Delay analysis for real-time streaming media in multi-hop ad hoc networks

Ralph El-Khoury
LIA/CERI, University of Avignon,
Agroparc BP 1228,
84911 Avignon, France.
Email: ralph.elkhoury@univ-avignon.fr

Rachid El-Azouzi
LIA/CERI, University of Avignon,
Agroparc BP 1228,
84911 Avignon, France.
Email: rachid.elazouzi@univ-avignon.fr

Eitan Altman
INRIA
BP 93, 2004 Route des Lucioles
06902 Sophia Antipolis, France.
Email: altman@sophia.inria.fr

***Abstract*—In this paper, we investigate an important issue for real-time multimedia over multi-hop ad-hoc networks on different layers. Such application requires receiver playback buffers to smooth network delay variation and reconstruct the periodic nature of the transmitted packets. Packets arriving after their schedule deadline are considered late and are not played out. This requires that the network is able to offer quality of service appropriate for the delay bounds of the real-time application constrains. Our primary contribution concerns the study of the end to end delay. Based on the latter results, we approximate the loss rate of a real-time traffic which requires a delay bounds constraints. Then, we propose a cross layer scheme from the network layer to the MAC layer to support real-time traffic. In this scheme, we reduce the loss rate by decreasing the packets that arrive after their schedule deadline. Through numerical and simulation results, we demonstrate the utility and efficiency of our approach.**

I. Introduction

A multi-hop wireless ad hoc network is a collection of nodes that communicate with each other without any established infrastructure or centralized control. Many factors interact with each other to make the communication possible like routing protocol and channel access. With the emerging of real-time applications in wireless networks, delay guarantees are increasingly required. In order to provide support for delay sensitive traffic in such network, an accurate evaluation of the delay is a necessary first step. Knowing the nature of the multi-hop ad hoc networks, many factors are crucial for the study of the end to end (e2e) delay. We cannot study separately the delay generated by a given layer without considering the others.

In this paper we investigate an important issue for real-time multimedia over multi-hop ad-hoc networks on different layers. In particular, we study the audio quality in interactive multimedia applications. Audio packets encounter variable delay while crossing the multi-hop ad-hoc network, which is mainly due to the variable queueing time in intermediate nodes. In order to play the receiver stream, an application must buffer the packets and play them out after a certain deadline to get again a periodic stream at the application level. Packets arriving after their corresponding deadline are considered lost and are not played out. Hence if the end to end delay increases, the number of packets arriving after their schedule deadline increases. In this paper, with a playout delay constraint, we propose a novel mechanism which reduces the loss rate of packets arriving after their scheduled.

We consider the framework of random access mechanism for wireless channel where the nodes having packets to transmit in their transmit buffers attempt transmissions by delaying the transmission by a random amount of time. We assume that time is slotted into fixed length time frames. In any slot, a node having a packet to be transmitted to one of its neighboring nodes decides with some fixed probability in favor of a transmission attempt. If there is no other transmission by the other nodes whose transmission may interfere with the node under consideration, the transmission is successful. As examples of this mechanism, we find Aloha and CSMA type protocols. In the heart of our work, we have considered a parameter that measures the aptitude of a node to forward packets coming from its neighbors. At any instant of time, a node may have two kinds of packets to be transmitted: (1) packets generated by the node itself: data or control packets, and (2) packets from other neighboring nodes that need to be forwarded. Yet, we consider two separate queues for these two types of packets and do a weighted fair queueing (WFQ) for these two queues. This type of configuration allows us to include in the model the cooperation level in forwarding packets. A cooperation between nodes to achieve optimal performances is essential. This paper can be studied from the perspective of game theory where the nodes are rational and can adjust their parameters judiciously. Many papers in the literature have studied the problem of cooperation in ad hoc networks, see [14], [15].

In [1] and [2], working with the above mentioned system model, we have already studied the impact of routing, channel access rates and weights of the weighted fair queueing on throughput, stability and fairness properties of the network. We obtained important insights into various tradeoffs that can be achieved by varying certain network parameters. The throughput maximization of the multi-hop wireless networks has been extensively studied in [4] and [5]. However, it is shown that the high throughput in the ad-hoc network is achieved at the cost of a high amount of delay. This problem has drawn our attention to the relation between the delay characteristic and the throughput. Moreover, most of the related study does not consider the problem of forwarding. In most recent literature,

the throughput, delay, energy consumption and the tradeoffs generated from these, have been investigated as a key measure of the network performance. A random network model was proposed by Gupta and Kumar [6] to study the capacity of ad hoc wireless networks and they have shown that the maximum throughput per-node of a static random ad hoc network scales in terms of the number of nodes: it decreases as the number of nodes increases. After that, considerable efforts was made to compute the capacity of the network by introducing several network characteristics, assumptions and constraints, see [7]–[9]. However, throughput analysis cannot be separated from the queue stability of nodes nor the end to end delay analysis. We can achieve the maximum throughput per-node but nodes may be unstable or the delay may tend to infinity. Therefore, a comprehension of the throughput-delay tradeoff is very important to achieve connections quality of service for different types of applications. In [10]–[13], the authors have studied this problem in static and mobile networks and different optimal throughput-delay tradeoffs are analyzed.

After the description of the network model in section II, we proceed in section III to the end to end delay analysis in multi-hop ad-hoc networks. We focus on the asymptotic proprieties of the delay and we obtain an analytical delay estimates. We use a Geo/G/1 queueing model to compute the waiting time of intermediate nodes between a source and a destination. However, we have verified via simulation that our model is pertinent in section V. Based on the analysis, in section IV we compute the rate of packets arriving before its scheduled playout time (delay constraint). We also present a cross layer framework to improve the audio quality. To draw benefit from the interaction of the MAC and routing layer to ameliorate the effective throughput, we consider our cross-layer scheme already presented in [3]. It consists on adjusting dynamically and judiciously the limit number of (re-) transmissions on each node and for each connection. The limit number of transmissions of each intermediate nodes depends on the number of hops between a source and destination and the delay constrains imposed by the playout. In addition, section V presents an evaluation of the performance and conclusion remarks.

II. Network Model

We model the ad hoc wireless network as a set of nodes deployed arbitrarily in a given area. We describe in the following the different implemented layers in each node and the related assumptions:

A. Physical Layer

We consider a one simple channel where the nodes use the same frequency for transmitting with an omni-directional antennas. A node j receives successfully a packet from a node i if and only if there is no interference at the node j due to another transmission on the same channel i.e. if there is no transmission from any node of the set $N(j) \cup j$ where $N(j)$ is the set of neighbors of node j. We assume that all the nodes in $N(j)$ has j as a neighbor. Note also that a node cannot receive and transmit at the same time.

