Quality-of-Service (QoS) Framework for Multi-rate Wireless Ad-hoc Network (MWAN)

Yow-Yiong Edwin Tan, Stephen McLaughlin, David I. Laurenson
Institute for Digital Communications, School of Engineering and Electronics
The University of Edinburgh, Alexander Graham Bell Building
Kings Buildings, Mayfield Road, Edinburgh EH9 3JL, U. K.
Tel: +44 (0) 131 650 5659, Fax: +44 (0) 131 650 6554
Email: yow.tan@ee.ed.ac.uk

***Abstract*— We propose MWAN, a multi-rate network model to deliver differentiated service in a wireless mobile ad-hoc network (MANET) with varying physical-layer link speed. The proposed architecture is modelled using a multi-dimensional Markov chain to support both real-time and non-real-time applications. It is demonstrated that various types of data arrival process can be modelled by a Markov Modulated Arrival Process (MMAP). Numerical analyzes are drawn to estimate the packet drop probability, effective throughput and packet queuing delay. We validate the scheme by simulating under different link utilizations with IEEE 802.11 Distributed Coordination Function (DCF) ad hoc mode. Analytical and simulation results are compared to determine the accuracy of the presented methods. Increased Quality-of-Service (QoS) performance is achieved for high priority traffic.**

I. Introduction

The increasing widespread use of wireless technologies has given rise to the need for QoS provisioning mechanisms for multimedia applications in wireless networks. However, to support QoS in MANET (without a fixed infrastructure or administrative support) is more challenging than in fixed and last-hop wireless access networks. Highly dynamic network topology and traffic load conditions, time-variant QoS parameters (e.g. throughput, latency and etc), and less communication bandwidth, smaller processing and power capacity than fixed network makes it difficult to support diverse application with an appropriate QoS.

The Internet Engineering Task Force (IETF) have proposed two internet QoS model namely: Integrated Service (IntServ) and Differentiated Service (DiffServ) [1]. In IntServ, stations classify incoming packets and network resources are explicitly identified and reserved. For DiffServ, traffic is categorized into classes while stations provide priority-based treatment without reserving resources exclusively. For instance, Assured Forwarding (AF) in DiffServ [2] provides differentiation service between traffic classes where the low-priority class experiences higher loss rates and delays than the high-priority class. AF is implemented with Random Early Discard (RED) [3] or an equivalent Active Queue Management technique. Different packet drop precedence levels indicate that packets receive different priorities in accessing buffers during periods of congestion. RED drops packets at random when average queue length exceeds a given minimum threshold.

Several QoS framework for MANETs have been proposed. In [4], the FQMM was presented that combines the reservation procedure for high priority traffic with service differentiation for low-priority traffic; INSIGNIA [5], an in-band signalling protocol, integrated with an ad-hoc routing protocol; Network feedback based on link and permissible throughput measurements were made to support higher layer and soft-QoS [6].

However, these schemes do not take into account the characteristics of MANET and drawbacks of IntServ and DiffServ remain. Therefore, to support a combination of real-time (e.g. voice or video) and non-real-time services (e.g. data or FTP), an accurate model has to be defined to investigate its applicability within the MANET. We consider a single server queuing system with a finite buffer and heterogeneous arrival streams. The arrival process is a Poisson or Markov Modulated Poisson Process (MMPP) while the service times (packet lengths) are iid with a general distribution. This classic problem of queuing theory where the probability of buffer overflow and packet dropping are computed. An additional property to consider is that existing wireless devices implement IEEE 802.11 a/b/g [7] standards which utilize multiple transmission rates depending on channel conditions, distance and transmitting power. Thus each mobile stations transmits data at an appropriate transmission rate using a particular modulation scheme based on the perceived Signal-to-Noise ratio (SNR) of the immediately previous frame in the frame exchange process. Provisioning of service delivery are dynamically varied by selecting links that can use higher bandwidth modulation schemes. In this paper, we integrated wireless channel modelling and data queuing analysis at the packet-level to provide a unique approach for studying the effect of physical-layer link speed on high-layer network performance. We verify the analytical model using simulation results. Our results quantify the impact of the buffer scheme, time varying channel and traffic sources on drop probability, throughput and delay.

