SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

June 2014, Vol. 57 062312:1–062312:13
doi: 10.1007/s11432-014-5103-y

End-to-end delay estimation for multi-hop wireless networks with random access policy

JIAO WanGuo, SHENG Min*, SHI Yan & LI YuZhou

The State Key Lab of Integrated Services Networks, Xidian University, Xi'an 710071*, China*

Received May 6, 2013; accepted June 13, 2013

Abstract End-to-end delay analysis is an important element of network performance analysis in multi-hop wireless networks. In this paper, we propose an analytical model for estimating the end-to-end delay performance of wireless networks employing a random access policy for managing node' transmissions on shared channels with time-varying capacity. To obtain the closed form expression, a new concept of residual effective capacity is presented using the definitions of effective bandwidth theory and effective capacity theory. This allows us to calculate the cumulative distribution function of the queuing delay. Based on this concept, we derive a formula to calculate the average end-to-end delay for multi-hop wireless networks, with the result including the effect of a random access protocol, which has not previously been considered. Finally, we validate our analysis through simulations and provide an example application for our results.

Keywords end-to-end delay, multi-hop wireless network, effective bandwidth, effective capacity, QoS guarantee

Citation Jiao W G, Sheng M, Shi Y, et al. End-to-end delay estimation for multi-hop wireless networks with random access policy. Sci China Inf Sci, 2014, 57: 062312(13), doi: 10.1007/s11432-014-5103-y

1 Introduction

With the rapid expansion of communication networks, the percentage of real-time traffic has grown. How to provide efficient quality of service (QoS) guarantees for real-time traffic is important in wireless network development [1–3]. As the end-to-end delay is a crucial QoS metric for real-time traffic and traffic with a delay requirement is increasing dramatically in wireless networks, the end-to-end delay is often used as the routing metric for designing a QoS routing protocol [4]. Therefore, analyzing the end-to-end delay performance is very important in research on multi-hop wireless networks. In this paper, we focus on end-to-end delay analysis for multi-hop wireless networks with a random access policy and time-varying channels.

End-to-end delay is the sum of the delays encountered by traffic flows or packets at each node along the path, determined by the length of the path and the sojourn time at each node on the path. The length of the path of a given communication pair mainly depends on the routing protocol adopted by the network. Given all route paths for a communication pair, the remaining work in estimating the end-to-end delay is to derive a formula to calculate the sojourn time at each node on the path. As a node in the network can be modeled by a queuing model, the sojourn time is obviously only affected by the traffic

*Corresponding author (email: msheng@mail.xidian.edu.cn)

arrival and node service processes. Because the traffic arrival process is a general stochastic process and the node service process, which is affected by the access protocol, transmission policy and time-varying channel [5], is difficult to be modeled, analysis of the end-to-end delay in a multi-hop wireless network is very challenging. Much effort has been spent on analyzing the end-to-end delay performance of wireless networks [6–13]. However, owning to the complexity of the problem, existing studies focus only on some of the factors.

A scaling law for the end-to-end delay in wireless random networks was derived in [6]. In this study, the authors assumed that there is no interference or collisions when two neighboring nodes transmit simultaneously and that each transmission channel is constant and error free, both of which are impractical. Various researchers [7–9] investigated the influence of the media access control (MAC) protocol on end-to-end delay performance. The authors in [7] analyzed the end-to-end delay in a linear network employing slotted-ALOHA or time division multiple access (TDMA). The optimal transmission policy at the MAC layer was analyzed in [8], while the end-to-end delay of a homogeneous network with a uniform distributed traffic load and node density was calculated in [9]. To ascertain which packet scheduling law is suitable for real-time traffic, various studies examined the effect of the service law on end-to-end delay performance [10,11]. In this paper, we assume all nodes use a first-in-first-out (FIFO) policy to service backlogged packets.

