Bozkurt *EURASIP Journal on Wireless Communications and Networking* (2016) 2016:52
DOI 10.1186/s13638-016-0545-0

EURASIP Journal on
Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Optimal delay analysis for real-time traffics over IEEE 802.11 wireless LANs

Aytül Bozkurt

Abstract

Supporting real-time communications over IEEE 802.11 wireless local area networks (WLANs) is very important yet challenging due to the limited channel capacity, unstable channel conditions, and the low transmission delay requirement of real-time traffic. In this paper, we propose a new analytical model to improve the delay and throughput performance of the real-time applications over WLANs. We model each node as an M/G/1/K queue and the random access process as a two-dimensional Markov chain. Taking into account the rate adaptation feature of real-time applications, we design an iterative searching algorithm to look for the optimal number of retransmission m in the MAC layer with concurrent exploration of the Markov chain and the M/G/1/K queuing models and the variation of the arrival rate. Performance results demonstrate that our analytical model can effectively improve the throughput and average delay under different conditions studied.

Keywords: IEEE 802.11 WLAN, Optimization of retransmission number, Real-time transmission, Quality of service

1 Introduction

Real-time multimedia applications such as voice over IP, video streaming, and video conferencing are getting widely use in business and everyday life [1]. In addition, many network services involve transmissions of real-time applications, for example, web browsing, P2P, Skype, and Youtube. Based on the forecast of Cisco Systems, two-thirds of the world's mobile data traffic will be video by 2016 [2]. With a quick growth of wireless devices and the wide deployment of wireless networks, it is important to enable wireless multimedia applications so people can access the steaming applications anywhere and at any time. However, there is a big challenge to transmit multimedia over wireless networks due to the limited channel capacity, unstable channel conditions, as well as interference and collisions.

In this work, we investigate the problem of better supporting real-time traffic over 802.11-based wireless LAN (WLAN). 802.11 MAC layer assumes a contention-based channel access mechanism, where the distributed control function (DCF) is applied to allocate the channel bandwidth to the users randomly. This random process that DCF employs is also called backoff process.

For higher quality of real-time services, it is desirable to limit the delay of transmissions, while ensuring higher average throughput and lower average delay.

Some recent efforts have been made to analyze transmission delay of the real-time applications over WLAN without providing an actual scheme to restrict the delay. On the other hand, various schemes have been proposed to reduce the transmission delay and collision probability and hence increase the network throughput [3–6] of general packets, including adaptive packetization at the MAC layer, transmission rate optimization, as well as scheduling and service differentiations. However, these studies often ignore the real-time requirement of real-time traffic and the impact of retransmissions on the delay. In 802.11 networks, a sender can retransmit a packet if it fails to arrive at the receiver, and there is a limit on the number of retransmissions allowed for the sender. Before each retransmission, the sender needs to backoff for a random duration within a window which is doubled for each failed transmission. Therefore, a higher number of retransmissions would lead to the increase of the backoff window size, which will on the one hand help reduce the collision probability, thus increasing the

Correspondence: aytulbozkurt@karabuk.edu.tr
Department of Mechatronics Engineering, Karabuk University, Balıklarkayası Mevkii, Karabuk 78050, Turkey

throughput and on the one hand result in a higher transmission delay. In addition, existing delay analyses are limited to the performance evaluation of backoff schemes without considering the real-time service features and arrival patterns.

In this work, we provide quantitative analyses on the performance of real-time transmissions to guide more efficient communications over wireless LAN. Different from the literature work so far, we investigate the optimization problem of retransmission number on transmission performance. Our work has the following features: (i) We concurrently consider the rate adaptation characteristics of real-time applications and the random access mechanism of WLAN, and (ii) We derive the optimal number of maximum retransmission value to ensure a higher transmission throughput without violating the transmission delay bound. We model the random access process as a two-dimensional Markov chain and for the analysis simplicity, exploit M/G/1/K queuing system. The transmission packet is modeled at each state with general packet distribution for service time. From the Markov chain model, we derive the maximum number of retransmission times allowed for a user based on the current traffic conditions and the collision probability, based on which we can calculate the service time of the queue and evaluate the average delay and throughput of the traffic. The two models interact and are solved iteratively with the simultaneous consideration of the rate adaptation of real-time applications to achieve a higher transmission throughput and the limiting of transmission delay to be within a threshold. Finally, numerical results are provided to assess our analytical model with a thorough comparative study of our proposed analytical model with "non-optimal" scheme that has a constant default retransmission number.

The rest of the paper is organized as follows. In Section II, we introduce the related work. In Section III, we present our analytical models and the algorithm to solve the problem iteratively. In Section IV, we provide numerical studies to evaluate the performance of our analytical models and the iterative searching algorithm. We conclude the paper in Section V.

2 Background

In recent years, many studies have been made to analyze the transmission performance over 802.11 WLAN. The authors in [7] analyze the performance of 802.11 taking into account the queue dynamics of a wireless station and the general probability distribution of packet sizes, while in [8], an M/MMG1/1/K queuing model is developed to reduce the complexity level by effectively restoring the independence between the service time and the packet inter-arrival time. The paper [9] evaluates the performance of DCF in binary symmetric channels (BSCs), concurrently considering factors such as the binary exponential backoff mechanism, the incoming traffic loads, and the distribution of incoming packet sizes. The paper [10] analyzes the delay and queue length characteristics following a discrete time G/G/1 queue model and assuming an arbitrary arrival pattern. The model is also extended for analyzing the performance of 802.11e by considering the burst packet transmissions. To support quality of service (QoS) in real-time applications, the concept of critical real-time traffic condition is introduced in [11] to characterize the marginal satisfaction of the real-time requirements. None of the solutions above, however, considers the derivation of performance metrics that can meet the targeted QoS desired by real-time applications. In [12], the authors presented a network analysis model to calculate MAC access delay and throughput by using M/G/1/K queuing model. Bianchi's model [13] is simple and fairly accurate model. Authors proved that 802.11e WLAN can guarantee QoS requirement of the real-time traffic as long as the network is tuned to operate in the non-saturated case and network traffic is not heavy. For saturated channel model, saturation throughput and computation of delay performance are analyzed by modifying channel busy condition and improving Ziouva and Antonakopoulous's model in [14], the more accurate analysis of the DCF are presented.

Supporting real-time quality of service (QoS) in wireless real-time control in [15] the concept of the critical real-time traffic condition, which is a non-saturation condition, is introduced and mathematical models are developed. All these modelling methods for non-saturation condition of an empty queue with random traffic generation, but in [14] authors defined the empty queue for periodic traffic and showed that developed models have been shown to be effective in evaluating the maximum achievable network performance. However, developed model does not combine Markov chain analysis empty queuing interactively in order to evaluate the network performance. For non-saturated traffic, the authors of [16] propose a comprehensive mathematical analysis with taking into account the heterogeneity of the traffic sources (i.e., with the different traffic sources with distinct arrival rates) with M/M//K/1 queues to estimate a set of networking parameters dependant on the traffic source type. However, an important aspect is missed in this analytical model: the distribution of the end-to-end delay. The derivation of the complete distribution of the end-to-end delay is discussed in [17] and proposed a clear and precise performance evaluation method for the total delay of the probability generating function (PGF) by selecting the most accurate

model for the MAC delay is available while improvements are needed for the queuing delay distribution derivation.

To achieve high throughput and QoS provisioning, the idea of resource reservation is a well-known technique in TDMA schemes. The paper [18] applies it in wireless CSMA networks to enhance 802.11e DCF and EDCA, which employ fully random backoff method, to resolve network collision. With the new method, named as semi-random backoff (SRB), analytical study and simulation results show that SRB performance is better than the default 802.11 DCF/EDCA and can achieve even higher performance gain over default 802.11 DCF/EDCA.