This work is organized as follows: section II and III give an overview of AF within DiffServ and assumptions made throughout the work, respectively. In section IV, we describe the queue, performance, and multi-rate system models implemented. Section V shows the analytical and simulation results and, finally section VI, presents the conclusions to the work done here.

II. Overview of Assured Forwarding within Differentiated Services

The DiffServ architecture uses the Type-of-Service (TOS) field in the IP header to classify flows. The architecture is scalable because it does not maintain a per-flow state and there is no requirement for end-to-end signaling. In the DiffServ model, the TOS byte (or DS byte) is divided into a 6-bit Differentiated Service Code Point (DSCP) and a 2-bit unused field [8]. DiffServ is realized by mapping the DSCP contained in the IP packet header to a particular treatment or per-hop behavior (PHB) at each network station along its path. Packets marked with the same PHB class would experience similar forwarding behavior in the core station. PHBs are typically implemented by different types of buffer management and packet scheduling techniques.

To date, two additional per-hop behaviors have been defined: the Expedited Forwarding PHB (EF PHB) and the Assured Forwarding PHB (AF PHB). EF PHB, which has high priority, is used to provide low-loss, low-latency, low-jitter, assured-bandwidth and end-to-end service. AF PHB, on the other hand, gives the customer the assurance of a minimum throughput, even during periods of congestion. AF PHB has four classes and three-drop precedence per class. AF is incorporated with the extension of RED with IN/OUT or RIO scheme [9], which uses a single first-in-first-out (FIFO) queue and two-drop precedences.

RIO, a variant of RED, achieves this by classifying packets as being inside (IN) or outside (OUT) depending on whether they conform to the allocated bandwidth profile. RIO can be viewed as the combination of two RED instances that provides different levels of drop precedence for two classes of traffic. Two sets of thresholds and packet dropping probabilities are selected so that OUT packets are dropped more aggressively than IN packets. Consequently as congestion sets in, OUT packets will be more likely to be dropped, preserving the QoS of the IN traffic.

Applying dropping as the packet marking principle in RIO, the probability that an arriving class i packet (where i is either Class A or Class B) is accepted into the queue containing k packets is defined as:

$$\alpha^i(k) = \begin{cases} 1 & k \leq min^i \\ 1 - \frac{d^i\ (k - min^i)}{(max^i - min^i)} & min^i < k \leq max^i \\ 0 & k > max^i \end{cases} \quad (1)$$

Here, d^i, min^i and max^i represent the maximum drop probability and the minimum and maximum thresholds, respectively.

To model the AF service, we define p as the probability that a packet is an IN packet and $\overline{p} = (1-p)$ is the probability of a packet to be an OUT packet. $\alpha(k)$ is the probability of all packet to be accepted into the queue.

$$\alpha(k) = p\ \alpha^{Class\ A}(k)\ +\ \overline{p}\ \alpha^{Class\ B}(k) \quad (2)$$

III. Assumptions

The analytical model developed in this paper is based on the following assumptions.

- We consider two classes of packet, namely Class A and B packets where Class A packets or voice calls have higher priority than Class B packets or data sources. In general, voice transmission has a low data rate requirement with stringent delay constraints, while data transmission demands higher rates with less stringent delay requirements.
- The arrival of Class A packets are modelled as a Poisson process with a negative exponential distribution and mean arrival rate β_l.
- Class B packets can be modelled by a two state Markov Modulated Poisson Process (MMPP), which is a doubly stochastic Poisson process where the rate process is determined by the state of a continuous-time Markov Chain (a high state and a low state). The mean sojourn times in the high and low state are $\frac{1}{r_1}$ and $\frac{1}{r_2}$, respectively. In the high state, packets are generated with a mean rate of β_1 pkts/sec while in the low state, it is β_2 pkts/sec. The underlying Markov chain whose state space is $S = \{s_1, s_2\}$ can be represented by its infinitesimal generator matrix Q_{MMPP} and rate matrix Λ as shown in (3) and (4) respectively.

