{i,i*}$ and usgi indicate removal of packet from usgi and returning the channel to the central pool of channels thus enabling $class_i$ to access the channel.

Transition no-chi is fired when the available channels is insufficient to serve the incoming packets. An inhibitor arc from cap to no-chi indicate non availability of channels.

Firing no-chi deposits the packets in retry- bufi with a buffer size set to 7. Arc connecting retry- bufi and $chchk_i$ represents the retrial of buffered packets for channel access.

Traffic classes are assumed to belong to delay sensitive applications with a maximum threshold on tolerable delay. Packets exceeding the threshold are dropped. Dropping of packets exceeding the delay limit is incorporated in the model using timed transitions $time\_o_i$ for reference class. Firing rate of $time\_o_i$ is set to μ_{to_i}, where $1/\mu_{to\_i}$ is the maximum tolerable delay for packets belonging to $class_i$. Firing $time\_o_i$ removes a packet from $1/\mu_{to\_i}$ indicating a packet drop. Probability of packet drop depends on the available channels, transmission rate of packets, buffer size.

Mean delay and normalized throughput: The underlying Continuous Time Markov Chain (CTMC) of the GSPN model discussed above can be obtained from reachability graph (Sahner *et al*., 1996). Since, the associated CTMC is very complex, we use SHARPE (Sahner *et al*., 1996) tool to obtain the performance metrics. The average throughput of a transition T is defined as the average rate at which packets are deposited by the transition in its output places. If o(t) is the average number of packets deposited by transition T in all of its output places up to a time t, then the throughput of a transition T, defined as Eq. 4:

$$\eta_T = \lim_{t\to\infty} \frac{\delta(t)}{t} \tag{4}$$

The throughput of traffic $class_i$, is given by Eq. 5:

$$\eta_i = \frac{\eta end_i}{\eta usr_i} \tag{5}$$

Average system throughput, n is given by Eq. 6:

$$\eta = \sum_{i=1}^{N} \eta_i \tag{6}$$

The mean delay, D_H, experienced by a HOL packet of traffic $class_i$, is the sum of the mean packet holding time and the sum of mean waiting times in places qi and usgi. Let the average number of packets in place P be #. $\hat{P}\hat{D}_H$ can be computed using Little's Theorem as Eq. 7:

$$\hat{D}_{H_i} = \frac{No(q_i)}{\eta_{usr_i}} + \frac{No(usg_i)}{\eta_{chchk_i}} + \frac{1}{\mu_i} \tag{7}$$

where, μi is the mean packet holding time for traffic $class_i$. The buffer in each queue is modelled as M/G/1 queue with mean service time $class_i$. The mean packet delay, $class_i$. can be determined by applying the Pollackzek-Kinchine mean value formula as Eq. 8:

$$\hat{D}_1 = \hat{D}_{H_i} = \left\lceil 1 + \frac{\rho_{bi}}{2(1-\rho_{bi})}(1+C_{Ri}^2) \right\rceil \quad (8)$$

where, $\rho_{bi} \triangleq \lambda_i \hat{D}_{Hi}$. If delay of HOL packet is represented by random variable, R_i, then Eq. 9:

$$C_{R_i}^2 = \frac{E\lfloor R_i^2 \rfloor}{\hat{D}_{Hi}^2} \quad (9)$$

For small loads, $E[R_i^2]$ can be obtained as Eq. 10:

$$E[R_i^2] = 2\left(\frac{nousg_i}{\eta_{chchk_i}}\right)^2 \quad (10)$$

RESULTS

We evaluate the system performance in terms of mean delay and normalized throughput for increasing traffic load, ρ given by $\sum_{i=1}^{N}\rho i$ where ρ_i corresponds to traffic load of $class_i$ for i = 1, 2,…, N. $\rho_i = \frac{\lambda_i}{\mu_i}$, where λ_I is the mean arrival rate and μ_i is the mean service rate of each traffic class. We assume N = 3 for our analysis. We ensure system stability by setting $\rho \leq 1$. The value of λ_i is chosen to vary from 0.0-0.3. $buf_i = 3$; $\mu_i = 1$ and cap = 10. We compare the analysis and simulation results for three traffic classes in terms of mean delay and normalized throughput. We also compare the performance of the proposed weighted priority scheme with fixed priority ($class_1$ highest followed by $class_2$ and $class_3$) and equal priority schemes. Priority adjustment is achieved by assigning w_i. The value of w_i = 0.33 for equal priority case. For fixed priority we assign w_1 = 0.5 w_2 = 0.3 w_3 = 0.2. Simulations are carried out using an event based simulator. The parameters used in the simulation are frequency band = 5 Mhz, propagation model assumed is two ray ground model, frame duration = 20 ms, cyclic prefix = 0.25 and packet length = 1025 bytes. We consider 9 rtPS, 3 nrtPS and 2 BE sources for simulation. The modulation setting chosen is 64-QAM 2/3. The simulation duration is chosen to be 100s.