The paper [19] introduce spatial reusability-aware single-path routing (SASR) and any path routing (SAAR) protocols for the IEEE 802.11 MAC to improve the end-to-end throughput by carefully considering spatial reusability of the wireless communication media in multi-hop wireless networks. Evaluation results show that the proposed two routing protocols can achieve more significant end-to-end throughput gains under higher data rates.

A good MAC protocol with multi-hop fair access can satisfy the upper bounds on network utilization and lower bounds on delay for multi-hop wireless networks. The authors in [20] propose a cooperative MAC protocol that integrates [21] relay selection, packet piggyback and medium access for the application of cooperative communication techniques and analyze the saturation throughput of the proposed protocol with the simulation results to validate the numerical results. Simulation results showed that throughput of the proposed protocol is better than those of existing CoopMAC and ZrcMAC protocols. They also show that the proposed protocol reduces reservation overhead and improves channel utilization.

To improve the multimedia streaming services, QoS for users over wireless networks is a common goal shared by content providers, network service providers, and smart device manufacturers. A survey on existing literatures on quality of experience (QoE) of the video streaming to explore the efforts to improve the QoE quality metrics and to inspire new research directions in defining better QoE is presented in [22]. The survey identifies four major challenges for QoE-driven mobile streaming video. However, as an open research issue, resource reservation and scheduling schemes are also required to be explored.

A set of studies have been made on the optimization of the initial size of the contention windows to evaluate the impact of the exponential backoff. In [25], the authors propose an analytical model based on closed networks to evaluate the performance of IEEE 802.11. The papers in [26, 27, 29] consider constant and optimal contention window size, respectively, and select the optimal contention window to maximize the throughput for networks of different scales. In [29], the authors incorporate the main QoS features of IEEE 802.11e into the discrete-time Markov chain model (DTMC) and jointly consider the state of MAC layer buffer and MAC differentiation for arbitrary traffic. Optimal configuration of the contention window is also proposed in [28] to improve the throughput performance of WLAN. The work in [28] investigates the video streaming performance based on a Markov chain model and signal transfer function of generalized state transition diagram. These models, however, assume the number of retransmissions is constant or unlimited, which lead to higher delay and lower throughput. The impact of the retransmission number on transmission performance is only considered by very few studies [23, 24]. However, the authors did not relate the performance of the retransmission limit with the actual transmission delay and throughput.

To the best of our knowledge, we propose the first analytical model to jointly optimize the retransmission strategy and the queuing process and consider their interaction to improve the delay and throughput performance, which are critical for real-time applications. In addition, we consider the rate adaptation feature and the delay limit of real-time transmission in searching for the optimal retransmission limit.

3 Analytical model

In this section, we will model the real-time transmission processes over 802.11 Wireless LAN, quantitatively analyze the performance of the system, and derive the parameters that can guide the engineering of the system for the optimal performance.

Without loss of generality, in the local area network, the set of stations are randomly distributed. A transmitted packet may be lost due to the collision or the transmission error and dropped after the maximum retransmission limit is reached. In deriving the optimal throughput, we consider that the network works in the saturation condition and each station always has packets to transmit, i.e., the probability of an empty queue is zero. Under a high traffic load, a packet queue may be full and additional arrival packets will be blocked from entering the queue thus the WLAN system.

In order to analyze the impact of random access of 802.11 on the actual service time of a packet, we model the backoff process as a two-dimensional Markov chain. We model the arrival and service process of each station as an M/G/1/K-PS queue, whose service time is derived from the Markov chain model based on the traffic and

interactions from all the stations in the 802.11 LAN system. Based on the average arrival rate and the service time, the average throughput and delay of a station can be obtained. We can then obtain the average throughput and delay of the overall system. Different from the literature work which generally has a constant limit on the retransmission times for each sender, we attempt to find an efficient retransmission limit in reference to the preset delay bound of a real-time application and taking advantage of the rate adaptation capability of traffic. This control of retransmission limit helps achieve a higher system throughput while not exceeding the target delay limit.

3.1 Markov chain model

The system has N number of stations, and each accesses the channel following the distributed coordinated function (DCF) of 802.11. In DCF, the channel status is monitored during the idle period, and a station can transmit if the channel is sensed free with duration of distributed inter-frame space (DIFS) time. If the channel is sensed busy, a station will backoff and not immediately compete for the channel access again and the backoff duration is set as a random period within a backoff window. The backoff window is initially set to the minimum value $CW_{\min}$, and will be doubled with each additional collision. We denote the number of backoff stages as j, and the size of the contention window at the jth backoff stage is $CW_j = 2j * CW_{\min}$, where $0 < j < m$ with m being the maximum number of retransmissions allowed with our derivation. At time t, if we have j backoff stages and the backoff counter is set as i, we have $s(t) = j$ and $b(t) = i$. The value of i is uniformly distributed in the range $[0, CW_{j-1}]$.

We model the random process $s(t)$, $b(t)$ as a discrete-time two-dimensional Markov chain with $b_{j,i} = Ps(t) = j, b(t) = i$ representing its steady-state probability. Figure 1 depicts the state transition diagram for each station. At the maximum number of backoff stages m, the maximum contention window size $CW_{\max} = CW_m = 2\ m * CW_{\min}$.

The probability of a packet being collided is denoted as p_c, and a collision can happen if there is at least one of the other stations also initiates the transmission at the same time. Whenever there is a collision, the Markov chain moves from the collision stage $j - 1$ to j, and starts from a counter randomly selected from the range $[0, CW_{j-1}]$ and the counter will reduce by one after each time slot if the medium is sensed idle. Generally, when m increases, the collision probability of the Markov chain will reduce.

To obtain the stationary distribution of the Markov chain $b_{j,i}$ from the transition diagram, we will first establish the balance equations. We have

$$b_{j-1,0} p_c = b_{j,0} \tag{1}$$

From which we obtain the following equations

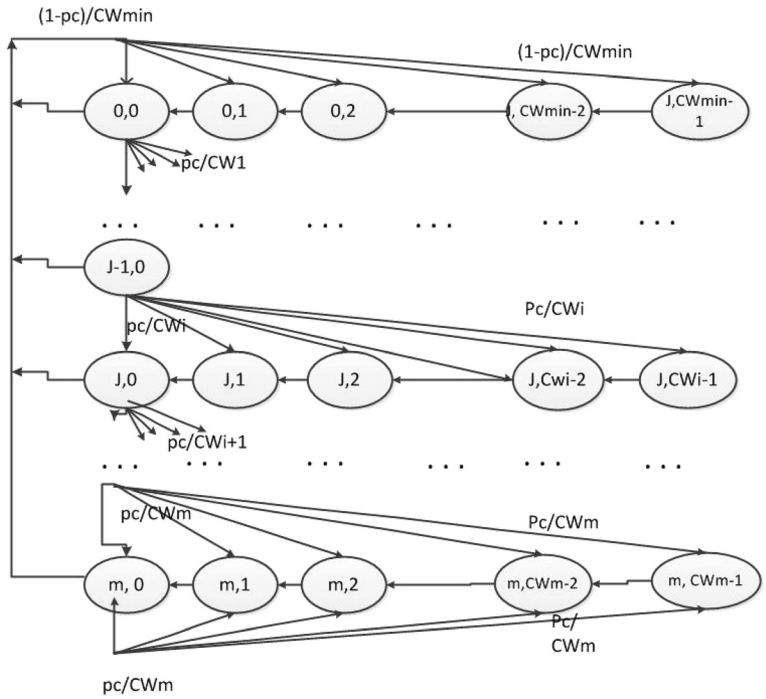

Fig. 1 Markov chain model for the 802.11e backoff process

$$b_{j,0} = p_c^j b_{0,0} \tag{2}$$

$$b_{m-1,0}p_c = (1-p_c)b_{m,0} \rightarrow b_{m,0} = \frac{p_c^m}{1-p_c} b_{0,0} \tag{3}$$