$$Q_{MMPP} = \begin{pmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{pmatrix} \quad (3)$$

$$\Lambda = \begin{pmatrix} \beta_1 & 0 \\ 0 & \beta_2 \end{pmatrix} \quad (4)$$

- Packets arriving at the transmitter enter the transmit buffer and are transmitted in a FIFO fashion, with a general service time distribution. It is assumed that the service times for both classes are independent and exponentially distributed based on the transmission time of the wireless link.
- We assume that the traffic sources are Transport Control Protocol (TCP) sources where each TCP segment is 512 bytes long.
- Finite-state Markov chain (FSMC) models [10], [11] are used to characterize the time-varying wireless channels whose bit-error rates (BERs) vary dramatically according to the received signal-to-noise ratio (SNR). The fitting is performed by partitioning the range of received signal-to-noise (SNR) into a set of non-overlapping intervals (states).
- We assume that multiple transmission modes are available, with each mode representing a pair of a specific modulation format, and a forward error correcting (FEC) code. Convolutional coded M_n-ary rectangular or square QAM are adopted here.

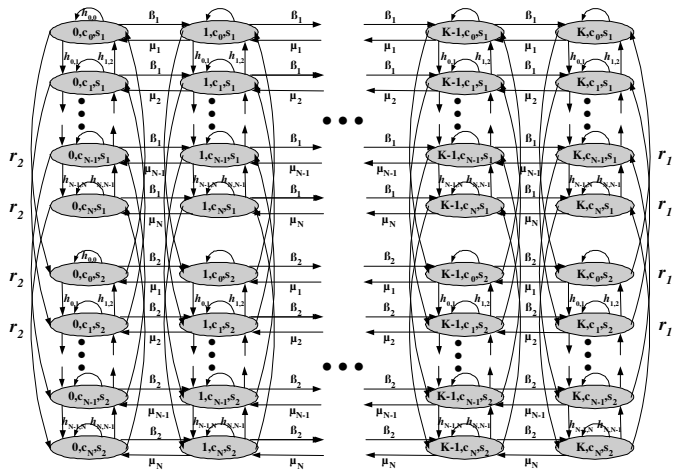


Fig. 1. State Transition Diagram

IV. Model and Analysis

A. Queue System Model

The overall system can be described by a three dimensional state transition diagram as shown in Figure (1). Each state is represented by a vector (k, c, s) where, k is the number of packets in a common buffer of size K in the device (including the ones in service and the ones in the queue buffer), c is the current Channel State Information (CSI) of the wireless link (a total of N states), s is the MMPP state of the source. In the steady-state condition, each state's probability is expressed as $p(k, c, s)$. The steady state probability distribution, $\underline{p}$, is determined by the following global balance equations:

$$\underline{p}\, Q_{BUFFER} = 0 \tag{5}$$

$$\Sigma\, \underline{p} = 1 \tag{6}$$

where

$$\begin{aligned}\underline{p} = [&p\,(0, n, s_1), p\,(1, n, s_1), ...p\,(K, n, s_1),\\ &p\,(0, n, s_2), p\,(1, n, s_2), ...p\,(K, n, s_2)]\end{aligned} \tag{7}$$

The system can be treated as a Continuous Time Markov Chain (CTMC) process. The CTMC describing this queuing model is a Quasi-Birth-Death (QBD) model. According to Matrix Analytical Methods (MAM) [12], the steady state probability of CTMC can be solved by exploiting the matrix-geometric properties. This is used to define the Q_{BUFFER} matrix that is sparse and block partitioned as shown in Equation (8)

Each of the square matrices B_0, B_1, A_0, A_1, A_2 has a dimension of 2(N - n + 1) by 2(N -n + 1)

$$Q_{BUFFER} = \begin{pmatrix} B_0 & A_0 & 0 & 0 & 0 \\ B_1 & A_1 & A_0 & 0 & 0 \\ 0 & A_2 & A_1 & A_0 & 0 \\ 0 & 0 & A_2 & A_1 & A_0 \\ & & \ddots & \ddots & \ddots \end{pmatrix} \tag{8}$$

B. Analysis for Performance Metrics

We can now evaluate the following performance metrics: packet drop probability, effective throughput and packet queuing delay denoted by Drop, EFFThrput and $Packet^{DELAY}$, respectively. Based on the Arrivals See Time Averages (ASTA) principle in queuing theory [13], these metrics can be given for Class A, Class B and All packets as follows:

$$Drop^{ClassA} = 1 - \sum_{k=0}^{K}\sum_{n=1}^{N} \alpha^{ClassA}(k)[p(k, n, s_1) + p(k, n, s_2)] \tag{9}$$

$$Drop^{ClassB} = 1 - \sum_{k=0}^{K}\sum_{n=1}^{N} \alpha^{ClassB}(k)[p(k, n, s_1) + p(k, n, s_2)] \tag{10}$$

$$Drop^{All} = 1 - \sum_{k=0}^{K}\sum_{n=1}^{N} \alpha(k)[p(k, n, s_1) + p(k, n, s_2)] \tag{11}$$