B. MAC layer

We assume a channel access mechanism only based on a probability to access the network i.e. when a node i has a packet to transmit from the queue Q_i or F_i, it accesses the channel with a probability P_i. For example, in IEEE 802.11 DCF, the transmission attempt probability is given by [17]

$$P = \frac{2(1-2\gamma)}{(1-2\gamma)(CW_{min}+1) + \gamma CW_{min}(1-(2\gamma)^m)},$$

where γ is the conditional collision probability given that a transmission attempt is made, and $m = log_2(\frac{CW_{max}}{CW_{min}})$ is the maximum of backoff stage. The transmission schedule overall the network depends on P_i. We assume that each node is notified about the success or failure of its transmitted packets. A packet is failure only when there is an interference on the intended receiver, in other terms, when a collision occurs on the receiver. We have considered previously infinite buffer size, therefore, there is no packet loss due to overflow at the queues. The only source of packet loss is due to collisions. For a reliable communication, we allow a limit number of successive transmissions of a single loosed packet, after that it will be dropped definitively.

C. Network layer

we assume that each node has two types of queues. The first one Q_i which carries the proper packets of the node i. The second one is the forwarded queue, noted by F_i, which carries the packets originated from source nodes and destined to destination nodes. We assume that each node has an infinite capacity of storage for the two queues. Packets are served with a first in first served fashion. When F_i has a packet to be sent, the node chooses to send it from F_i with a probability f_i. In other terms, it chooses to send from Q_i with probability $1 - f_i$. When one of these queues is empty then we choose to send a packet from the none empty one with a probability 1. Consider that each node has always packets to be sent from queue Q_i, whereas F_i can be empty. Consequently, the network is considered saturated and depends on the channel access mechanism. This assumption has permitted us to study the stability limit and property of the forwarding queues in [2].

Network layer handles the two queues Q_i and F_i using the WFQ scheme, as described previously. Also, this layer maintains routing algorithms. So, each node acts as a router, it permits to relay packets originated from a source s to a destination d. It must carries a routing information which permits sending of packets to a destination via a neighbor. In this paper, we assume that nodes form a static network where routes between any source s and destination d are invariant in the saturated network case. Proactive routing protocols as OLSR (Optimized Link State Routing) construct and maintain a routing table that carries routes to all nodes on the network. These kind of protocols corresponds well with our model. We

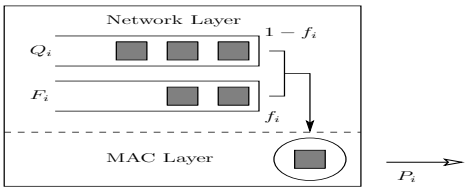


Figure 1. Network layer and MAC layer of node i

use the notation $R_{s,d}$ to denote the set of nodes between a node s and d (s and d not included).

D. Cross-layer representation of the model

The model of figure 1 represents our model. Different layers are clearly separated. Attempting the channel begins by choosing the queue from which a packet must be selected. And then, this packet is moved from the corresponding queue from the network layer to the MAC layer where it will be transmitted and retransmitted, if needed, until its success or drop. In this manner, when a packet is in the MAC layer, it is itself attempted successively until it is removed from the node.

We define a *cycle of transmissions* as the number of slots needed to transmit a single packet until its success or drop. We distinguish two types of cycles: The *forwarding cycles* related to the packets of F_i and the *source cycles* related to the packets coming from Q_i. Also, each cycle is affected to a connection. The beginning of each cycle represents the choice of the queue from which we choose a packet and the choice of the connection where to send it. Whereas, the slots that constitute the cycle represents the attempts of the packet itself to the channel, including its retransmissions. Hence, the distinction of the network and MAC layer is now clear.

E. Main Notation

We summarize the main notation of the paper in the following two lists:

1) MAC layer notation:
   - P_i is the probability of transmission on the channel of the node i.
   - $P_{i,s,d}$ is the probability that a transmission from node i on the path from s to d is successful.
   - $K_{i,s,d}$ is the maximum number of successive collision allowed for a single packet sent from the node i on the path from s to d. After a $K_{i,s,d}$ failure, the packet is dropped i.e. it is removed from the node i.
   - $L_{i,s,d}$ is the expected number of attempts till successful or a drop from node i on the path from s to d.
2) Network layer notation:
   - f_i is the prob ability to send a packet from the queue F_i when it carries a packet.
   - $R_{s,d}$ is the set of intermediate nodes in a path between a node s and a node d. s and d are not in this set.
   - $R_{i,s,d}$ is the set of nodes $R_{s,i} \bigcup i$ in the path s, d.
   - π_i is the probability that the queue F_i has at least one packet to be forwarded in the beginning of each cycle.
   - $\pi_{i,s,d}$ is the probability that the queue F_i has a packet at the first position ready to be forwarded to the path $R_{s,d}$ in the beginning of each cycle. Then, $\pi_i = \sum_{s,d:i\in R_{sd}} \pi_{i,s,d}$.

We use the notation $P_{i,d}$ to denote the probability that the node i chooses the path $R_{i,d}$ (whose destination is d) for sending packets from Q_i.

III. Delay Analysis

We have been interested to compute an accurate value of the average end to end packet delay of forwarding queues of each connection in an ad hoc network and to find closed-form expressions. For that, we have proposed two methods for delay calculation. The first one is based on Markov chains with three dimensions, from which we find the forwarding queue average size and delay of forwarded packets in each node. The second one is a decomposition method based residual service time which is used to find an expression of the delay function of all the network parameters.

The queue F_i is Geo/G/1 with the following characteristics: (1) Packets arrives according to a geometric process with average a_i given from [2]:

$$a_i = \sum_{s,d:i\in R_{s,d}} (1-\pi_s f_s).P_{s,d}.\frac{P_s}{\overline{L}_s}.\Big[(1-(1-P_{s,s,d})^{K_{s,s,d}}). \prod_{k\in R_{i,s,d}\setminus i} (1-(1-P_{k,s,d})^{K_{k,s,d}})\Big]$$

(2) The MAC layer has always a packet in service due to the presence of a saturated queue Q_i. We define the service time of a given packet in the MAC layer of node i as the time from the instant the packet reaches the MAC until it is successfully transmitted or dropped. (3) Service times of packets are identically distributed, independent of the arrival process and each other, with a general distribution.

A. Accurate Delay in a single node using Markov chain

An arriving packet to the queue F [1] has to wait the service time of all the packets in the queue F and some of them from Q. Each connection has its own service time which depends on the topology (number of neighbors), the transmission probability of nodes and the limit number of transmissions. In our method, we need to identify the connection identity of the packet being served in the MAC layer and we need also to know how much transmission has been accomplished by this same packet.

[1] In this sub-section, we will calculate the delay for a single node i, for that we omit the index i that identifies the node itself to facilitate the notation and the reading. In addition, we will represent a connection (s, d) by a single identity j. In other words, we note $X_i \equiv X$ and $X_{i,s,d} \equiv X_j$ where X can be one of the element of the set $\{\pi, K, P, f, F...\}$. For example, $P_i \equiv P$ and $P_{i,s,d} \equiv P_j$.

Formally, to describe the Markov chain, let $q(t)$ be the number of packets to be forwarded presents in the node at time t. Let $k(t)$ be the number of transmissions of a packet at time t. At each time the node serves a packet that corresponds to one of the connections passing through F or originated from Q. For that, we use the process $c(t)$ to distinguish between them. So, $c(t)$ identifies the connection number of a packet in the MAC layer. It is defined as $c(t) = j$ for $0 \leq j \leq \mathsf{C}$ where C is the maximum connection number. Also, let $c(t) = 0$ identifies the packet in the MAC that comes from Q.