For each i in $[0,CW_{j-1}]$, the steady-state probability $b_{j,i}$ is given by

$$b_{j,i} = \frac{CW_j - i}{CW_j} \begin{cases} (1-p_c)\left(b_{1,0} + b_{0,0}\right), j = 0 \\ p_c, b_{i-1,0} + (1-p_c)b_{i+1,0} 0 < j < m \\ p_c\left(b_{m,0} + b_{m-1,0}\right) j = m \end{cases} \tag{4}$$

$$b_{0,0} = \frac{2(1-2p_c)(1-p_c)}{(1-2p_c)(CW_{\min}+1) + (p_c CW_{\min}(1-(2p_c)^m)} \tag{5}$$

The probability of the successful transmission in a slot time is defined as P_s. If at least one of the stations transmits during a slot time, the channel is busy and the probability of transmission is denoted as P_{tran}.

When there are total N stations contending for the channel, at a given time slot, the collision probability p_c, the success probability P_s, and the transmission probability P_{tran} are given respectively as follows:

$$P_c = 1-(1-\tau)^{N-1} \tag{6}$$

$$P_s = \frac{N\tau(1-\tau)^{N-1}}{P_{\text{tran}}} \tag{7}$$

$$P_{\text{tran}} = 1-(1-\tau)^N \tag{8}$$

The parameter τ is the probability that a station would transmit a packet in a given time slot. At the steady state, the collision probability p_c depends on τ, while τ also depends on the backoff duration thus p_c. τ can be calculated from the Markov chain as the total probability that a counter reaches 0 from any of the m transmission attempts and is represented as follows:

$$\tau = \sum_{j=0}^{m} b_{j,0} \tag{9}$$

Equations (8) and (9) are solved iteratively to determine the unknown parameters τ and p_c until a converging condition is met. After the m-th backoff stage, a packet is discarded in the Markov chain model.

3.2 M/G/1/K-PS queuing analysis

To analyze the average transmission delay and throughput for each station, we apply the M/G/1/K queuing model, where K represents the maximum capacity of the queue at a station. Packets arriving after K packets are already in the queue are dropped. Call arrivals of real-time applications are assumed to follow the Poisson process and the arrival rate to a station n is given by λ(packets/s) and the arrival rate matrix for all N stations is $\lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$.

The steady-state probability of the queue with k packets is $\pi(k)$, where $k = 0, 1, 2, \ldots, K$. Each packet is transmitted using the full channel capacity C.

Let $E[L]$ denote the average packet size (in bytes) in the MAC layer protocol. From the queuing model, the average packet transmission delay can be found by considering the variation in the service rate of the packets due to the random access of the channel. The service capability, C^*, depends on the average channel capacity and the average data payload size $\bar{P}$. It can be calculated as $C^* = C/\bar{P}$, and the traffic load or traffic intensity from the station n is:

$$\rho = \frac{\lambda}{C^*} = \frac{\lambda \bar{P}}{C} \tag{10}$$

In order for the system to be stable, the offered traffic call load should be smaller than one. If the offered traffic load reaches the saturation value 1, it may lead to infinite packet delay. The steady-state probability $\pi(k)$ can be calculated as [30, 31].

$$\pi(k) = \frac{(1-\rho).\rho^k}{(1-\rho^{k+1})}, \ldots\ldots\ldots 0 \le k \le K \tag{11}$$

$$\bar{P} = E[L].P_s.P_{\text{tran}} \tag{12}$$

The packet blocking probability from the WLAN server is given by

$$P_b = P[k = K] = \frac{(1-\rho)\rho^K}{(1-\rho^{K+1})} \tag{13}$$

The average duration it takes to transmit a packet in one transmission stage, $E[T_{\text{slot}}]$, is given as

$$E[T_{\text{slot}}] = (1-P_{\text{tran}})\delta + P_{\text{tran}}P_s T_{\text{suc}} + P_{\text{tran}}(1-P_s)T_c \tag{14}$$

where δ is the duration of the empty slot time, T_{suc} and T_c represent the duration for successful transmission and duration of collision, respectively. T_{slot} is calculated as the summation of the average empty slot time, the average successful transmission period, and the average collision period. The probability that the channel is empty for a slot time is given by $P_I = (1 - p_{\text{tran}})$. When the average number of backoff stages is obtained, the average time a packet spends in the backoff process is divided into average duration $E[T_{\text{slot}}]$ occurred in one backoff stage. The average number of backoff stages of the system is:

$$\alpha = [\delta P_I + T_c P_c + T_{\text{suc}} P_s]/E[T_{\text{slot}}] \quad (15)$$

The mean throughput, S (kbps), is calculated according to the Little's law and is equal to the ratio of the average number of arriving packets with the average packet size that can be served and the average number of packets accepted into the system queue. S is calculated as

$$S = \frac{\lambda(1-P_b)(1-p_c^{(\alpha+1)}}{\sum_{k=0}^{K} \pi(k).k} \quad (16)$$

3.3 Determination of the maximum number of retransmission times (m)

With a limited channel capacity and N stations randomly competing in channel access, the transmission end-to-end delay can be varied in WLAN. In order to meet the quality of service requirements of real-time transmissions, there is a need to reduce the average transmission delay while ensuring a bigger throughput especially under the high network load. The MAC transmission delay is defined as the total duration that a packet is successfully transmitted by the MAC access layer or discarded after reaching the maximum number of retransmission times. The MAC transmission delay is impacted by several factors, including the MAC layer random access delay which is impacted by backoff and collisions, the actual packet transmission delay determined by the packet size and the channel capacity, and the delay caused by channel errors and the resulting retransmissions.

The throughput and the average random backoff delay depend on the number of backoff stages and the probability that a packet is successfully transmitted or discarded after the maximum m-th retransmission. As m increases, the window size will increase, and thus the collision probability will reduce. This will help increase the successful transmission probability and thus the throughput. On the other hand, a large window will also lead to a higher backoff delay, which may lead to the overall increase of the transmission delay and the reduction of throughput. In this work, instead of simply setting m to the fixed default number given by 802.11, we will look for the optimal m considering the tradeoff between reducing the collision probability and increasing the transmission delay. We will look for the optimal m based on the upper-bound of the mean delay time of real-time applications, and also considering the rate adaptation feature of applications. The value m will impact the average MAC service time and throughput under varying traffic load conditions.

The data rate of a real-time service can adapt to the change of bandwidth, which however would also lead to the change of the average delay. The change of the data rate of the real-time service can be modeled by the change of the arrival rate λ of the M/G/1/K queue at a station n. Therefore, the determination of the maximum number of backoff stages would also need to consider the arrival rate. The queuing model is used to evaluate the delay and throughput experienced by a packet at each station, while the service rate of the queue will be impacted by competition and random access from traffic of all stations. The Markov model on backoff stages is used to evaluate the performance of the channel access due to contentions from all stations, and an optimal retransmission limit m will be derived under the condition that the user transmission delay remains to be below the target delay limit of real-time service.