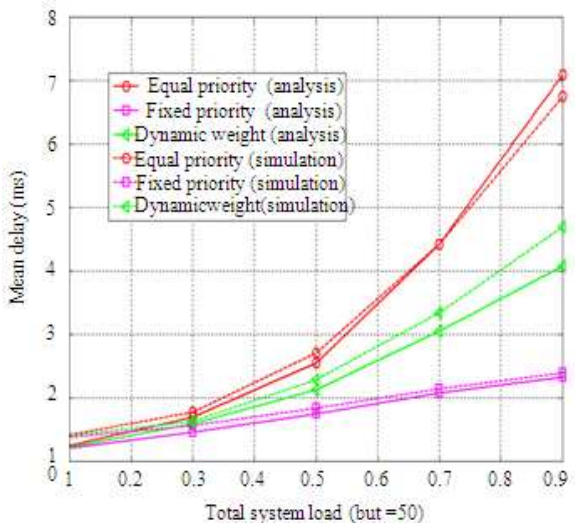

Fig. 3: Comparison of mean delay for $class_1$ traffic with different priority schemes

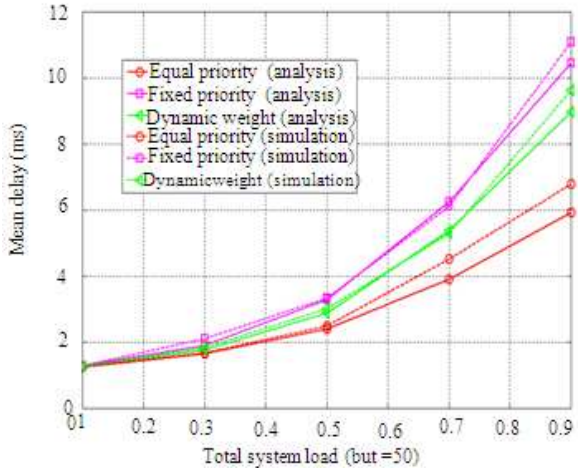

Fig. 4: Comparison of mean delay for $class_3$ traffic with different priority schemes

Figure 3-4 present a comparison of mean delay with various priority schemes for $class_1$ and $class_3$ traffic respectively.

DISCUSSION

As observed from the figures we find that mean delay with fixed priority is the least for $class_1$ and highest for $class_3$. This implies that at higher load of lower priority traffic class, the delay increases due to insufficient bandwidth available. Also, with equal priority allocation mechanism, we find that $class_1$ has highest delay and $class_3$ has the lowest delay.

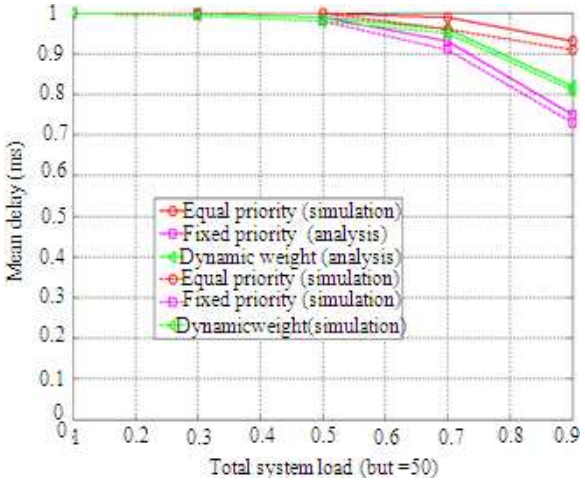

Fig. 5: Comparison of normalized throughput for $class_3$ traffic with different priority schemes

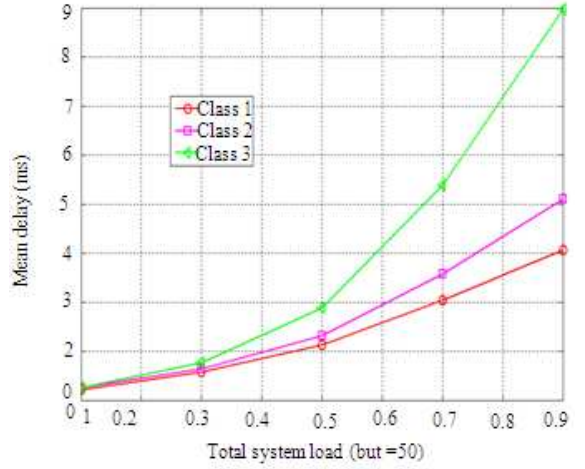

Fig. 6: Mean Delay of three traffic classes

This is not acceptable since $class_1$ traffic is assumed to be delay sensitive application and hence cannot tolerate excessive delay. With dynamic weight assignment mechanism, we achieve a balance between the two mechanisms. Also, from the Figure we find that the analytical results match closely with the simulations, thus validating our analytical approach.