It can be shown that $s(t) = (q(t), k(t), c(t))$ for $t \geq 0$ is a discrete time Markov chain when considering independent arrivals of packets. The arrival process to the node follows a geometric distribution with mean a. The finite state space is then $\mathbb{S} = \{(h,i,j) : h \geq 0; 1 \leq i \leq K_j; 0 \leq j \leq \mathsf{C}\} - \{(0,i,j) : 1 \leq i \leq K_j, 0 < j \leq \mathsf{C}\}$ where the queue size is infinite, the maximum limit of transmissions is K_j and the maximum number of connections is C. The set $\{(0,i,j) : 1 \leq i \leq K_j, 0 < j \leq \mathsf{C}\}$ corresponds to some cases that does not exist and cannot occur: when there are no packets to be forwarded in the node, it is impossible to have $c(t) > 0$.

For each node in the network, π and π_j are determined from the rate balance equations in steady state presented in paper [2]. They depend on the parameters and evolution of all the nodes in the network. π and π_j which are calculated using the notion of cycles, give a partial view of the queue F load. In this section, we will find the real load in queue F with a slot by slot vision.

Moreover, the probability to find a packet belonging to the forwarding connection j in the MAC layer is $\pi_j f$. And with a probability $y := 1 - \pi f$ this packet belongs to the node own connection. Recall that π_j includes also the probability that the connection j packet is in the first position. A node chooses a packet from F, when it exists, with probability f, then the probability that this packet is for connection j is $\phi_j f$ where $\phi_j := \frac{\pi_j}{\pi}$.

Our first objective is to find the stationary probability

$$s(h,i,j) = \lim_{t \to \infty} P\{q(t) = h, k(t) = i, c(t) = j\}$$

when this limit exist and where the combination of h, i and j corresponds to the states in the space $\mathbb{S}$. The transition probability matrix is denoted $\mathbf{P}$ with dimension $(\mathsf{C}+1) \times (\mathsf{C}+1)$. We will not give the details of the transition matrices due to lack of space.

The unique solution of the system $s(h,i,j)$ is determined from the following system:

$$\mathbf{sP} = \mathbf{s}, \quad \mathbf{s\overline{1}} = 1 \tag{1}$$

where $\mathbf{s}$ is the stationary probability vector. The solution of the system can be easily obtained numerically using Matlab tools and by entering the necessary parameters.

The exact probability to find a packet in the forwarding queue of a node i (when observing the F queue slot by slot) is given by

$$\pi^{slot} = 1 - \sum_{i=0}^{K_0} s(0,i,0) \tag{2}$$

The average number of packets to be forwarded (in the system) can be easily given by:

$$\overline{N} = \sum_{h=1}^{\infty} h. \sum_{j=0}^{\mathsf{C}} \sum_{i=1}^{K_j} s(h,i,j) \tag{3}$$

Therefore, the mean sojourn time of a forwarding packet in a given node i is obtained from Little's formula, and by reusing the initial notations of section II, we get: $\widehat{D}_i = \frac{\overline{N}_i}{a_i}$.

We derive a closed-form for the special case $f = 1$ and $K = 1$ (we use a simplified notation as previously). By resolving the system of equations given from equation (1), We find the following

$$s(i,1,0) = (\frac{a}{1-\beta})^i s(0,0,0) \tag{4}$$

$$s(i,1,1) = \frac{a^{i-1}}{P}[\frac{1-\beta}{P^{i-1}} - \frac{P}{(1-\beta)^{i-1}}]s(0,0,0) \tag{5}$$

Also, from the equation of normalization, we get ,

$$\pi^{slot} = \frac{2a}{P+a}, \quad \overline{N} = \frac{a(2P-a)}{P(P-a)}, \quad \widehat{D} = \frac{(2P-a)}{P(P-a)} \tag{6}$$

we can also derive π^{slot} in terms of π (where $\pi = \frac{a}{P}$ for $K = 1$ and $f = 1$, given from the stability condition where the arrival a equals the departure $\pi f \frac{P}{L}$) as follow:

$$\pi^{slot} = \frac{2\pi}{\pi + 1} \tag{7}$$

B. Delay in a single node using a decomposition method

They have been many works that had used a decomposition method based residual service time to find the waiting time in M/G/1 queues. Here, our purpose is to derive the expression of the delay of a Geo/G/1 forwarding queue with the presence of a saturated source queue Q_i. We are interested to find the delay function of several parameters belonging to different layers, so it will be easy to study their impact in a multi-hop wireless network.

Let, (1) $\overline{W}_i$ be the mean waiting time, (2) $\overline{N}_i^F$ be the mean number of packets in the queue F_i (without the MAC packet) (3) $\overline{\mathbf{R}}_i$ is the mean residual service time of a packet in MAC seen by an arrival packet. Let $\tau_{i,s,d}$ and $\tau_{i,s,d}^{(2)}$ be the first and second moment respectively corresponding to the path $R_{s,d}$. Therefore, the average service time of the forwarded connections and the node own connections are respectively:

$$\tau_i^F = \sum_{s,d} \frac{\pi_{i,s,d}}{\pi_i} \tau_{i,s,d} \quad \text{and} \quad \tau_i^Q = \sum_d P_{i,d} \tau_{i,i,d} \tag{8}$$

Then, an average service time for any packet in MAC layer is:

$$\tau_i = \pi_i f_i \tau_i^F + (1 - \pi_i f_i)\tau_i^Q = \frac{\overline{L}_i}{P_i} \tag{9}$$

where $\overline{L}_i$ is the average number of transmissions of a node i to any of its neighbors, see [2].

The waiting time of an arrival packet can be decomposed in two terms: (1) the mean residual time of packet in service

(2) the mean time to serve all packets that will be transmitted before it. It is clear that:

$$\overline{W}_i = \overline{\mathbf{R}}_i + \overline{\mathbf{B}}_i \tag{10}$$

Where $\overline{\mathbf{B}}_i$ corresponds to the second terms described above and can be decomposed according to the waiting time generated by packets from F_i and Q_i. In the following, we determine $\overline{\mathbf{R}}_i$ and $\overline{\mathbf{B}}_i$.

-Mean residual service time: an arrival packet to F_i can find in the service a packet corresponding to one of the possible paths $R_{s,d}$ for all (s,d). Then, when a packet in the service is for the route $R_{s,d}$ the mean residual service time is $\overline{\mathbf{R}}_{i,s,d}$. For any packet in service, the mean residual service time is:

$$\overline{\mathbf{R}}_i = \sum_{s,d} \pi_{i,s,d} f_i \overline{\mathbf{R}}_{i,s,d} + \sum_{d} P_{i,d}(1-\pi_i f_i)\overline{\mathbf{R}}_{i,i,d} \tag{11}$$

In the literature, the M/G/1 wire line queueing system gives an expression for the $\overline{\mathbf{R}}_{i,s,d}$, but in our wireless model it turns to be a little different because antennas cannot transmit and receive in the same time. For that we have derived the accurate expression of this residual time based on the renewal theory and the method presented in [16]. Therefore,

$$\overline{\mathbf{R}}_{i,s,d} = \frac{\tau^{(2)}_{i,s,d}}{2\tau_{i,s,d}} + \frac{1}{2} \tag{12}$$

see the appendix for a proof.