The interactions of the queuing model and the Markov chain model at a station n are depicted in Fig. 2, which shows the following steps:

Step 1: Initialize $m = 1$ and $p_c = pc_{\text{init}}$.
Step 2: With m and pc_{init}, calculate pc_{cur} and Ps_{cur}.
Step 3: Calculate the service time $E[T_s]$ and the average delay T_{delay} by using P_s and the state probabilities $\pi(k)$, $k = 1, 2, \ldots, K$ in M/G/1/K queue system. Find new m and p_c with a recalculation for each arrival rate increase. At different arrival rates, algorithm is run to determine the corresponding outputs (m, T_{delay}, p_c, $E[T_s]$, S).
Step 4: Repeat Steps 2 and 3 with updated m, while the average delay is smaller than the delay upper bound. Otherwise stop the algorithm.

Our proposed searching scheme for determining the value m is given in Algorithm 1. In the proposed algorithm, firstly, $\pi(k)$ is computed by increasing λ in each searching step for $m = 1$ as shown in Steps (1–5). With this steady-state probability $\pi(k)$, the blocking probability of real-time service call users P_b and the mean throughput S are calculated in Steps (6–16).

As discussed earlier, Eqs. 8 and 9 are solved iteratively to find the two unknown parameters, the collision probability p_c, and the transmission probability τ. These two values are obtained in Steps 10–20. The two parameters are used in the determination of the average MAC service time $E[T_s]$ and the average total delay $E[T_{\text{delay}}]$.

>The initial value of m is 1 in Step 5. The value m is searched by increasing its value by 1 in each searching step while the average delay of the call users is smaller than the upper bound of the average delay or when m reaches the largest number of retransmissions at backoff

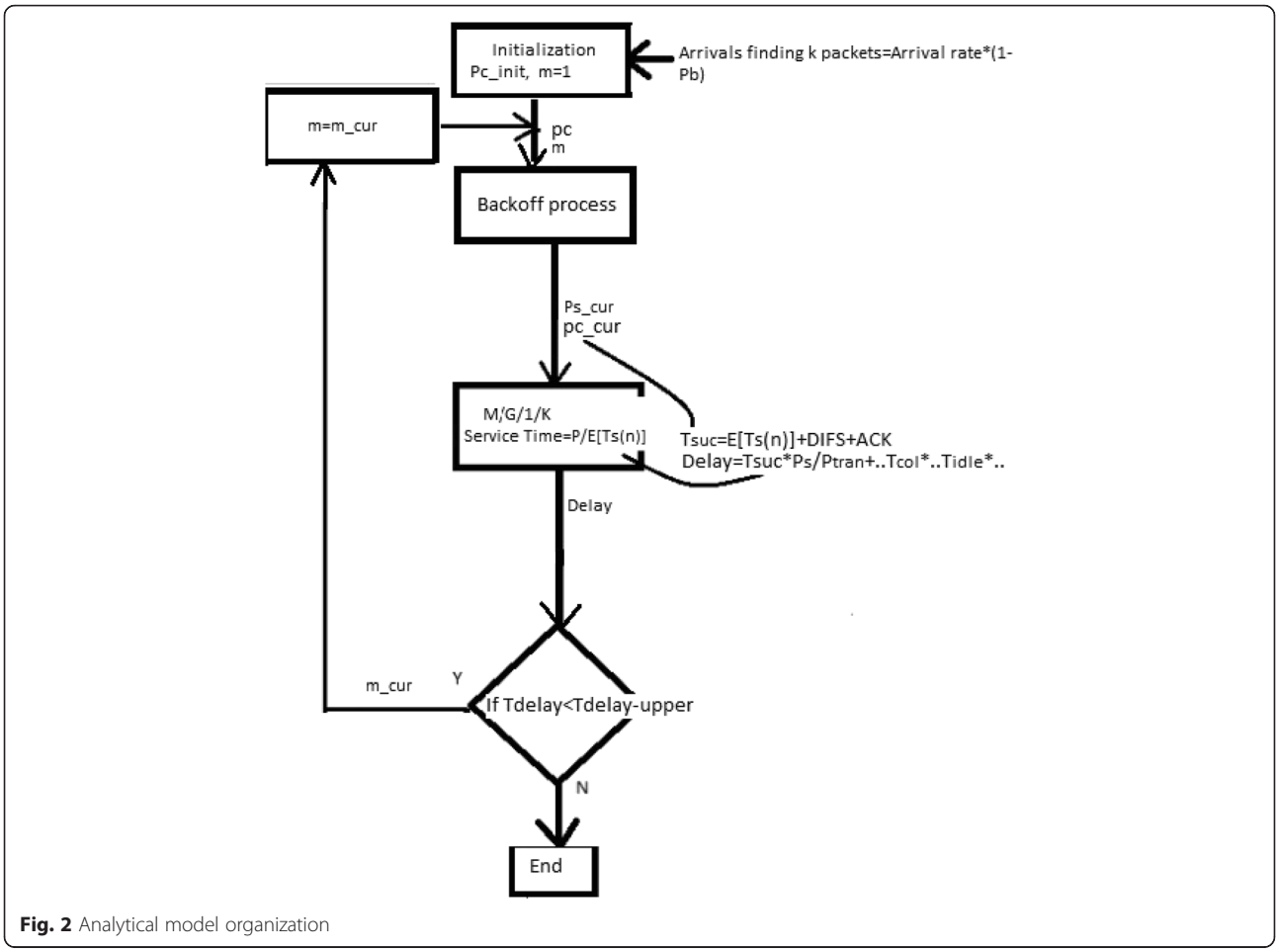

Fig. 2 Analytical model organization

stage that MAC layer allows L_{retry}. The default value of L_{retry} in the IEEE 802.11 is 7. If the optimal number of backoff stages is larger than the default value, the m value will be set as L_{retry}. Steps 22–25 search for the maximum m in each step, and Step 28 outputs the obtained result.

The average MAC service time is defined as the time from a packet reaches the head of M/G/1/K queue at the station to the time it successfully departs from the queue, and is calculated as the time to transmit the average payload size within the retransmission limit m. By obtaining the optimal number of m with our searching algorithm, the collision probability p_c will be reduced upon a high load while not wasting the bandwidth unnecessarily with an excessively large backoff window. Accordingly, the successful transmission probability P_s will increase. With the increase of the average throughput, the average service time can be also reduced. The average service time $E[T_s]$ is calculated as:

$$E[T_s] = \frac{\bar{P}}{S} \tag{17}$$

The duration due to successful transmission is calculated as:

$$T_{\text{suc}} = E[T_s] + \text{SIFS} + \text{ACK} + \text{DIFS} \tag{18}$$

The time cost in the collision T_c is calculated as:

$$T_c = \text{DIFS} + H + \sigma \tag{19}$$

where H and σ represent respectively the transmission time of the packet header and the propagation delay. We define the total delay of a packet as the summation of the duration of idle period, the duration that a channel is sensed busy due to both collisions and successful transmissions, and the duration due to retransmissions.

More specifically, $E[T_{\_\text{delay}}]$ can be expressed as the summation of the average number of idle backoff slots multiplied by the average idle slot duration as a result of backoff at state (ji), the average busy duration as the result of both

successful transmissions and failed transmissions due to collisions, and also retransmission duration, as follows:

Algorithm 1 Determining algorithm of optimal number of the backoff stage for a station n

1: *Initialization:*

2: p_{c_init}, $iterate_{max}$, T_{delay_upper}, m_{init}, *error*

3: *Input:* $\lambda_n=(:\lambda_{n(1)}, \lambda_{n(2)}, \ldots \lambda_{n(L)}$

4: **if** $E(T_{delay}) \geq T_{delay\_upper}$ **then**

5: stop.