All the works mentioned above focused only on the effect of higher layers, and failed to consider the features of the wireless channel, which is time varying and unreliable. To remedy this oversight, the authors in [12] studied the influence of a retransmission policy on end-to-end delay performance under a constant data rate and determined error rate. The statistical upper bound of a network with the general service process in the area of information theory was derived in [13]. To capture the effect of the time-varying property of a wireless channel, the authors in [14] considered a channel with time-varying capacity by extending the work in [15] to a multi-hop scenario. In [14,15], the packet arrival interval was assumed to be constant. In [16], the analysis was improved by including traffic burstiness in the analytical model with the help of the concept of residual effective capacity. To further improve the work in [16], we include the effect of the MAC protocol in the end-to-end delay analysis.

In this paper, we study the influence of the traffic arrival process and time-varying channel on end-to-end delay performance. Therefore, the traffic arrival process is assumed to be a general stochastic process while the channel capacity is time varying. Owning to the limitation of the queuing theory, we use network calculus, which is widely used to analyze queue performance in communication networks, as the analysis tool [17]. For a given traffic arrival process with a QoS requirement and channel with constant data rate, the queuing delay distribution can be obtained by using effective bandwidth (EB) function [18,19], which is an important method for modeling the traffic arrival process in network calculus. Correspondingly, the queuing delay distribution of a system with constant arrival rate and time-varying service rate can be calculated by employing the effective capacity (EC) function [15] which is an extension of EB. However, neither EB nor EC theory can solve the problem with a general arrival process and time-varying channel. In this paper, we propose the concept of residual effective capacity to release the constraint on traffic arrival rate or service rate. Although part of this work has been published as a conference paper [16], this paper improves the accuracy of the results by adding the effect of the MAC protocol and also presents further simulations.

In the remainder of the paper, we first introduce some mathematical background and our network model in Section 2. End-to-end delay analysis is presented in Section 3, followed by the simulation results and applications in Section 4. We conclude our paper in Section 5.

2 Preliminaries and network model

In this paper, we use network calculus theory as analysis tool. Before presenting our analytical model, in this section we firstly give a brief introduction to the effective bandwidth theory and effective capacity theory, which are important parts of network calculus theory and will be used in our context. And then describe the network model which will be investigated in this paper.

2.1 Preliminaries

Effective bandwidth is a widely used stochastic traffic model for stochastic QoS analysis in computer networks, and is used to estimate the minimum bandwidth required to guarantee a certain probabilistic QoS requirement of multiple flows under multiplexing [19]. The authors in [18] calculated the EB function of a general Markovian traffic model for both a single-source and multiplex system. Later, Chang et al. introduced the physical meaning of the EB function and applied the EB function to QoS guarantees in asynchronous transfer mode (ATM) networks [19]. According to [19], the EB function gives the minimum amount of bandwidth required to satisfy that demanded by a time-varying source to maintain a certain queue quality. Given the arrival process model of the traffic flow, the EB function can be calculated by directly applying the results presented in [18] or using the definition in [19]. For example, if the arrival process is a two-state Markov Modulated Poisson Process (MMPP2) with parameters $(r_1, r_2, \lambda_1, \lambda_2)$, according to the results in [18], the effective bandwidth function $\alpha^{(b)}(u)$ can be directly calculated as

$$\begin{aligned}\alpha^{(b)}(u) =&\frac{1}{2u}\Big(-(r_1+r_2)+(\lambda_1+\lambda_2)u\\&+\sqrt{(r_1+r_1-(\lambda_1+\lambda_2)u)^2-4(\lambda_1\lambda_2u^2-r_1\lambda_2u-r_2\lambda_1u)}\Big),\ u\geqslant 0,\end{aligned} \tag{1}$$

where u is the QoS exponent indicating the service level, and r_j and λ_j are the transition and arrival rates of state j, $j=1,2$, respectively.