For a formal calculation of $\tau_{i,s,d}$ and $\tau^{(2)}_{i,s,d}$, and for a given node i and path $R_{s,d}$, we define a *mini-cycle* as the number of time slots needed to transmit a packet. So, a cycle is formed by L mini-cycles where L is a random variable representing the number of transmissions in a cycle. X_1, X_2,...X_L is a sequence of random variables representing the length of a mini-cycle for each transmission in a cycle. All variables have the same mass function $P\{X_i = x\} = (1-P)^{x-1}P$ (P is the transmission probability) which is a geometric distribution with an expectation $E[X] = \frac{1}{P}$ and a second moment $E[X^2] = \frac{2-P}{P^2}$. Let S be the random variable representing the time service of a packet in MAC layer. Therefore, $S = \sum_{i=1}^{L} X_i$ and $E[S] = E[L].E[X]$. Then, for a specified node i and path $R_{s,d}$, $E[S]$ can be written as

$$\tau_{i,s,d} = \frac{L_{i,s,d}}{P_i} \tag{13}$$

where $L_{i,s,d} = \frac{1-(1-P_{i,s,d})^{K_{i,s,d}}}{P_{i,s,d}}$, see [1].

To find the second moment $E[S^2]$, let $Y_l = \sum_{i=1}^{l} X_i$ be the service time for a given number of transmissions $L = l$. Then, from the conditional probability properties

$$E[S^2] = \sum_{l=1}^{K} E[Y_l^2] P\{L = l\} \tag{14}$$

where K is the maximum number of transmissions allowed. The calculation of $E[Y_l^2]$ gives $E[Y_l^2] = lE[X^2] + (l^2 - l)E[X]^2$. By replacing it in equation (14), we get,

$$E[S^2] = E[L^2]E[X^2] + E[L](E[X^2] - E[X]^2) \tag{15}$$

$E[L]$ the average number of transmissions is already known. For a given node i and path $R_{s,d}$, $E[L] \equiv L_{i,s,d}$ and $E[L^2] \equiv L^{(2)}_{i,s,d}$. This latter is given by $L^{(2)}_{i,s,d} = \sum_{l=1}^{K_{i,s,d}} l^2(1-P_{i,s,d})^{l-1} + K^2_{i,s,d}(1-P_{i,s,d})^{K_{i,s,d}}$, therefore,

$$L^{(2)}_{i,s,d} = L_{i,s,d} + \frac{2(1-P_{i,s,d})}{P^2_{i,s,d}} - \frac{2(1-P_{i,s,d})^{K_{i,s,d}}(K_{i,s,d} - (1-P_{i,s,d})(K_{i,s,d}-1))}{P^2_{i,s,d}}$$

By replacing the values of $E[X]$ and $E[X^2]$ in equation (15), and for a given node i and path $R_{s,d}$, we get:

$$\tau^{(2)}_{i,s,d} = \frac{L^{(2)}_{i,s,d} + L_{i,s,d}(1-P_i)}{P_i^2} \tag{16}$$

In this manner, we have all the necessary to get the residual service time $\overline{\mathbf{R}}_i$.

-Waiting time due to packets in buffer F_i and packets coming from Q_i: $\overline{\mathbf{B}}_i$ can be written as follow,

$$\overline{\mathbf{B}}_i = \overline{N}_i^F \tau_i^F + (\overline{N}_i^F + 1)\overline{n}_i^Q \tau_i^Q \tag{17}$$

Where $\overline{n}_i^Q$ be the mean number of Q_i packets that are served before a packet in the head of queue F_i.

After the departure of a forwarding packet, a head of queue F_i packet, if it exists, has to wait V (random variable) number of cycles for serving packets from Q_i before it can accesses the MAC layer. The probability to wait k cycles is: $P\{V = k\} = (1-f_i)^k f_i$. V is then a geometric distribution with expected value $E[V] \equiv \overline{n}_i^Q \approx \frac{1-f_i}{f_i}$. This is an approximation of n_i^Q since V cannot take very large values in practice.

From Little's formula, $\overline{N}_i^F = a_i \overline{W}_i$. Then, by replacing it in equation (17), we find the waiting time in the forwarding queue F_i from equation (10):

$$\overline{W}_i = \frac{\overline{\mathbf{R}}_i + \tau_i^Q \frac{1-f_i}{f_i}}{1 - a_i(\tau_i^F + \tau_i^Q \frac{1-f_i}{f_i})} \tag{18}$$

Therefore, the average total delay in a single node i is:

$$\widetilde{D}_i = \overline{W}_i + \tau_i^F \tag{19}$$

We add the service time τ_i^F to $\overline{W}_i$, not τ_i, because the delay time spent by the packet due to packets from Q_i is added in the expression $(\overline{N}_i^F + 1)\overline{n}_i^Q \tau_i^Q$. For an addition accuracy, a packet belonging to the path $R_{s,d}$ waits the waiting time in queue F_i and the service time of its corresponding path, thus $\widetilde{D}_{i,s,d} = \overline{W}_i + \tau_{i,s,d}$.

As we can see and guess from the expression (19), the delay is a decreasing function of f_i. one could believe that giving priority to the forwarding queue by increasing f may penalize the delay in the source queue Q_i. Recall that the saturation condition of the queue Q_i is mainly due to control packets and not necessary due to data traffic. The intuition of penalizing the delay is only true when the schedule mechanism between the two queues Q_i and F_i is symmetric, but it is not the case in our paper. In fact, when there are no packets in the queue

F_i, a packet is chosen from Q_i with probability 1, elsewhere the probability is $1-f$. Therefore, the rate of serving Q_i in layer 3 is $1-\pi_i f$ but the rate of serving F_i (in layer 3) is just f. From our analysis in paper [2], we have seen that y_i is independent on f, which means that if f varies the delay in Q_i remains unchanged with f. From this analysis, we deduce that setting $f=1$ is the best possible configuration with our scheduling mechanism. In fact, with $f=1$ the delay of the forwarding queue is maximized, whereas the throughput and energy consumption remain unchanged.

C. End to end Delay

Here, we define the average end to end delay of a packet on a path $R_{s,d}$ to be the average time from the instant the packet reaches the MAC layer of the source to the instant it is received by the destination. In previous sections, we have derived the average waiting time spent by a given forwarding packet in a node i without considering if this packet will be successfully transmitted or dropped at the end of the service in the MAC layer. Yet, this delay time is for both successful and dropped packets. However, in the e2e delay case, the dropped packets due to the finite number of transmissions must not be included in the calculation.