6: **else**

7: $m=1$

8: **end if**

9: **while** $E(T_{delay}) < T_{delay\_upper}$ **do**

10: **for** $\lambda_n=(:\lambda_{n(1)}, \lambda_{n(L)}$ **do**

11: *compute* $\pi(n)$

12: *compute* p_c

13: **for** $s=1$ *to iterate max* **do**

14: $\pi_{cur}(s)=\frac{(1-\rho_n).\rho_n^k}{(1-\rho_n^{k+1})}$

15: *compute* $\tau_{cur}(s)$

16: *compute* $pc_{cur}(s)$

17: *compute* $Sn_{cur}(s)$

18: *compute* Pb_s

19: $diff1(s)=\pi_{cur}(s)-\pi(l)$ *and* $diff2(s)=p_{c\_cur}(s)-p_c(l)$

20: **if** $idiff1(s) > error1$ *and* $diff2(s) > error2$ **then**

21: $\pi(n) \leftarrow \pi_{cur}(s)$

$p_c(n) \leftarrow p_{c\_cur}(s)$

22: **end if**

23: **end for**

24: *compute* $T_{delaycur}$

25: **if** $T_{delaycur} < T_{delayupper}$ **then**

26: $m=m+1$

27: $T_{delay}=T_{delaycur}$

28:

29: **end if**

30: **end for**

31: **end while**

$$E[T_{\text{delay}}] = E[I]\delta + E[B]\left[\frac{P_s}{p_{\text{tran}}}T_{\text{suc}} + \frac{p_c}{p_{\text{tran}}}T_c\right] + E[R](T_c + \text{SIFS} + \text{ACK}) \quad (20)$$

With retransmissions, the probability of successful transmission of a packet increases, correspondingly, the channel idle probability decreases. At stage j, the idle probability is given by $p_{\text{tran}}^j.(1-p_{\text{tran}})$. The expected number of idle slots is calculated as,

$$E(I) = \sum_{j=0}^{m} \frac{P_{\text{tran}}^j(1-P_{\text{tran}})}{1-P_{\text{tran}}^{(m+1)}} \sum_{h=0}^{j} \frac{CW_h-1}{2} \quad (21)$$

where, the first term represents the idle probability when the maximum retry limit is reached. The second component on the right side of the equation gives the average number of backoff slots for j retransmissions.

The number of retries/retransmissions to transmit a packet successfully is calculated as,

$$E(R) = \sum_{j=0}^{m} jp_c^j.\frac{(1-p_c)}{1-p_c^{m+1}} \quad (22)$$

The expected number of backoff slots within the limit of the maximum allowable number backoff stages m is calculated as,

$$E(T_{\text{total}}) = \sum_{h=0}^{m}\left(\frac{CW_h-1}{2}\right) \quad (23)$$

Channel busy probability P_{busy} is determined by the ratio of the total busy periods to the total duration, which includes the idle period, the busy period due to successful and collided transmissions and the period due to retrying:

$$P_{\text{busy}} = \frac{\left[\frac{P_s}{P_{\text{tran}}T_{\text{suc}}+\frac{p_c}{1-p_c}T_c}\right]}{\left(1-P_{\text{tran}}E[T_{\text{slot}}] + \left[\frac{Ps}{T_{\text{suc}}}P_{\text{tran}} + \frac{p_c}{1-p_c}T\right]_c + (A)\right.} \quad (24)$$

$$A = T_c + \text{SIFS} + \text{ACK}$$

4 Numerical results

In this section, we evaluate the performance of real-time applications in an 802.11 WLAN based on our analytical models, and present numerical results under various network conditions. The default parameters of the WLAN are set following the 802.11 protocol: ACK = 20 μS, SIFS = 10 μS, DIFS = 50 μS, and the initial backoff window is set as CW = 32. The default number of stations in the network is 40, unless otherwise mentioned. The wireless network channel rate is set as 2 Mbps. The packet size is fixed as

1000 byte. We compare the performance of our scheme, called "optimal" in the figures, with a non-optimal scheme that has a fixed number of retransmissions. Also, for a fair comparison, our proposed model is compared with an existing DCF analytical model [12] through numerical results and in comparison. Although the authors [12] consider many aspects of the MAC protocol such as the AIFS, countdown procedure, the backoff process, and collision, it lacks computation of optimal retry limit number, which is a very important condition affecting the delay performance of backoff mechanism. Related model in [12] is referred to as a Xiang, Yu-Ming, and Jun's model (XYJ's) model in the rest of this paper. We set the m value of the non-optimal scheme to be 3 or 4 which is in the middle range of 0 and the maximum transmission limit 7 set by 802.11, using retry limit m for (XYJ) model 4 and considering that the service time is composed of RTS + CTS + l_d + ACK + 3SIFS + AIFS, where l_d is packet length that has 1000 bytes. Variable number of wireless traffic users (with an average bitrate of 22.4 kb/s) for (XYJ) model, which is going from 1 to 80, is chosen. All the users experience the same Poisson arrival rate λ(packets/s) and the m value in our scheme varies under different conditions to approach the optimal system performance. Following, we present our performance studies under different scenarios.

4.1 Impact of initial window size on real-time service delay

We first evaluate the impact of initial window size on the average delay of the real-time application with the number of users in the network $N = 20$, 40, and 80, respectively. The upper bound of the average delay is set to $T_{\text{delayupper}} = 1$ s.

Figure 3 shows the average delay when the initial window size increases, with the number of stations $N = 20$.

For the default size of initial window, $CW = 32$, the average delay is 0.3698 s. As the initial window size varies from 4 to 78, the average delay of the non-optimal scheme and XYJ model increases linearly with the initial window size, while the average delay of the optimal scheme we propose remains in the range of 0.40–0.42 s and is much lower than the non-optimal scheme. Also, our model can obtain a better delay performance than XYJ model. In our searching algorithm, with each setup of initial window size, m is optimized to maintain the delay QoS requirement. Hence, when $CW = 52$, $CW = 72$, and $CW = 78$, we can observe the reduction of the average delay due to changes in the optimal value of m. In the non-optimal scheme and XYJ scheme, there is only one m value for the different initial window sizes.

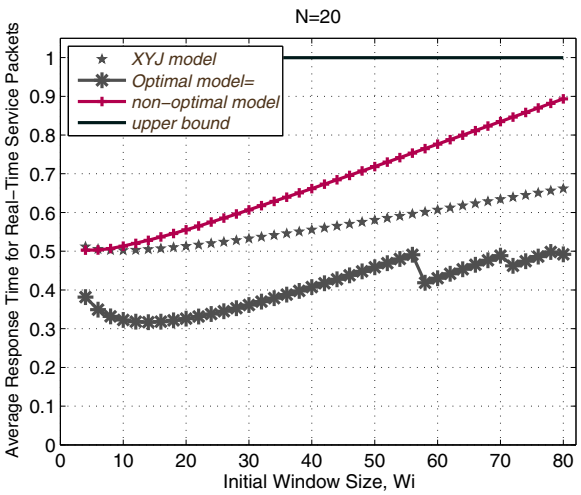

Fig. 3 Average delay versus initial window contention size (*CW*_min) for $N = 20$

This indicates that the window size has a big impact on the average delay and our proposed scheme of adapting m and application rate can effectively maintain a stable and low transmission delay.

Figures 4 and 5 also show the impact of the initial window size for different number of users N, such as $N = 40$ and $N = 80$. For a larger N, the initial window size setting has a larger impact, and the average delay increases much faster with the increase of the initial window sizes. At the default size of initial window, $CW_{\min} = 32$, the average delay for the optimal scheme is 0.3741 and 0.3784 s for $N = 40$ and 80, respectively. The average delay value of the non-optimal scheme and XYJ scheme is much more than that of the optimal scheme and the difference increases as N becomes larger. When $N = 80$, at the default window size of $CW_{\min} = 32$, the average delay for the non-optimal scheme almost doubles that of the optimal one.