When the EB function of the traffic arrival process is given and the service rate is a constant c, the queuing performance of this system can be derived according to [19] as

$$\Pr(q(\infty)\geqslant x)\approx e^{-\theta_b x}, \tag{2}$$

where $q(\infty)$ is the steady state queue length, and θ_b is the unique solution of $\alpha^{(b)}(u)=c$.

In [15], Wu et al. extended EB theory to model wireless channel at the data link layer and proposed the EC theory. The EC function characterizes the maximum arrival rate that the wireless channel can handle when satisfying the QoS requirement. EC has been extended to model the capacity of secondary users in cognitive networks [20]. Let $r(t)$ be the instantaneous channel capacity at time t. The service provided by the channel during $[0,t]$ is $S(t)$ and is calculated as $S(t)=\int_0^t r(\tau)\mathrm{d}\tau$, which depends on the instantaneous channel capacity only and is independent of the traffic arrival process. The EC function of a wireless channel with instantaneous capacity $r(t)$ of connection with QoS exponent u is shown in [15]:

$$\alpha^{(c)}(u)=\frac{-1}{u}\lim_{t\to\infty}\frac{1}{t}\log E\left[e^{-u\int_0^t r(\tau)\mathrm{d}\tau}\right],\ \forall u\geqslant 0. \tag{3}$$

If the EC function of a wireless channel is known, the queueing delay of a system with constant traffic arrival rate μ can be calculated as [15]:

$$\Pr(d(\infty)\geqslant x)\approx \gamma_c e^{-\theta_c x}, \tag{4}$$

where $d(\infty)$ is the steady state queueing delay, and γ_c is the probability that the buffer is nonempty at a randomly chosen time, and θ_c is the unique solution of $\alpha^{(c)}(u)=\mu$.

However, while (2) requires a constant service rate, Eq. (4) requires a constant traffic arrival rate. Therefore, neither of these is suitable for the general case with a general traffic model and a time-varying service process. To solve this problem, we propose the concept of residual effective capacity, which is used in investigating the queuing delay performance of the general case and is discussed in detail in the next section.

2.2 Network model

There are N randomly distributed nodes in a multi-hop wireless network. Each node in the network is immobile and the paths between communication nodes are predefined and time-invariant. In a multi-hop wireless network, each node has to handle two types of traffic: newly generated traffic and relay traffic