For that, in the decomposition method the e2e delay in a path $R_{s,d}$ can be written as:

$$\widetilde{D}_{s,d} = \frac{L^{succ}_{s,s,d}}{P_s} + \sum_{i=1}^{|R_{s,d}|} (\overline{W}_i + \tau^{succ}_{i,s,d}) \tag{20}$$

where $\tau^{succ}_{i,s,d}$ is the average service time of successfully transmitted packets in this same path $R_{s,d}$. $\tau^{succ}_{i,s,d}$ has the same form as $\tau_{i,s,d}$ and can be written:

$$\tau^{succ}_{i,s,d} = \frac{L^{succ}_{i,s,d}}{P_i} \tag{21}$$

where $L^{succ}_{i,s,d}$ is the average number of successful transmissions. It is given by,

$$\begin{aligned} L^{succ}_{i,s,d} &= \sum_{l=1}^{K_{i,s,d}} l(1-P_{i,s,d})^{l-1}P_{i,s,d} \\ &= \frac{L_{i,s,d} - K_{i,s,d}.(1-P_{i,s,d})^{K_{i,s,d}}}{(1-(1-P_{i,s,d})^{K_{i,s,d}})} \end{aligned} \tag{22}$$

In the markov chain method, let Δ_i be the average time to deduct from the delay $\widehat{D}_i$ in each node to get an accurate e2e delay of only successful packets that have reached the destination. Therefore,

$$\Delta_{i,s,d} = \tau^F_i - \tau^{succ}_{i,s,d} \tag{23}$$

Therefore, the e2e delay in a path $R_{s,d}$ can be written as:

$$\widehat{D}_{s,d} = \frac{L^{succ}_{s,s,d}}{P_s} + \sum_{i=1}^{|R_{s,d}|} (\widehat{D}_i - \Delta_{i,s,d}) \tag{24}$$

IV. Playout delay control in VoIP

Delay, jitter and packet loss are the main factors impacting audio quality in interactive multimedia applications. In ad-hoc network, the audio packets transmitted from a source to a destination can encounter variable delay while crossing the intermediate nodes. In order to play the receiver stream, an application must buffer the packets and play them out after a certain deadline to get again a periodic stream at the application level. Packets arriving after their corresponding deadline are considered lost and are not played out. For that, we need to fix a delay limit needed for some type of application.

Let $T_{s,d}$ be the maximum delay limit for a connection between s and d allowed for each packet in this connection. The instantaneous delay of each packet must not exceed $T_{s,d}$, if so, the destination node must drop this packet. For large $T_{s,d}$, we get less packet dropped and the end to end delay may be greater. Then clearly we see the tradeoff generated by setting the value of $T_{s,d}$. The end to end effective throughput of a given connection $R_{s,d}$ can be written as $thp_{s,d}.P(delay \leq T_{s,d})$ where $thp_{s,d}$ is the e2e throughput without the constraint on the delay, and $P(delay > T_{s,d})$ is the probability to drop a packet when its delay exceeds $T_{s,d}$. To facilitate the analysis, we consider the minimum effective throughput $thpu_{s,d}$ as a metric for measure of the quality of service when the constraint on the delay is respected, then $thpu_{s,d} = thp_{s,d}.(1-\frac{D_{s,d}}{T_{s,d}})$ where $P(delay > T_{s,d}) < \frac{D_{s,d}}{T_{s,d}}$. In addition, we have already find the throughput $thp_{s,d}$ in [2], then one could study the optimization of the effective throughput. To draw benefit from the interaction of the MAC and routing layer to ameliorate the effective throughput, we consider our cross-layer scheme already presented in [3]. It consists on adjusting dynamically and judiciously the limit number of (re-) transmissions $K_{i,s,d}$ on each node and for each connection. The following is a brief description of the scheme.

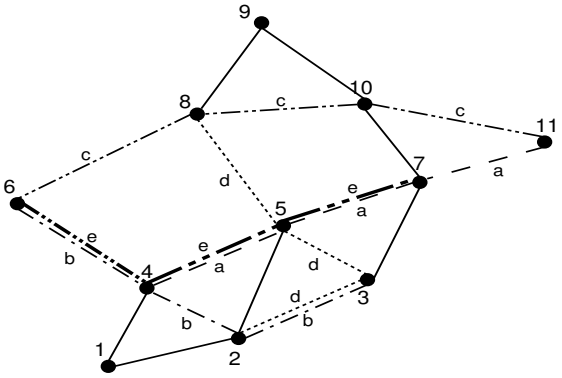
Consider that each node has a default value of the limit number of transmissions set to K. Each node i in the set $R_{s,d} \cup \{s\}$ computes the corresponding $K_{i,s,d}$ in such a manner that this latter is higher or equal to the previous $K_{j,s,d}$ where j is the previous node of i in the path $R_{s,d} \cup \{s\}$. Furthermore, the average values of $K_{i,s,d}$ (for $i \in R_{s,d} \cup \{s\}$) must be set to K i.e. $\frac{1}{|R_{s,d} \cup \{s\}|} \sum_{i \in R_{s,d} \cup \{s\}} K_{i,s,d} = K$. Also, the values of $K_{i,s,d}$ (for $i \in R_{s,d} \cup \{s\}$) are determined based on the position of the node i in the path $R_{s,d}$ i.e. it is based on the number of hop that separates it from the source or the destination. We add to this scheme a *reset technique* when the average queue size or the load of F_i exceeds some value. In fact, when the average queue value in dynamic case becomes not profitable in comparison to the static case, we reset the value of $K_{i,s,d}$ to K.

The first purpose of the dynamic scheme is to give more chance of success to packets that had come near to their destination, so the waste of bandwidth throughout a path becomes lower. The second purpose is to drop the packets

in the intermediate nodes instead of dropping them in the destination node due to the delay constraint.

V. Performance Evaluation

We consider an asymmetric static wireless network with 11 nodes as shown in figure 2. Five connections are established a, b, c, d and e as indicated in the same figure (a dashed or complete line between two nodes in this figure means that there is a neighboring relation). We choose the parameters $K_{i,s,d} \equiv K$, $f_i \equiv f$ and P_i in a manner of enabling stability, for all i, s and d. We fix $f = 0.8$ and $K = 4$ except contraindication. Let $P_2 = P_3 = P_7 = P_8 = 0.3$ $P_4 = P_{10} = 0.4$ and $P_5 = 0.5$ be the fixed transmission probabilities for nodes 2, 3, 4, 5, 7, 8 and 10 while $P_i \equiv P$ for all other i. Many nodes need to have a fix transmission probabilities so to get a stable queues for all nodes.



connection a : 4 - 5 - 7 - 11
connection b : 3 - 2 - 4 - 6
connection c : 11 - 10 - 8 - 6
connection d : 8 - 5 - 3 - 2
connection e : 6 - 4 - 5 - 7

Figure 2. Wireless network

A. Validation of the results

We present some numerical results and simulations to show the accuracy of the method studied previously. For that, a discrete time simulator that implements the model of section II is used to simulate the network and to validate the numerical results. In figures 3 to 9, we compare the two methods (the decomposition method labeled DM and the based Markov chain method labeled MC) presented previously with the simulation results (labeled $SIMU$) and with each other. The two methods are sufficiently close to each other and with the simulation results. This is also true in the case where nodes forward to different neighbors on different paths. In figures 4, 6 and 8, we show a zoom on figures 3, 5 and 7.

B. Playout delay Control with dynamic retransmission limit

We apply here the dynamic scheme to the connections of the network. We aim to compare the performances of the network with static and dynamic retransmissions and show what is happening in the network in the limit case where the loss rate is $P(delay > T_{s,d}) = \frac{D_{s,d}}{T_{s,d}}$.