4.2 Impact of delay upper bound

In this study, the required delay is set to be below 0.5 s. In Fig. 6, as the number of real-time users N varies in the range 0 and 80, the average delay of our proposed algorithm is 0.2 to 0.5 s, which is bounded below the delay limit. In contrast, the average delay of the non-optimal scheme increases almost exponentially with the number of stations when a fixed value m is used for different number of stations. Our proposed model also outperforms XYJ model. This is because m is obtained by iteratively for different number of traffic users, the number of retransmission is optimized and the users do not suffer the longer delays due to increasing m. In contrast, the optimal m values at $N = 10$, $N = 18$, $N = 30$, $N = 60$,

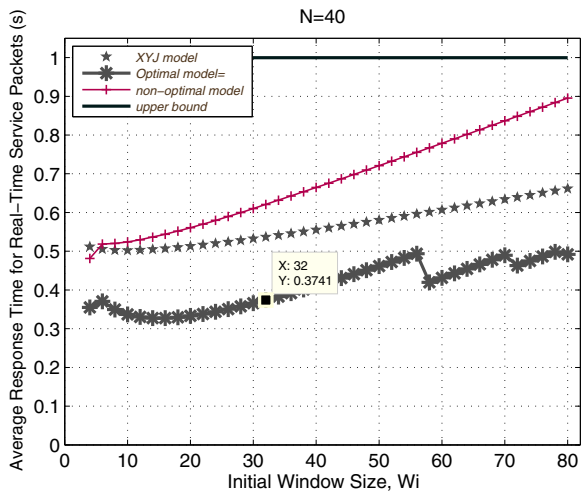

Fig. 4 Average delay versus initial window contention size (CW_{min}) for $N = 40$

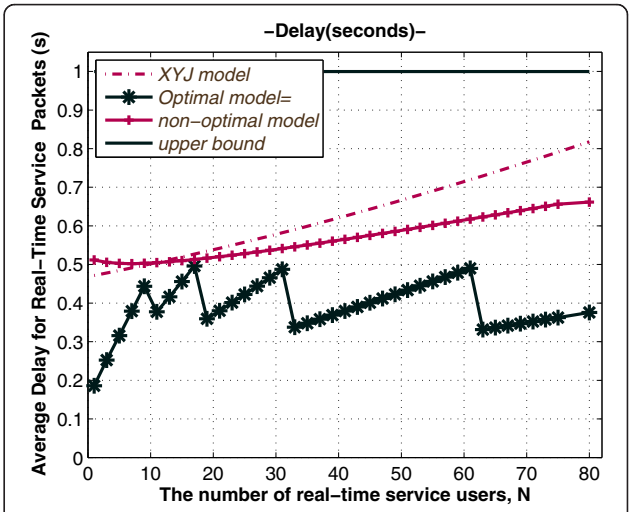

Fig. 6 Average delay versus the number of real-time call users, N for $CW_{min} = 32$

and $N = 78$ are obtained as $m = 3$, $m = 4$, $m = 5$, $m = 6$, and $m = 7$, respectively.

In each change of m to a higher value, the collision probability p_c will first reduce, and consequently, the delay will reduce at each changing point. With a given m, the delay starts to increase with the number of users until m is increased to the next value. We can observe the fluctuation of delay at each m changing point from the figure.

Figure 7 shows the throughput versus the number of stations. As the number of stations increases, the average throughput for each station decreases for both schemes. The optimal scheme has much higher average throughput compared to XYJ scheme and the non-optimal scheme at a higher load, i.e., when the number of stations exceeds the default number of stations $N = 40$.

When the number of stations N increases from 0 to 80, the average throughput of the optimal scheme decreases from 1976 to 901.2 Kbps while the throughput of the non-optimal scheme decreases from 1924 to 439.7 Kpbs.

At a lower number of stations, the optimal m is smaller than the value of the non-optimal scheme, which leads to a higher collision probability thus a slightly lower throughput.

4.3 Impact of the arrival rate

In Fig. 8a, b, we study the impact of the arrival rate at the system capacity $C = 2000$ Kbps. We vary the arrival rate by

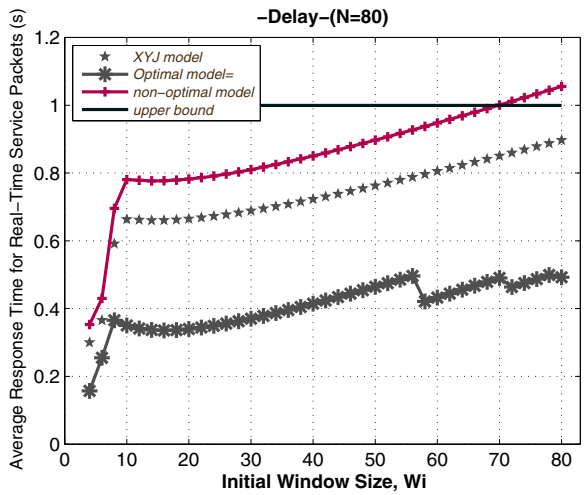

Fig. 5 Average delay versus initial window contention size (CW_{min}) for $N = 80$

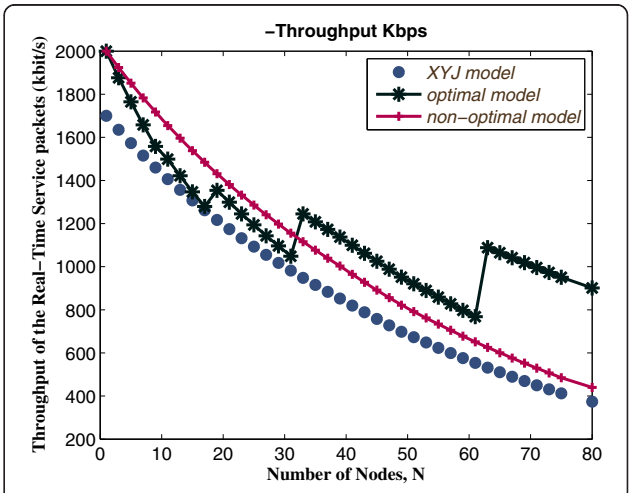

Fig. 7 Throughput for packets (kbits/s) versus number of nodes, N

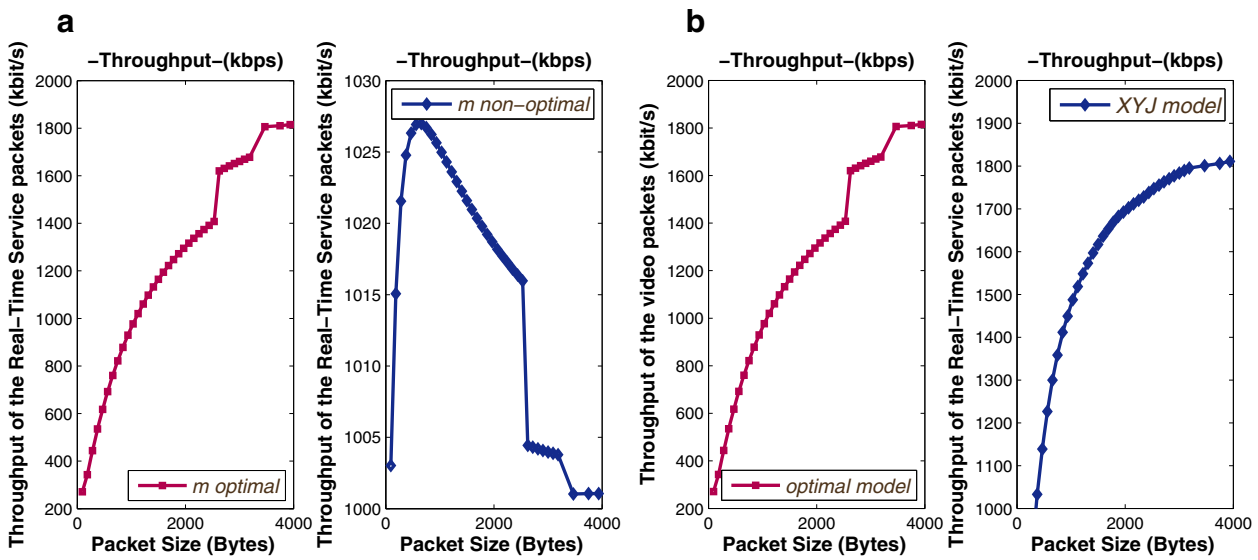

Fig. 8 Throughput for packets (kbits/s) versus packet size. **a** Throughput for non-optimal scheme. **b** Throughput for XYJ scheme