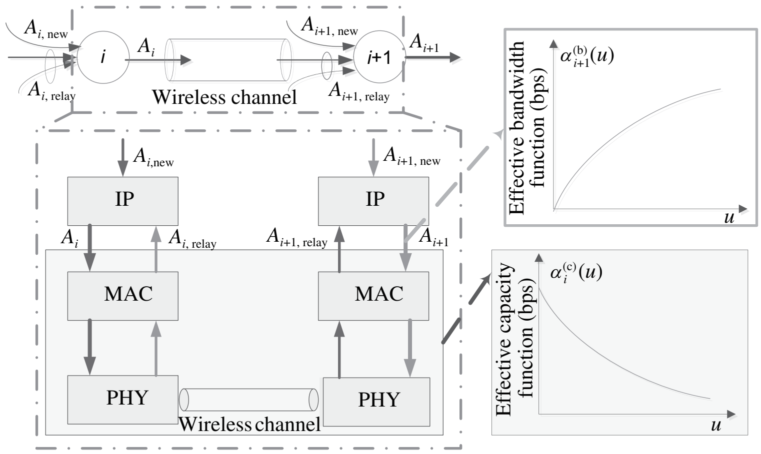

Figure 1 The node model with EB and EC

as shown in Figure 1. $A_{i,\text{new}}(t)$ and $A_{i,\text{relay}}(t)$ represent the numbers of newly arrived packets and relay packets, respectively, at node i, $i$in$\{1, 2, \ldots, N\}$, during [0, t]. The total number of packets arriving at node i during [0, t] is $A_i(t) = A_{i,\text{new}}(t) + A_{i,\text{relay}}(t)$. We assume $A_{i,\text{new}}(t)$ is a stationary random process. Relay traffic from neighboring nodes depends on the routing protocol, network topology, and traffic property, amongst others. Owning to the sufficiently randomized events in the network, it is plausible to declare that the arrival process of relay traffic is a stationary random process [21,22]. Similar to the assumptions adopted in previous works [5-19], we assume that the traffic arrival process is independent of the node service process. Therefore, the packet arrival process is a stationary process with average rate λ_i and is independent of the channel capacity varying process. Packets from different sources are buffered in a single queue and are fairly served using a first-in-first-out (FIFO) policy. To simplify the analysis, we assume that the buffer size of each node is infinite and that each packet has the same length L.

One important difference between the model in this paper and that in [16] is that a random access protocol is employed at the MAC layer. Because in wireless networks, such as ad hoc networks, the nodes in the same sensing range share one wireless channel, a transmission collision occurs when two nodes in the same sensing range transmit packets simultaneously. Hence, a protocol is needed to manage node transmissions. The main function of the random access protocol at the MAC layer is scheduling node transmissions to minimize the probability of transmission collisions thereby maximizing the utility of the network resource. Under the random access protocol control, a node in the network can only occupy the channel at a certain opportunity. Therefore, compared with the traditional physical channel model, the property of the MAC protocol can be modeled by a tuple $\langle \mathrm{P_{tr}}, \mathrm{P_{col}} \rangle$, where $\mathrm{P_{tr}}$in$(0, 1)$ and $\mathrm{P_{col}}$in$(0, 1)$ are the probabilities of a node transmitting in a random slot and a collision occurring, respectively. Note that the influence of the number of neighbors is included in the process of calculating $\mathrm{P_{tr}}$ and $\mathrm{P_{col}}$ [7,23,24].

It is worth pointing out that $\mathrm{P_{tr}}$ has different meanings for different MAC protocols. For example, for ALOHA, P_{tr} is defined as the probability that the channel is idle and the node is nonempty, which is the same as the probability of the backoff counter being 0 when the channel is idle in the CSMA/CA protocol. Existing works have investigated the values of $\mathrm{P_{tr}}$ and $\mathrm{P_{col}}$ for these well-known channel allocation protocols in great detail [23,24]. In this paper, we do not discuss the process of deriving the values of these two parameters. We denote the probability that node i, $i$in$\{1, 2, \cdots, N\}$, has the opportunity to start a packet transmission in a random slot by $P_{i,\text{tr}}$. Accordingly, the probability that a collision occurs is $P_{i,\text{col}}$. We assume that the values of $P_{i,\text{tr}}$ and $P_{i,\text{col}}$ are given.

At the physical layer, according to Shannon's capacity theory, the capacity of the wireless channel of node i at time t is $r_i(t) = B \log(1 + \mathrm{SNR}_i(t))$, where B is the channel bandwidth and $\mathrm{SNR}_i(t)$ is the signal-to-noise ratio (SNR) at the receiver at time t when the corresponding transmitter is owned by node i. The value of $\mathrm{SNR}_i(t)$ depends on the channel fading model and the distance between the transmitter and the receiver. The mean value of $\mathrm{SNR}_i(t)$ is $\mathrm{SNR}_{\text{avg},i}$.

3 End-to-end delay analysis

Because the end-to-end delay is the sum of the sojourn time that the packet experiences at each node along the path, the fundamental work in calculating the end-to-end delay is to analyze sojourn time. Unlike previous works [14,15], which assumed that traffic flows in a network are fluid flows and that the sojourn time is equal to the queuing delay, in this paper, the sojourn time consists of two components: packet service time and queuing delay. If the channel capacity of node i is $r_i(t)$, the packet service time is $\frac{L}{r_i(t)}$ at time t for the ideal MAC protocol. However, because we consider the effect of a random access policy in this paper, the packet service time cannot be directly calculated if $r_i(t)$ is given. Owning to the natural limitations of (2) and (4), the queuing delay cannot be derived by directly applying these two formulas. To estimate the queueing delay, we must solve two problems: one is extending EC or EB theory for application to the general case, such as the network model in this paper, while the other is making them suitable for packetized networks. To solve these two problems, we propose residual effective capacity, which can help us take good advantage of EC and EB theory to derive the queuing delay in the general case. In this section, we first introduce the concept of residual effective capacity before analyzing the packet service time and end-to-end delay performance of the network.