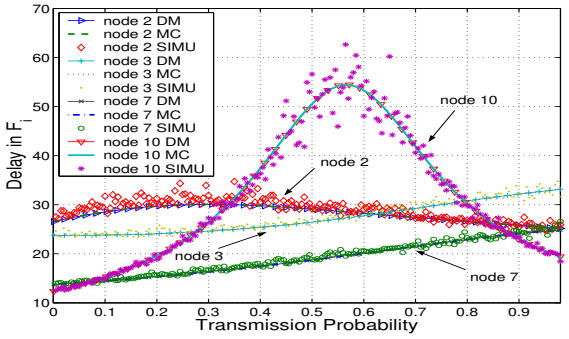


Figure 3. Delay in forwarding nodes 2, 3, 7 and 10 of the DM, MC and $SIMU$.

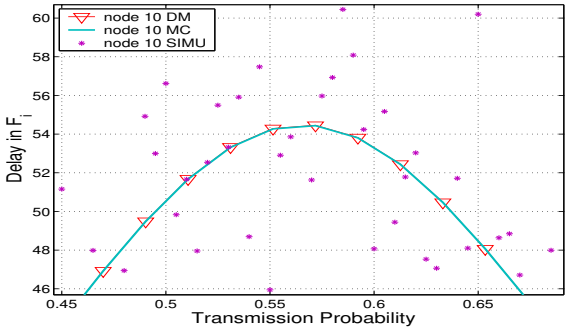


Figure 4. Zoom on figure 3.

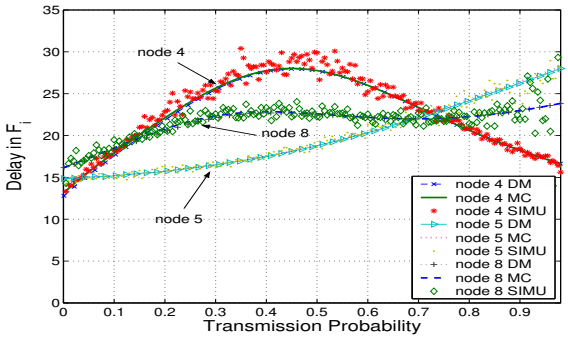


Figure 5. Delay in forwarding nodes 4, 5 and 8 of the DM, MC and $SIMU$.

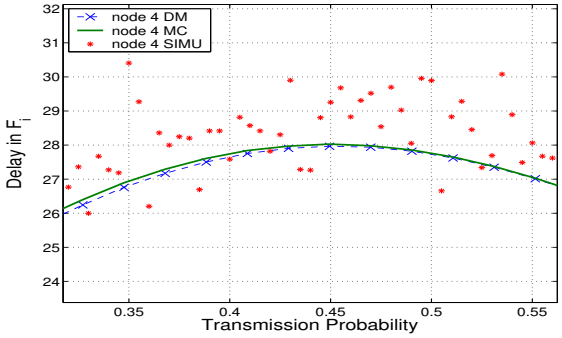
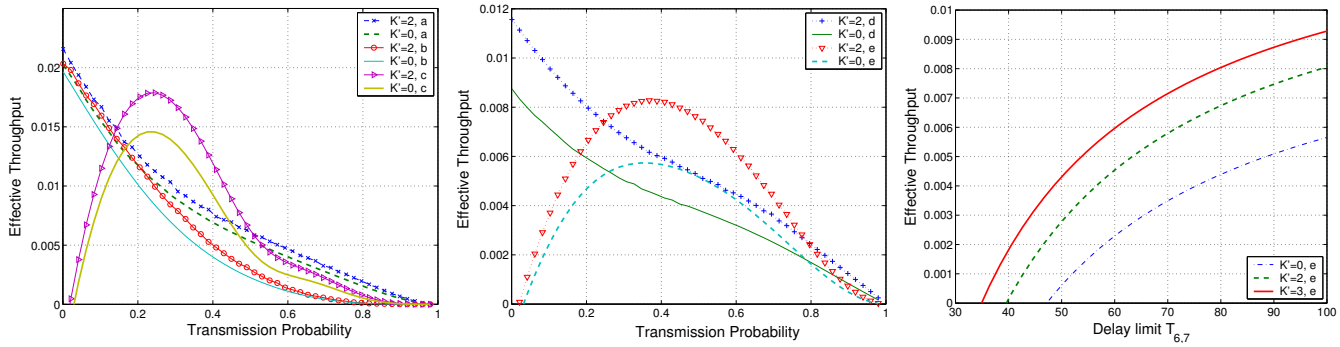


Figure 6. Zoom on figure 5.

Let K' be the step that we fix to increase the value of $K_{i,s,d}$



a b c

Figure 10. (a) and (b) show the minimum effective throughput of connections a, b, c, d and e versus the transmission probability for $T_{s,d} = 100$ for all the connections. (c) shows the minimum effective throughput of connection e versus the maximum delay limit $T_{6,7}$ for different configuration of the limit number of transmissions and $P = 0.305$.

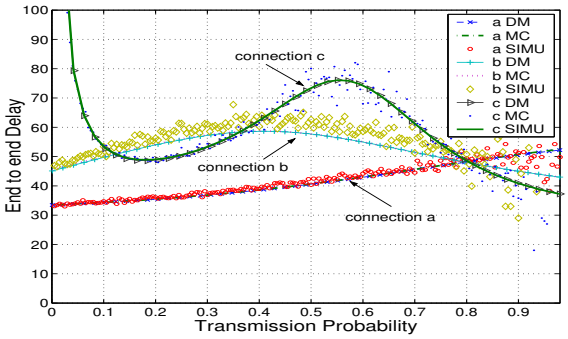


Figure 7. Connections a, b and c end to end delay of the DM of equation (20), MC of equation (24) and $SIMU$.

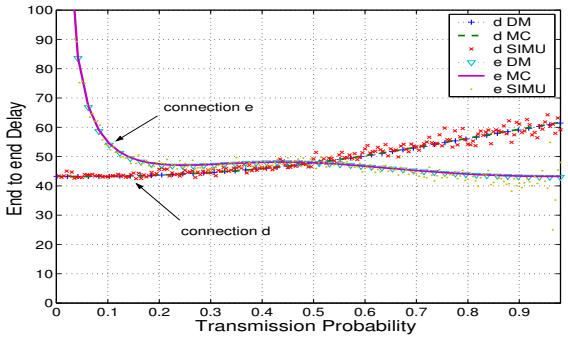


Figure 9. Connections d and e end to end delay of the DM of equation (20), MC of equation (24) and $SIMU$.

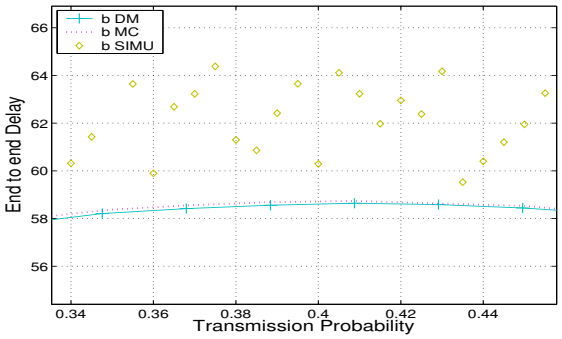


Figure 8. Zoom on figure 7.