We compare the normalized throughput of $class_3$ traffic with different allocation mechanisms in the Fig. 5. Throughput of given traffic class decreases with increased load.

As for $class_3$, we find the throughput decreasing considerably at higher traffic loads. With dynamic weight mechanism, we achieve an increased throughput at higher traffic loads. For a traffic load of 0.9, throughput of $class_3$ is increased from 0.72 with fixed priority mechanism to 0.8 with dynamic weight mechanism.

In Fig. 6, we compare the mean delay of $class_1$, $class_2$ and $class_3$ traffic with dynamic weight assignment mechanism. We observe the mean delay of $class_1$ being least compared to $class_2$ and $class_3$. Hence, the proposed mechanism preserves priority requirements while maintaining fairness in resource allocation.

CONCLUSION

We presented a GSPN model for performance evaluation of IEEE 802.16 BWA systems with multiple traffic classes. We have also proposed a dynamic weight assignment mechanism to achieve fair bandwidth allocation. Performance of the proposed system is evaluated in terms of mean delay and normalized throughput. Our model is validated using simulations. The model can be generalized to incorporate multiple access networks. Use of colored Petri net can be explored to model the behavior of traffic classes.

REFERENCES

Anderson, H.R., 2003. Fixed Broadband Wireless System Design. 1st Edn., John Wiley and Sons, Chichester, West Sussex, England ; Hoboken, NJ., ISBN: 0470844388, pp: 510.

Chen, J., W. Jiao and H. Wang, 2005. A service flow management strategy for IEEE 802.16 broadband wireless access systems in TDD mode. Proceedings of the International Conference on Communications, May 16-20, IEEE Xplore Press, pp: 3422-3426. DOI: 10.1109/ICC.2005.1495056

Jayaparvathy, R., S. Anand, S. Dharmaraja and S. Srikanth, 2007. Performance analysis of IEEE 802.11 DCF with stochastic reward nets. Int. J. Commun. Syst., 20: 273-296. DOI: 10.1002/dac.821

Lin, Y.N., Y.D. Lin, Y.C. Lai and C.W. Wu, 2009. Highest Urgency First (HUF): A latency and modulation aware bandwidth allocation algorithm for WiMAX base stations. Comp. Commun., 32: 332-342. DOI: 10.1016/j.comcom.2008.11.003

Liu, N., X. Li, C. Pei and B. Yang, 2005. Delay character of a novel architecture for IEEE 802.16 systems. Proceedings of the 6th International Conference on Parallel and Distributed Computing, Applications and Technologies, Dec. 05-08, IEEE Xplore Press, pp: 293-296. DOI: 10.1109/PDCAT.2005.112

Ma, L., 2009. Current Technology Developments of WiMax Systems. 1st Edn., Springer, ISBN-10: 1402092997, pp: 316.

Raja, K. and S. Kumanan, 2007. Resource leveling using petrinet and memetic approach. Am. J. Applied Sci., 4: 317-322. DOI: 10.3844/ajassp.2007.317.322

Sahner, R.A., K.S. Trivedi and A. Puliafito, 1996. Performance and Reliability Analysis of Computer Systems: An Example-Based Approach using the SHARPE Software Package. 1st Edn., Kluwer Academic Publishers, Boston, ISBN: 0792396502, pp: 404.

Sayenko, A., O. Alanen, J. Karhula and T. Hamalainen. 2006. Ensuring the QoS requirements in 802.16 scheduling. Proceedings of the 9th ACM International Symposium on Modeling Analysis and Simulation of Wireless and Mobile Systems, Oc. 02-06, Torremolinos, Spain, pp: 108-117. DOI: 10.1145/1164717.1164737

Sharieh, A., M. Itriq and W. Dbabat, 2008. A dynamic resource synchronizer mutual exclusion algorithm for wired/wireless distributed systems. Am. J. Applied Sci., 5: 829-834. DOI: 10.3844/ajassp.2008.829.834

Shreedhar, M. and G. Varghese, 1996. Efficient fair queuing using deficit round-robin. IEEE/ACM Trans. Network, 4: 375-385. DOI: 10.1109/90.502236

Wongthavarawat, K. and A. Ganz, 2003. IEEE 802.16 Based last mile broadband wireless military networks with quality of service support. Proceedings of the IEEE Military Communications Conference, Oct. 13-16, IEEE Xplore Press, pp: 779-784. DOI: 10.1109/MILCOM.2003.1290211