Before giving a definition for residual effective capacity, we define some notations that are used later. As shown in Figure 1, $S_i(t)$ is the service provided by the channel, which is defined in the same way as in [15]. The EB and EC functions of node i are referred as to $\alpha_i^{(b)}(u)$ and $\alpha_i^{(c)}(u)$, respectively.

3.1 Residual effective capacity

To understand the delay of a path, we need to ascertain the amount of channel capacity remaining after servicing the traffic demand. We call this remaining capacity the residual capacity $S_i^{(\mathrm{res})}(t)$ over time interval $[0,\, t]$, which can be represented as $S_i^{(\mathrm{res})}(t) = \max\{0, (S_i(t) - A_i(t))\}$. If $\lim_{t\to\infty}(S_i(t)-A_i(t)) \leqslant 0$, the network is unstable. In this paper, to investigate the delay performance of a stable wireless network, we enforce $\lim_{t\to\infty}(S_i(t) - A_i(t)) > 0$. The corresponding EC function of $S_i^{(\mathrm{res})}(t)$ is the residual effective capacity function denoted by $\alpha_i^{(\mathrm{res})}(u)$. To calculate $\alpha_i^{(\mathrm{res})}(u)$, we develop a property of EC function as described in Theorem 1.

Theorem 1. If the channel capacity $S(t)$ is equal to the sum of a finite sub-channels capacities $S^{(k)}(t)$, then its EC function $\alpha^{(c)}(u)$ is equal to the sum of sub-channels' EC functions $\alpha^{(ck)}(u)$. That is, if $S(t) = \sum_{k=1}^{K} S^{(k)}(t)$, $\alpha^{(c)}(u) = \sum_{k=1}^{K} \alpha^{(ck)}(u)$.

Proof. Without loss of generality, we only consider the case with two sub-channels; other cases can easily be proven by extending the argument. If $K = 2$, $S(t) = S^{(1)}(t) + S^{(2)}(t)$, and the theorem can be proved by proving $\alpha^{(c)}(u) = \alpha^{(c1)}(u) + \alpha^{(c2)}(u)$.

The asymptotic log-moment generating function of $S(t)$ for $\forall u \geqslant 0$, as defined in [15], is

$$\begin{aligned}\Lambda^{(c)}(-u) &= \lim_{t\to\infty}\frac{1}{t}\log E[e^{-uS(t)}] = \lim_{t\to\infty}\frac{1}{t}\log E[e^{-u(S^{(1)}(t)+S^{(2)}(t))}]\\ &= \lim_{t\to\infty}\frac{1}{t}\log(E[e^{-uS^{(1)}(t)}]E[e^{-uS^{(2)}(t)}])\\ &= \lim_{t\to\infty}\frac{1}{t}\log E[e^{-uS^{(1)}(t)}] + \lim_{t\to\infty}\frac{1}{t}\log E[e^{-uS^{(2)}(t)}]\\ &= \Lambda^{(c1)}(-u) + \Lambda^{(c2)}(-u). \end{aligned} \tag{5}$$

By substituting (5) into (3), the EC function of $S(t)$ can be calculated as

$$\alpha^{(c)}(u) = \frac{-\Lambda^{(c)}(-u)}{u} = \frac{-(\Lambda^{(c1)}(-u) + \Lambda^{(c2)}(-u))}{u} = \frac{-\Lambda^{(c1)}(-u)}{u} + \frac{-\Lambda^{(c2)}(-u)}{u} = \alpha^{(c1)}(u) + \alpha^{(c2)}(u). \tag{6}$$

Eq. (6) indicates that the theorem for $K = 2$ is correct. For other cases with $K \geqslant 3$, the theorem is obviously correct based on the results of $K = 2$. This completes the proof.