The corresponding effective throughput for the packet is:

$$\begin{aligned}EFFThrput^{ClassA} = \lambda_p \sum_{k=0}^{K}\sum_{n=1}^{N} &\alpha^{ClassA}(k) \cdot \\ &[p(k, n, s_1) + p(k, n, s_2)]\end{aligned} \tag{12}$$

$$\begin{aligned}EFFThrput^{ClassB} = \overline{\lambda_p} \sum_{k=0}^{K}\sum_{n=1}^{N} &\alpha^{ClassB}(k) \cdot \\ &[p(k, n, s_1) + p(k, n, s_2)]\end{aligned} \tag{13}$$

$$EFFThrput^{All} = \lambda \sum_{k=0}^{K}\sum_{n=1}^{N} \alpha(k) \cdot [p(k, n, s_1) + p(k, n, s_2)] \tag{14}$$

Using Little's formula, the average delay experienced by the packets in traversing the network is

$$Packet^{DELAY} = \frac{\sum_{k=0}^{K}\sum_{n=1}^{N} k[p(k,n,s_1) + p(k,n,s_2)]}{EFFThrput} \tag{15}$$

where, $\alpha^{ClassA}(k)$, $\alpha^{ClassB}(k)$ and $\alpha(k)$ are the probabilities of Class A, Class B and All packets to be accepted into the queue, respectively.

C. Analysis for Channel State Information (CSI)

1) Let $CSI = \{c_0, c_1, ..., c_{N-1}, c_N\}$ denote the state space of the Markov modulated wireless link with stationary transitions for N states. The state space CSI is that of N different transmission modes with corresponding bit-error rate (assuming that the transitions only happen between adjacent states). For flat fading channels, where it varies from frame to frame, the general Nakagami-m model is adopted [14]. The probability density function (pdf) of the received signal-to-noise ratio (SNR), γ per frame, is given as

$$f_\gamma(\gamma) = \frac{m^m \gamma^{m-1}}{\overline{\gamma}^m \Gamma(m)} exp(-\frac{m\gamma}{\overline{\gamma}}) \tag{16}$$

where, $\overline{\gamma} \approx E\{\gamma\}$ is the average received SNR; $\Gamma(m) \approx \int_0^\infty t^{m-1} exp(-t)dt$ is the Gamma function; and m is the Nakagami fading parameter ($m \geq 1/2$).
The probability that the link is in state c_n at time t is

$$\begin{aligned} q_n &= \int_{\gamma_n}^{\gamma_{n+1}} f_\gamma(\gamma)d\gamma \\ &= \frac{\Gamma(m, m\gamma_n/\overline{\gamma}) - \Gamma(m, m\gamma_{n+1}/\overline{\gamma})}{\Gamma(m)} \end{aligned} \tag{17}$$

where, $\Gamma(m,x) \approx \int_x^\infty t^{m-1} exp(-t)dt$ is the complementary incomplete Gamma function.

2)Since transition happens only between adjacent states, probability of transition exceeding two consecutive states [15] is

$$h_{m,n} = 0 \qquad |m - n| \geq 2 \tag{18}$$

The adjacent-state transition probability [15] is

$$\begin{aligned} h_{n,n+1} &= \frac{N_{n+1}T_f}{q_n} \qquad n = 0, ..., N-1 \\ h_{n,n-1} &= \frac{N_n T_f}{q_n} \qquad n = 1, ..., N \end{aligned} \tag{19}$$

where, T_f is the frame duration; N_n is the cross-rate of mode n (either upward or downward), which can be estimated [16] as

$$N_n = \sqrt{2\pi \frac{m\gamma_n}{\overline{\gamma}}} \frac{f_d}{\Gamma(m)} (\frac{m\gamma_n}{\overline{\gamma}})^{m-1} exp(-\frac{m\gamma_n}{\overline{\gamma}}) \tag{20}$$

where, f_d denotes the mobility-induced Doppler spread.