changing the packet size. With the fixed maximum retransmission times, Fig. 8a shows that the throughput of the non-optimal scheme starts to decrease when the packet size is beyond 1000 bytes, i.e., when the system load reaches certain level. Also, as the packet size of XYJ model increases, its throughput performance decreases. In contrast, the optimal scheme proposed adapts the maximum retransmission times m when the traffic load changes with the variation of packet sizes, and the system throughput is seen to increase further.

We observe the smaller rate for the optimal scheme when the network load is low as a result of changing retransmission times m to ensure the delay to be below the threshold required by the user. However, when the network load is high, the optimal scheme can achieve a much higher throughput than the non-optimal scheme.

4.4 Variation of link capacity

Compared to the XYJ model and the non-optimal scheme, Fig. 9 indicates that the proposed optimal scheme is much more effective in improving the throughput by taking advantage of the higher system capacity.

At a higher system capacity, the throughput of the optimal scheme is much higher as it has a lower collision probability with m set to a larger value, while the XYJ model and non-optimal scheme keeps the m value to be the same. On the other hand, when the channel capacity is low, the throughput of the optimal scheme is slightly lower as it has a higher collision probability with its use of a smaller m value thus smaller backoff window sizes.

4.5 The impact of the number of users on the collision probability

In Fig. 10, as expected, the increase of the number of users leads to a higher collision probability for XYJ model and the non-optimal scheme.

Whereas in the optimal scheme, with the change of the number of stations, the m value adapts accordingly to ensure the average delay to not exceed the target value. The collision probability is also maintained within the range 0.3–0.6. The collision probability performance is also demonstrated for different values of initial contention window size, $CW_{\min} = 16$, $CW_{\min} = 24$, and $CW_{\min} = 32$.

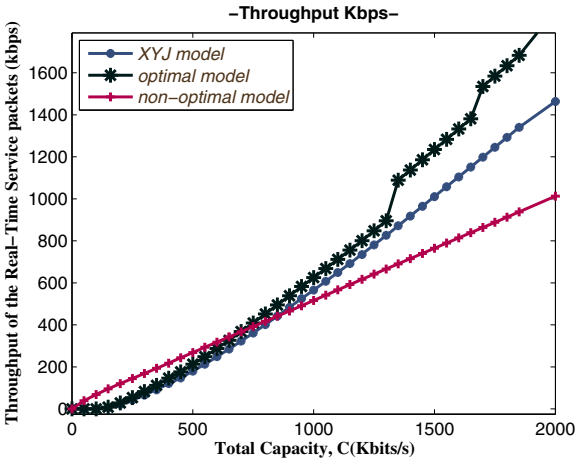

Fig. 9 Throughput for packets (kbits/s) versus the total link capacity, C (Kbits/s)

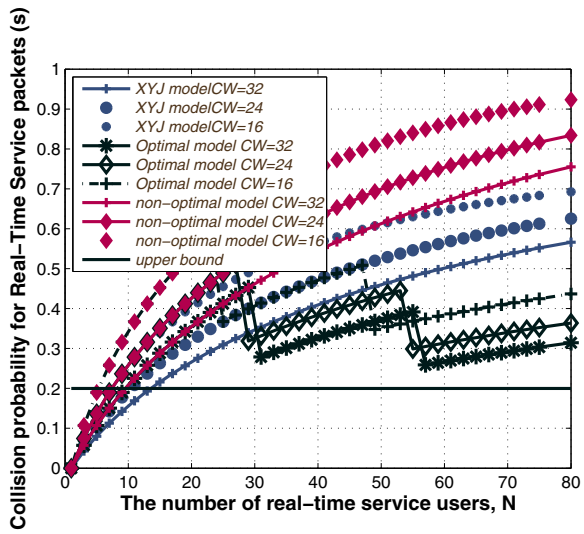

Fig. 10 The collision probability versus the number of users *N*

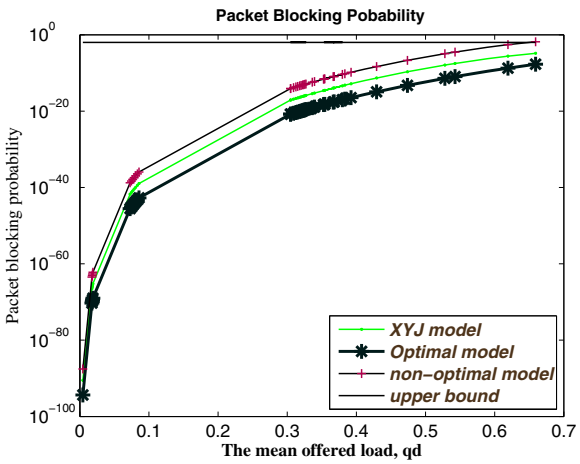

Fig. 11 Packet blocking probability for packets (kbits/s) versus mean offered load, ρ_n

In the figure, as the initial contention window size increases, the collision probability decreases.

4.6 The impact of the offered traffic load on the blocking probability

When the offered traffic load exceeds the throughput that the system can achieve, the incoming packet service requests are blocked. As the optimal scheme can ensure a higher system throughput compared to the XYJ model and the non-optimal scheme, it allows a lower packet blocking probability accordingly as shown in Fig. 11. In this study, the upper bound of the blocking probability is set as $P_b = 0.01$.

4.7 The impact of packet size on the channel busy probability

We consider a channel to be busy when there are successful transmissions or collided transmissions. In Fig. 12, as the packet size increases, we can observe that the channel busy probability also increases which indicates that the wireless channel utilization is at a higher level. For example, the busy probability is 0.90 when the packet size is 1000 bytes, 0.905 when the packet size is 1200 bytes.

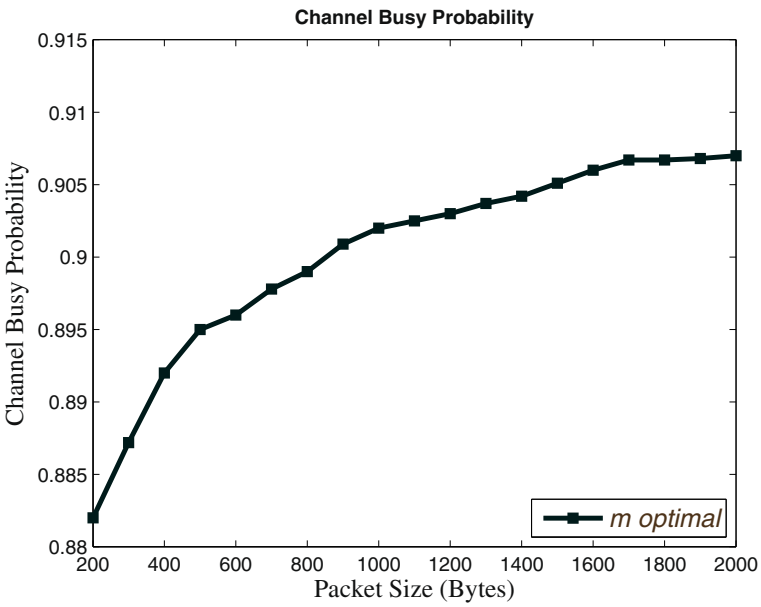

Fig. 12 Channel busy probability

For 2000 bytes packets, the maximum performance is gained at the channel busy probability of 91 %.