To calculate $\alpha_i^{(\mathrm{res})}(u)$, we investigate the relation between the traffic arrival process and channel capacity from a channel capacity point of view. Let us use an example to explain this relation. Assume that a certain channel with capacity $S(t)$ is shared by two different priority sources, and that the higher priority source generates $A_h(t)$ bits over interval $[0,\ t]$. Because service scheduling is preemptive, the channel capacity for the lower priority source is equal to $\max\{0, (S(t) - A_h(t))\}$. This indicates that the traffic load of the higher priority source is the negative value of channel capacity of the lower priority source. Now we are back to the scenario where each node occupies the whole capacity of a wireless channel. When we analyze residual effective capacity, the traffic load can be regarded as negative channel capacity, which is the same as the case in the previous example. Because traffic load is negative channel capacity when the node calculates residual effective capacity, according to Theorem 1, the residual effective capacity function of node i can be calculated as

$$\alpha_i^{(\mathrm{res})}(u) = \alpha_i^{(c)}(u) - \alpha_i^{(b)}(u). \tag{7}$$

Because (7) does not impose any restriction on the traffic model, it is always valid for all traffic and channel models. Therefore, we can use (7) to calculate the residual capacity function for the general case with any traffic model and wireless channel model.

As shown in Figure 1, $\alpha_i^{(b)}(u)$ is monotonically increasing and $\alpha_i^{(c)}(u)$ is monotonically decreasing as u increases. In addition, when $\alpha_i^{(\mathrm{res})}(u)$ is equal to 0, it implies that all the available capacity has been completely used. The QoS exponent u obtained under this situation, which is called the critical value θ^*_{Bi}, represents the QoS level of the node. The QoS level of a path can be computed based on the QoS of each node along the path. Then the network can determine whether to accept a new flow with a certain QoS requirement. Therefore, the critical value not only helps us estimate delay performance, it can also be used as a metric to make admission control decisions to guarantee traffic QoS and improve network utility.

3.2 Service time analysis

As indicated by (4) and (7) the estimation of queuing delay requires an understanding of the packet service process. Therefore, in this subsection, we investigate the packet service process in our network scenario. Despite the channel capacity being given, the service process cannot be obtained directly using $r_i(t)$ because the packet service time of node i at time t may not equal $\frac{L}{r_i(t)}$ owning to the random access protocol. Therefore, before estimating the queuing delay and end-to-end delay, we first analyze the packet service time of nodes when the access channel opportunity is controlled by the MAC protocol.

Because the nodes in the same sensing range share limited wireless channels, the actual service capacity of node i at time t may not be equal to $r_i(t)$. Let the actual channel capacity of node i at time t be $r'_i(t)$. Because the queuing delay calculation is based on the EC function of the service capacity, we must first derive the value of $r'_i(t)$ and the EC function of $r'_i(t)$.

Despite the channel being idle, if the node does not transmit packets on this channel, the channel is not available for the node. In other words, the channel is available for node i only when the node is scheduled by the access protocol. Let $C_i(t)$ denote a random variable representing the available channel capacity for node i at time t. Because the transmission probability $(P_{i,\mathrm{tr}})$ of node i is given, $C_i(t)$ can be expressed as

$$C_i(t) = \begin{cases} r_i(t), & P_{i,\mathrm{tr}}; \\ 0, & 1 - P_{i,\mathrm{tr}}. \end{cases} \tag{8}$$