The probability of staying at the same state [10] is

$$h_{n,n} = \begin{cases} 1 - h_{n,n+1} - h_{n,n-1} & \text{if } 0 < \text{n} < \text{N} \\ 1 - h_{0,1} & \text{if n = 0} \\ 1 - h_{N,N-1} & \text{if n = N} \end{cases} \tag{21}$$

The channel state transition matrix $\{(N+1) \text{ x } (N+1)\}$, is banded as

$$\mathbf{h}_c = \begin{bmatrix} h_{0,0} & h_{0,1} & \dots & 0 \\ h_{1,0} & h_{1,1} & h_{1,2} & \vdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & h_{N-1,N-2} & h_{N-1,N-1} & h_{N-1,N} \\ 0 & \dots & h_{N,N-1} & h_{N,N} \end{bmatrix} \tag{22}$$

V. Performance Analysis

A. Simulation Setup

The simulation is performed using the ns-2 simulator. Matlab analysis is also used to approximate (9) - (15) to validate the accuracy of our model. We consider a simulated multi-hop network with 10 mobile ad-hoc stations where each station has a transmission range of 250m. We use a random waypoint mobility model [17] in which each mobile station selects a random destination at an arbitrary speed up to a maximum speed of 20 m/s and pauses for a given pause time when the destination is reached. When the pause timer expires, the mobile station picks another destination and speed randomly throughout the simulation duration of 500 seconds. The network area has a rectangular shape covering 600m x 400m that minimizes the effect of network partitioning. AODV [18] is used for routing in the simulated network. Since the analytical model is based on a Markov Modulated Arrival Process (MMAP), two different types of traffic are used: (i) Poisson arrivals, and (ii) Markovian traffic consisting of a superposition of 32 independently Pareto distributed on/off sources with parameter a = 1.4 or a Hurst parameter H = 0.8. Source traffic is sent via a one-way TCP Reno protocol, with an ACK sent for each packet. The maximum number of retransmissions is three. The RIO active queue management scheme, which provides different levels of drop precedence for two classes of traffic, is used. We classified Class A packets as IN and Class B packets as OUT. According to RIO when congestion occurs, OUT packets are discarded with a higher probability than IN packets, thus protecting the QoS of the IN traffic.

The IEEE 802.11a and 802.11b MACs provide a physical-layer multi-rate capability where higher data rates than base rate (2 Mbps) are possible when the signal-to-noise ratio (SNR) is sufficiently high. Here, we adopt the Opportunistic Media Access (OAR) [19] to opportunistically send multiple back-to-back data packets whenever the channel quality is good.

B. Results and Analysis

As shown in Figure 2 - 4, the simulation results almost exactly match the analytical results. In Figure (2), its shows the packet drop probability against traffic offered load. As we can see the drop probability for Class B packets rapidly increase towards 0.98 when offered load is 350 kbit/s. These results seem logical as the main idea of RIO is to protect the high priority packets when the offered load is high. This can be further explained that RIO begins to drop packets prematurely even before the buffer size exceeds it limit to alleviate the congestion within the network so that more resources can be shared among the higher priority packets.

Figure (3), depicts the effective throughput of both traffic classes where Class A packets increase steadily, saturating at 0.38 Mbit/s when offered load is increased beyond 370 kbit/s. However, due to the increased dropping rate, only small amount of Class B packets are received at the destination.

Since Class A traffic represents real-time traffic that has stringent delay requirements, we can see that packet delay in Figure (4) is maintained at 0.8 despite an increase in offered load. The higher delay experienced by Class B traffic are those minority packets that managed to be delivered to the destination. By evaluating the system beyond it s theoretical capacity, we are able to achieve different performance for the two independent traffic classes.