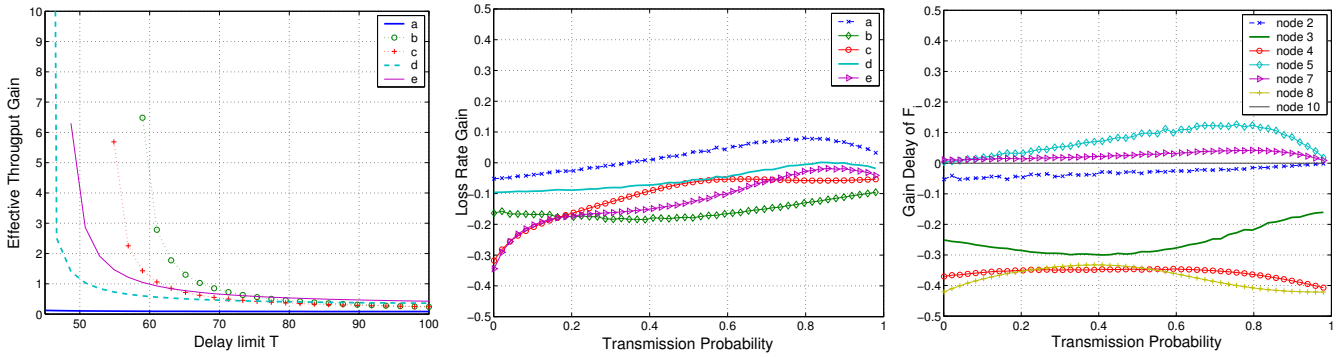
on each node of a given path $R_{s,d}$. For example, by setting $K' = 0$, all nodes have their limit number of transmissions set to K. By setting $K' = 2$ and $K = 4$, the connection a in figure 2 has: $K_{4,4,11} = 2$, $K_{5,4,11} = 4$ and $K_{7,4,11} = 6$. Remark that a node can have different limit number of transmissions for its different connections, for example the node 4 has: $K_{4,3,6} = 6$, $K_{4,6,7} = 4$ and $K_{4,4,11} = 2$. In figures 10 (a) and (b), we observe an amelioration of the effective throughput by configuring dynamically the limit number of transmissions.

Furthermore, in figure 10 (c), we plot the connection e effective throughput and observe an amelioration of the performances. The gain ratio of the static-dynamic retransmissions for different metrics is

$$\frac{X(K' \neq 0) - X(K' = 0)}{X(K' = 0)} \tag{25}$$

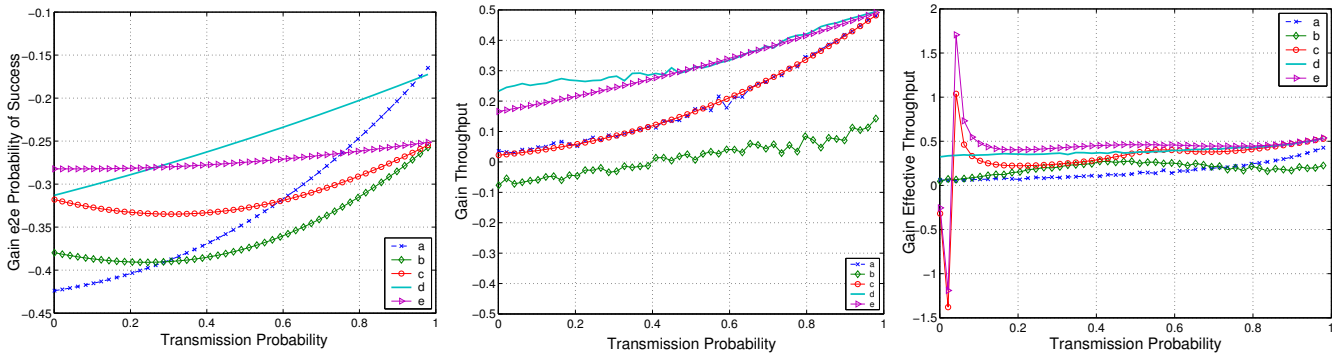
where $X(K' \neq 0)$ (resp. $X(K' = 0)$) is the metric for $K' \neq 0$ (resp. $K' = 0$). X can be $thpu_{s,d}$ the effective throughput, the loss rate of a connection, the delay in F_i etc. Figure 11 (a) shows the gain ratio of $thpu_{s,d}$ and its variation function of the delay limit for $K' = 2$. Here we consider one single delay limit for all connections: $T_{s,d} \equiv T$. For small T where we have a hard constraint of the delay, it is necessary to reduce the flow of packets in the intermediate nodes in a connection while for large T the gain ratio converges to a non zero values. Therefore, we can choose a corresponding configuration of the retransmission limit number function of the delay limit and the hop number of a given connection. This remarkable amelioration observed in the previous figures is mainly due to the fact that:

- Packets coming from each source are been limited on the first hops of each connection. If the network cannot support transporting more packets on a connection (due to congestion and increasing waiting time in the network), it is better to limit the flow of new entering packets in the network. This is a load moderating issue. Furthermore,



a b c

Figure 11. (a) shows the gain ratio when comparing the static and dynamic retransmission scheme function of the delay limit T, here $P = 0.305$ and $K' = 2$. (b) shows the probability of loss gain in the destination for all the connections. (c) shows the gain of the delay in the forwarding queues of each node. (b) and (c) are plotted function of the transmission probability for $T = 100$ and $K' = 2$.



a b c

Figure 12. (a) shows the end to end probability of success gain. (b) shows the gain of the rate of arrival packets to the destination. (c) shows the gain of the effective throughput received. (a), (b) and (c) are plotted function of the transmission probability for $T = 100$, for $K' = 2$ and for all the connections.

it is better also to drop the packets in the intermediate nodes of a connection instead of dropping them in the destination due to the delay constraint. This may lead to gain in bandwidth.

- the dynamic scheme privileges the forwarded packets that come near the destination. It is better to encourage these packets to reach their destination, if not, the network will suffer more wasted bandwidth.

Figures 11 and 12 illustrate clearly the network reaction of the dynamic scheme in comparison to the static case. The gain as it appears from equation (25) can be positive or negative. A negative loss rate gain in figure 11 (b) shows that less packets are dropped in the destination of all the connections (except connection a for high P) in the dynamic scheme while using the delay constraint. A remarkable amelioration in the delay of the source nodes 3, 4 and 8 is deduced from figure 11 (c). This inform on load moderation and then more quantity of packets can enter the network. Unfortunately, the chance that a packet arrives to the destination (in lower layers) becomes lower in the dynamic scheme due to the severe drops on source nodes, see figure 12 (a). However, The balance gain between entering new packets and the end to end probability of success is shown in figure 12 (b) which represents the arrival rate of packets to the destination (in lower layers). Finally, the effective throughput of figure 12 (c) which is a combination of the arrival rate and the accepted rate of packets (in higher layers) shows an amelioration that reaches 50% for moderated transmission probability.

We have presented in this section an analysis of the performance per connection and we have shown the importance of adjusting the retransmissions dynamically for traffics with delay constraints.