In a distributed wireless network, all nodes in the same sensing range use local information to make access decision. Hence, there may be more than one node transmitting simultaneously, which causes a collision. If a collision occurs in the sensing range of node i, the transmission of node i is unsuccessful. To guarantee a reliable transmission, the packet is retransmitted when the channel become available. In other words, this part of the available capacity is wasted and cannot provide any service for backlogged packets at node i. The probability of waste available capacity is equal to the collision probability. According to the analysis in [23,24], the value of the collision probability $P_{i,\mathrm{col}}$ of node i is related to the access protocol and the number of neighbors. For example, for the IEEE 802.11 distributed coordination function (DCF) protocol, $P_{i,\mathrm{col}} = 1 - \prod_{j=1}^{N_i} (1 - P_{j,\mathrm{tr}})$, where N_i is the number of node i's neighbors, while in a TDMA system, $P_{i,\mathrm{col}} = 0$. Given $P_{i,\mathrm{col}}$, according to (8), the actual service capacity of node i at time t can be calculated as

$$r'_i(t) = \begin{cases} r_i(t), & P_{i,\mathrm{tr}}(1 - P_{i,\mathrm{col}}); \\ 0, & 1 - P_{i,\mathrm{tr}}(1 - P_{i,\mathrm{col}}). \end{cases} \tag{9}$$

Substituting (9) into (3), the EC function of node i at the MAC layer denoted by $\alpha_i^{(c)}(u)$ can be expressed as

$$\alpha_i^{(c)}(u) = \frac{-1}{u} \lim_{t\to\infty} \frac{\log E\left[\exp(u\sum_{\tau=0}^{t} r'_i(\tau))\right]}{t}. \tag{10}$$

Although (10) can help us obtain the queuing delay of packets at node i, it does not give us the packet service time. In packetized networks, packet service time is an important part of the sojourn time. Next, we derive a formula for calculating the average packet service time $\bar{d}_i$ of node i.

As a packet may suffer l+1 MAC overhead time as a result of accessing the channel, that is, l retransmissions and one successful transmission during the service process, $\bar{d}_i$ can be expressed as

$$\bar{d}_i = \sum_{l=0}^{\infty} P_{i,\mathrm{col}}^l (1 - P_{i,\mathrm{col}}) \left(E[\frac{L}{r_i(t)}] + lE[T_{i,\mathrm{col}}] + (l+1)E[T_{i,\mathrm{over}}] \right), \tag{11}$$

where $E[\frac{L}{r_i(t)}]$ is the average packets transmission time of node i, $E[T_{i,\mathrm{col}}]$ is the average time owning to transmission collision, and $E[T_{i,\mathrm{over}}]$ is the mean value of MAC overhead time as a results of accessing the channel.

With a random access protocol, for node i, four cases can occur: 1) none of the nodes in node i's sensing range transmits; 2) at least one neighbor of node i transmits when node i does not transmit; 3) at least one neighbor as well as node i transmit at the same time (transmission collision); and 4) no other node transmits when node i transmits (successful transmission). Of these scenarios, cases 1 and 2 result in time wasted for node i where the time associated with these two cases is $T_{i,\mathrm{over}}$. Obviously, collision time occurs in case 3. For the l-th collision, the collision time $T_{i,\mathrm{col}}^{(l)}$ is $T_{i,\mathrm{col}}^{(l)} = \frac{L}{\min_{j\mathrm{in}[1,N_i^{(l)}]}(r_i(t), r_j(t))}$, where $N_i^{(l)}$ is the number of nodes taking part in this collision. Let the neighbors of node i be ordered according to their average packet transmission time recorded by Ω_i. The elements in Ω_i satisfy $E[\frac{L}{r_1(t)}] \geqslant E[\frac{L}{r_2(t)}] \geqslant \cdots \geqslant E[\frac{L}{r_j(t)}] \geqslant \cdots \geqslant E[\frac{L}{r_{N_i}(t)}]$. Now, $E[T_{i,\mathrm{col}}]$ can be calculated as