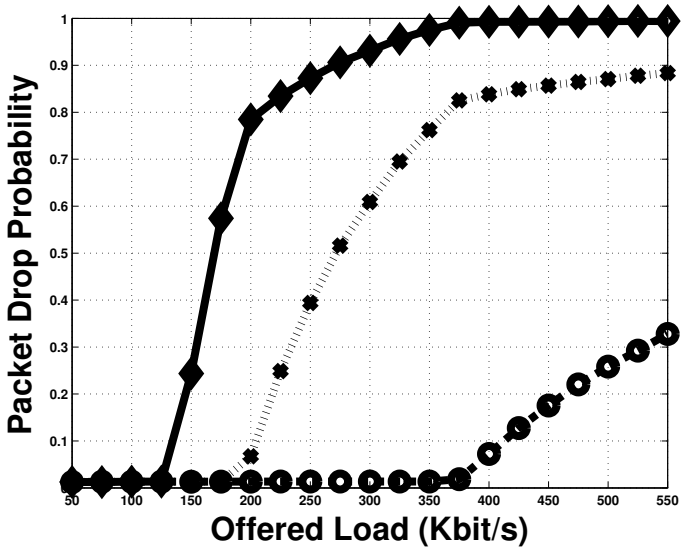


Fig. 2. Packet Drop Probability versus Offered Load

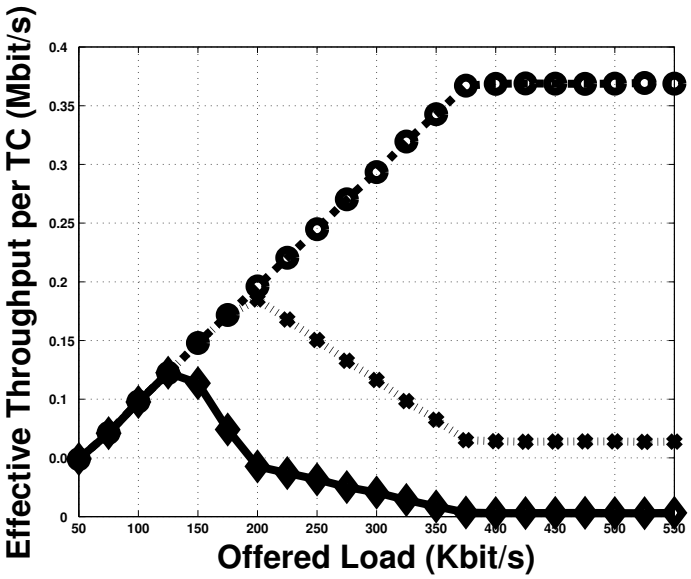


Fig. 3. Effective Throughput versus Offered Load

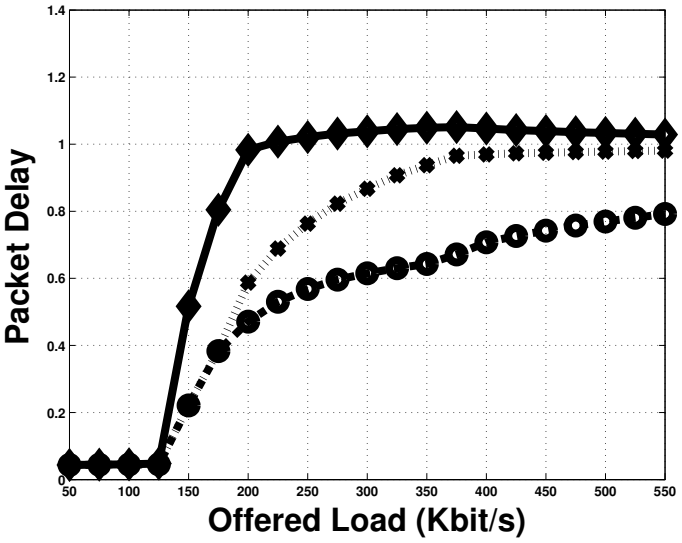
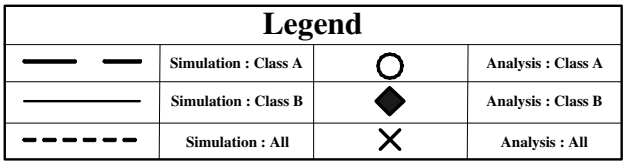


Fig. 4. Packet Delay versus Offered Load



VI. Conclusion

This paper has presented an analytical wireless DiffServ model where the RIO scheme is used to achieve simultaneous low packet drop rate and packet queuing delays. Arrivals are modelled as a general batch Markov arrival process in which the thresholds and packet dropping probabilities are selected so that real-time and non-real-time traffic observe different QoS performance. We have also considered the impact of the varying physical-layer link speed in a realistic MANET environment.

Acknowledgements

The work reported here has formed part of the Personal Distributed Environment of the Core 3 Research Programme of the Virtual Centre of Excellence in Mobile and Personal Communications Mobile VCE (http://www.mobilevce.com) whose funding support is grateful acknowledged.