5 Conclusions

In this paper, we provide an analytical model on the performance of real-time applications transmission over WLAN. Motivated by the rate adaptation feature of real-time applications, our analytical model novelly evaluates the random access performance of real-time services based on two-dimensional Markov chain model by taking into account the impact of the maximum optimal retransmission number on the service time of the packet transmission. We analyze the transmission throughput and delay of each station based on the M/G/1/K queuing model, and the service time is calculated based on the collision probability derived from the Markov chain model. We propose a searching algorithm that can iteratively solve the two models to look for the optimal retransmission number limit with the adaptation of arrival rate of real-time services, while ensuring the average transmission delay to be below the target delay limit.

Based on our analytical models, we evaluate the delay and throughput performance of the real-time applications under various traffic load conditions and system parameter changes. Under the same channel settings, the numerical results demonstrate that our analytical models can achieve much better performance results than the XYJ model and the non-optimal scheme with a fixed retransmission number.

Competing interests

The authors declare that they have no competing interests.

Received: 8 December 2015 Accepted: 4 February 2016
Published online: 17 February 2016

References

1. Allot, Allot mobile trends: global mobile broadband traffic report. (2011)
2. Cisco Visual Networking Index: global mobile data traffic forecast update, 2011-2016
3. B Fan, L Shen, T Song, The design and implementation of a wireless real-time video transmission system over WLAN, in *Proc. International Conference on Information Science and Engineering (ICISE)*, 2009, pp. 684–687
4. H Zheng, G Chen, L Yu, Video transmission over IEEE 802.11n WLAN with adaptive aggregation scheme, in *Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2010, pp. 1–5
5. K Medepalli, FA Tobagi, Towards performance modeling of IEEE 802.11 based wireless networks: a unified framework and its applications, in *Proceedings of IEEE INFOCOM*, 2006, pp. 1–12
6. NS Shankar, MV Schaar, Performance analysis of video transmission over IEEE 802.11a/e WLANs. IEEE J. Trans. Vehic. Tech. **56**(4), 2346–2362 (2007)
7. CG Park, HS Jung, DH Han, Queueing analysis of IEEE 802.11 MAC protocol in Wireless LAN, in *Proc. IEEE International Conference on Mobile Communications and Learning Technologies, ICN/ICONS/MCL 2006*, 2006, pp. 139–145
8. M Ozdemir, AB McDonald, "A queuing theoretic model for IEEE 802.11 DCF using RTS/CTS", in Proc. 13th IEEE Workshop on Local and Metropolitan Area Networks, 2004. LANMAN 2004, pp. 33-38.
9. Y Zheng, K Lu, D Wu, Y Fang, Performance analysis of IEEE 802.11 DCF in binary symmetric channels, in Proc. IEEE Global Telecommunications Conference, 2005. GLOBECOM '05, Dec. 2005, pp. 3144-3148
10. O Tickoo, B Sikdar, A queueing model for finite load IEEE 802.11 random access MAC, in *Proc. 2004 IEEE International Conference on Communications*, 2004, pp. 175–179
11. T Guosong, T Yu-Chu, Modelling and performance evaluation of the IEEE 802.11 DCF for real-time control. Computer Networks **56**(1), 435–447 (2012)
12. B Xiang, MY-Ming, X Jun, "Performance investigation of the M/G/K/1-based IEEE 802.11E EDCA under non-saturation conditions based on the M/G/1/K Model", in Proc. IEEE International Conference on Communications, Circuits and Systems (Harbin, 23-25 May 2007), pp. 298-304.
13. G Bianchi, Performance analysis of the IEEE 802.11 distributed function. IEEE J. Selected Areas Commun. **18**(3), 535-547(2000).
14. C-E Weng, H-C Chen, Performance evaluation of IEEE 802.11 DCF using Markov chain model for wireless LANs. Comp. Standards Interfaces. **44**, 144-149(2016).
15. G. Tian, Y-C Tia, Modelling and performance evaluation of the IEEE 802.11 DCF for-real time control, Computer Networks **56**(1), 434-447 (2012).
16. K Kosek-Szott, A comprehensive analysis of IEEE 802.11 DCF heterogeneous traffic sources. Ad Hoc Netw. **16**, 165-181 (2014).
17. Q Wang, KJ-Runser, J-L Scharbarg, C Fraboul, Y Sun, J Li, Z Li, A through analysis of the performance of delay distribution models for IEEE 802.11 DCF. Ad Hoc Netw. **24**, 21-33 (2015).
18. Y He, R Yuan, J Sun, W Gong, Semi-random backoff: towards resource reservation for channel access in Wireless LANs. IEEE/ACM Trans. Netw. **21**(1), 204-217 (2013).
19. T Meng, F Wu, Z Yang, G Chen, AV Vasilakos, Spatial reusability-aware routing in multi-hop wireless networks. IEEE Transaction on Computers. **65**(1), 244-255 (2016).
20. K Liu, X Chang, F Liu, X Wang, AV Vasilakos, A cooperative MAC protocol with rapid relay selection for wireless ad hoc networks. Comp. Netw. **91**, 262-282 (2015).
21. L Fei, Q Gao, J Zhang, Q Xu, Relay selection with outdated channel state information in cooperative communication systems. IET Commun. **7**, 1557–1565 (2013)
22. G-M Su, X Su, Y Bai, M Wang, AV Vasilakos, H Wang, QoE in video streaming over wireless networks: perspective and research challenges. Springer Wireless Netw. 1-23 (2015)
23. M Schaar, DS Turaga, Cross-layer packetization and retransmission strategies for delay-sensitive wireless multimedia transmission. IEEE J. Trans.Mult. **9**(1), 185–197 (2006)
24. Y Zhaos, SC Ahalt, J Dong, Content-based retransmission for video streaming system with error concealment. Visual Inform. Process. XIII **5438**, 63–70 (2004)
25. E Karamad, F Ashtiani, Performance analysis of IEEE 802.11 DCF and 802.11e EDCA based on queueing networks. Commun. IET. **3**(5), 871–881 (2009)
26. C-E Weng, C-Y Chen, C-H Chen, C-H Chen, in Proc. International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA). Optimal Performance Study of IEEE 802.11 DCF with Contention Window (Barcelona, Spain, 2011), pp. 505-508.
27. I Inan, F Keceli, E Ayanaoglu, Analysis of the 802.11e enhanced distributed channel access function. IEEE J. Trans.Commun. **57**(6), 1753–1764 (2009)
28. XW Yao, WL Yang, SH Yang, Video streaming transmission: performance modelling over wireless local area networks under saturation condition. Commun. IET. **6**(1), 13–21 (2012)
29. Z-N Kong, DHK Tsang, B Bensaou, D Gao, Performance analysis of IEEE 802.11e contention-based channel access. IEEE J. Sel. Areas Commun. **22**(10), 2095–2106 (2004)
30. J Cao, M Andersson, C Nyberg, M Kihl, Web server performance modeling using an M/G/1/K*PS Queue, in Proc. IEEE 10th International Conference on Telecommunications (Zagrep, Croatia, 2009), pp. 1501-1506.
31. RB Cooper, *Introduction to Queueing Theory*, 2nd edn., 1981