VI. Conclusion

Our primary contribution in this paper is the study of the end to end delay in multi-hop ad hoc networks by introducing the impact of different layer. We have presented two methods to find this delay. The first one is based on Markov chains with 3 dimensions from which we have found the distribution of the forwarding queue size and then the average delay. The second one is based on a decomposition method using the residual service time. Packets arrive to the forwarding queues with a geometric distribution with an average already found in our previous works and which depends on many parameters of the network. On other hands service time follows a general distribution.

As an application of our results, we consider the case of real-time traffic which requires delay constraints and then by using the delay analysis we approximate the loss rate of this

traffic. We also propose a cross layer scheme to improve the quality of service of such traffic. It basically consists on adjusting dynamically the limit number of transmissions function of the number of hops. Numerical and simulation results validate the utility and efficiency of our scheme.

REFERENCES

[1] A. Kherani, R. El Azouzi et E. Altman "Stability-Throughput Tradeoff and Routing in Multi-Hop Wireless Ad-Hoc Networks," in *Proc. of Networking Conference*, 15, 19 MAY 2006, Coimbra, Portugal (Best paper award).

[2] R. El Khoury and R. El Azouzi, "Modeling the effect of forwarding in a multi-hop ad hoc networks with weighted fair queueing," in *Proc. of the 3rd International Conference on Mobile Ad-hoc and Sensor Networks*, Springer's LNCS, 12-14 December 2007, Beijing, China (Best paper award).

[3] R. El Khoury and R. El Azouzi, "Dynamic Retransmission Limit Scheme for Routing in Multi-hop Ad hoc Networks," in *Proc. of the Workshop on interdisciplinary systems approach in performance evaluation and design of computer & communication systems* , ACM, 26 October 2007, Nantes, France.

[4] M. Grossglauser and D. Tse, "Mobility Increases the Capacity of Adhoc Wireless Networks", *IEEE/ACM Transactions on Networking*, vol. 10, no. 4, August, 2002, pp. 477-486.

[5] S. R. Kulkarni and P. Viswanath, "A deterministic approach to throughput scaling in wireless networks," *IEEE Trans. on Information Theory*, vol. 50, no. 6, pp. 1041-1049, June 2004.

[6] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 388-404, March, 2000

[7] M. Franceschetti, O. Dousse, D. Tse, and P. Thiran, "Closing the gap in the capacity of random wireless networks," in *Proc. IEEE ISIT*, pp. 439, Chicago, June 2004.

[8] R. Gowaikar, B. Hochwald, and B. Hassibi, "Communication over a Wireless Network with Random Connections," *IEEE Trans. on Inform. Theory*, July 2006.

[9] P. R. Jelenkovic, P. Momcilovic and M. S. Squillante, "Buffer Scalability of Wireless Networks," in *Proc. of IEEE INFOCOM*, Barcelona, April 2006.

[10] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Throughput-Delay Trade-off in Wireless Networks", in *Proc. IEEE INFOCOM*, Hong Kong, 2004.

[11] A. El Gamal, J. Mammen, B. Prabhakar, and D. Shah, "Optimal Throughput-Delay Scaling in Wireless Networks - Part I: The Fluid Model," *IEEE Trans. on Information Theory*, June 2006.

[12] M. J. Neely and E. Modiano, "Capacity and Delay Tradeoffs for Ad-Hoc Mobile Networks," *IEEE Trans. on Inform. Theory*, June 2005.

[13] G. Sharma, R. R. Mazumdar and N. B. Shroff, "Delay and Capacity Trade-offs in Mobile Ad Hoc Networks: A Global Perspective," in *Proc. IEEE INFOCOM* , 23-29 April 2006, Barcelona, Catalunya.

[14] V. Srinivasan, P. Nuggehalli, C. F. Chiasserini, and R. R. Rao, "Cooperation in wireless ad hoc networks," in *Proc. IEEE INFOCOM*, 2003.

[15] A. Urpi, M. A. Bonuccelli, and S. Giordano, "Modelling cooperation in mobile ad hoc networks: a formal description of selfishness," in *Proc. of WiOpt Workshop*, 2003.

[16] R. B Cooper, "Introduction to Queueing Theory," North Holland, 2d ed edition (1981).

[17] Y. Yang, J. C. Hou, and L. C. Kung, "Modeling the effect of transmit power and physical carrier sense in multi-hop Wireless networks," in *Proc. INFOCOM*, Alaska, 2007

APPENDIX

Let us consider a discrete renewal process in which a renewal occurs in a departure of a packet from the forwarding queue F of a given node. Departures occur on times T_1, T_2,..., with $S_i = T_i - T_{i-1}$ $(i = 1, 2, ...; T_0 = 0)$ and S_1, S_2,... is the sequence of service time with independent, identically distributed random variables. Let R_t be the elapsed time from an arbitrary (arrival of a packet to the queue) instant t to the next departure. It is the residual service time at instant t. We need to find $E[R]$ function of the first and second moment of S.

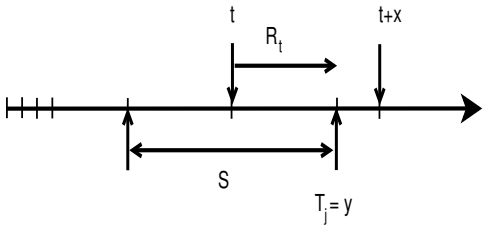


Figure 13. Discrete renewal process with service and residual time

Let the j^{th} departure occurs in $(t, t+x]$ i.e. $t+1 \leq T_j \leq t+x$. Then, the residual service time at time t will not exceed x, so $R_t \leq x$.

In order to have $R_t \leq x$, we must have a departure at time $T_j = y$ between $(t, t+x]$, *and* that $S > \xi$ where $\xi = (t+x)-y$. S cannot be 1 because an arrival cannot see a time service of one slot due to the fact that antennas cannot receive while transmitting. Therefore,

$$\begin{aligned} P\{R \leq x\} &= \lim_{t\to\infty} P\{R_t \leq x\} \\ &= \lim_{t\to\infty} \sum_{j=1}^{\infty} \sum_{y=t+1}^{t+x} P\{S > t+x-y\} P\{T_j = y\} \end{aligned} \tag{26}$$

Remark that $\lim_{t\to\infty} \sum_{j=1}^{\infty} P\{T_j = y\} = \frac{1}{E[X]}$ and S is independent of t. And by making a variable change $\xi = (t+x) - y$, we can write

$$P\{R \leq x\} = \frac{1}{E[S]} \sum_{\xi=0}^{x-1} P\{S > \xi\} \tag{27}$$

Therefore, we can find $P\{R = x\}$ easily,

$$P\{R = x\} = \frac{1}{E[S]} P\{S \geq x\} \tag{28}$$

Hence, the expectation of the residual service time can be calculated,

$$\begin{aligned} E[R = x] &= \sum_{x=1}^{\infty} xP\{R = x\} \\ &= \frac{1}{E[S]} \sum_{x=1}^{\infty} xP\{S \geq x\} \\ &= \frac{1}{2E[S]} (E[S^2] - E[S]) \\ &= \frac{E[S^2]}{2E[S]} + \frac{1}{2} \end{aligned} \tag{29}$$