References

[1] S. Blake et al., "An Architecture for Differentiated Services", *IETF RFC 2475*, December 1998.

[2] J. Heinanen, F. Baker, W. Weiss and J. Wroclawski, "Assured forwarding PHB", *IETF RFC 2597*, June 1999.

[3] S. Floyd and V. Jacobson, "Random early detection gateways for congestion avoidance", *IEEE/ACM Transactions on Networking*, vol. 1, no. 4, pp. 397 - 413, August 1993.

[4] H. Xiao, W. G. Seah, A. Lo and K. C. Chua, "A flexible quality of service model for mobile ad hoc networks", *in Proceedings of IEEE Vehicular Technology Conference (VTC 2000 - Fall)*, vol. 1, no. 4, pp. 397 - 413, May 2000.

[5] S. B. Lee and A. T. Campbell, "INSIGNIA: In-band signalling support for QOS in mobile ad hoc networks", *in Proceedings of 5th International Workshop on Mobile Multimedia Communications (MoMuC 1998)*, vol. 1, no. 4, pp. 397 - 413, October 1998.

[6] M. Kazantzidis, M. Gerla and S. Lee, "Permission throughput network for adaptive multimedia in AODV MANETs", *in Proceedings of IEEE Internation Conference on Communications (ICC 2001)*, vol. 6, pp. 1900 - 1904, June 2001.

[7] IEEE 802 LAN/MAN Standards Committee, "Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications", IEEE Std 802.11, 1999.

[8] K. Nicholas, S. Blake, F. Baker and D. L. Black, "Definition of the Differentiated Services Field in the IPv4 and IPv6 headers", *IETF Draft, draft-ietf-diffserv-header-04.txt*, October 1998.

[9] D. D. Clark and W. Fang, "Explicit allocation of best-effort packet delivery service", *IEEE/ACM Transactions on Networking*, vol. 6, no. 4, pp. 362 - 373, August 1998.

[10] H. S. Wang and N. Moayeri, "Finite-state Markov channel - A useful model for radio communication channels," *IEEE Transactions on Vehicular Technology*, vol. 44, no. 1, pp. 163 - 171, February 1995.

[11] H. Steffan, "Adaptive generative radio channel models," *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, vol. 1, pp. 268 - 273, September 1994.

[12] G. Latouche and V. Ramaswami, "Introduction to Matrix Analytic Methods in Stochastic Modeling", 1st edition, *ASA-SIAM Series on Statistics and Applied Probability*, Philadelphia, USA, 1999.

[13] B. Melamed and W. Whitt, "On arrivals that see time averages", *Ops. Res. Society of America*, vol. 38, no. 1, pp. 156 - 172, January 1990.

[14] G. L. Stuber, "Principles of Mobile Communication", *Norwell, MA: Kluwer Academic*, 2nd Edition, 2001.

[15] J. Razavilar, K. J. R. Lui and S. I. Marcus, "Jointly optimised bit-rate/delay control policy for wireless packet networks with fading channels", *IEEE Transactions on Communications*, vol. 50, No. 3, pp. 484 - 494, March 2002.

[16] M. D. Yacoub, J. E. V. Bautistu and L. Guerra de Rezende Guedes, "On higher order statistics of the Nakagami-m distribution", *IEEE Transactions on Vehicular Technology*, vol. 48, No. 3, pp. 790 - 794, May 1999.

[17] J. Broch, D. A. Maltz, D. B. Johnson, Y. C. Hu and J. Jetcheva, "A Performance Comparison of Multi-Hop Wireless Ad Hoc Network Routing Protocols", *in Proceedings of 4th Annual IEEE/ACM Internation Conference on Mobile Computing and Network (MobiCom 1998)*, pp. 1 - 13, October 1998.

[18] C. E. Perkins and E. M. Royer, "Ad-hoc On Demand Distance Vector Routing", *in Proceedings of 2th IEEE Workshop on Mobile Computing Systems and Applications (WMCSA 1999)*, pp. 90 - 100, February 1999.

[19] B. Sadeghi, "Opportunistic Media Access for Multirate Ad Hoc Networks," *in Proc. of ACM MOBICOM '02*, pp. 90 - 100, September 2002.