$$\begin{aligned} E[T_{i,\mathrm{col}}] =& \sum_{j=1}^{N'_i} \frac{P_{j,\mathrm{tr}} \prod_{m=1}^{j-1} (1 - P_{m,\mathrm{tr}})}{P_{i,\mathrm{col}}} E\left[\frac{L}{r_j(t)}\right] \\ &+ \frac{\prod_{j=1}^{N'_i} (1 - P_{j,\mathrm{tr}})(1 - \prod_{k=N'_i}^{N_i} (1 - P_{k,tr}))}{P_{i,\mathrm{col}}} E\left[\frac{L}{r_i(t)}\right], \end{aligned} \tag{12}$$

where N'_i is the number satisfying $E[\frac{L}{r_{N'_i}(t)}] \geqslant E[\frac{L}{r_i(t)}] \geqslant E[\frac{L}{r_{N'_i+1}(t)}]$.

The probabilities that cases 1 and 2 occur are $(1 - P_{i,\mathrm{tr}}) \prod_{j=1}^{N_i} (1 - P_{j,\mathrm{tr}})$ and $(1 - P_{i,\mathrm{tr}})(1 - \prod_{j=1}^{N_i}(1 - P_{j,\mathrm{tr}}))$, respectively. The corresponding times associated with cases 1 and 2 are one slot length and $T'_{i,\mathrm{tr}}$, respectively. $T'_{i,\mathrm{tr}}$ is determined by the minimum sequence of nodes in case 2. Similar to the analysis of $E[T_{i,\mathrm{col}}]$, the mean value of $T'_{i,\mathrm{tr}}$ can be derived as $E[T'_{i,\mathrm{tr}}] = \sum_{j=1}^{N_i} \frac{P_{j,\mathrm{tr}} \prod_{k=1}^{j-1} (1 - P_{k,tr})}{1 - \prod_{m=1}^{N_i} (1 - P_{m,\mathrm{tr}})} E[\frac{L}{r_j(t)}]$. Therefore, the average value of $T_{i,\mathrm{over}}$ can be calculated as

$$E\left[T_{i,\mathrm{over}}\right] = \sum_{n=0}^{\infty} (1 - P_{i,\mathrm{tr}})^n \left(\sum_{m=0}^{n} C_n^m \left(\prod_{j=1}^{N_i} (1 - P_{j,\mathrm{tr}}) \right)^m \left(1 - \prod_{j=1}^{N_i} (1 - P_{j,\mathrm{tr}}) \right)^{n-m} \left(m + (m - n)E[T'_{i,\mathrm{tr}}] \right) \right). \tag{13}$$

By Substituting (12) and (13) into (11), $\bar{d}_i$ can be obtained.

3.3 End-to-end delay analysis

Now, having obtained the EC function of the service process, we can use the residual effective capacity function to derive the formula for estimating the average end-to-end delay. First we describe how to find the queuing delay using the QoS exponent based on residual effective capacity function. Let the time that a random packet spends in node i be W_i, and the cumulative distribution function (CDF) of W_i be $F_{iW}(x)$.

By setting $\alpha_i^{(\mathrm{res})}(u) = 0$, the QoS exponent θ_{Bi}^* of node i can be obtained. Using the large deviation theory and the results in [15], $F_{iW}(x)$ can be derived as

$$F_{iW}(x) = \Pr(W_i \leqslant x) = 1 - \gamma_i e^{-\theta_i^* x}, \tag{14}$$

where $\theta_i^* = \alpha_i^{(b)}(\theta_{Bi}^*) \times \theta_{Bi}^*$, and γ_i is the probability that the buffer is nonempty at a randomly chosen slot. According to queuing theory, γ_i is also equal to the probability that node i is busy at a randomly chosen time. That is, $\gamma_i = \lambda_i \times \bar{d}_i$, where $\bar{d}_i$ is calculated in (11). According to probability theory, the average queuing delay can be derived as

$$E[W_i] = \int_0^\infty x \mathrm{d}F_{iW}(x) = \frac{\gamma_i}{\theta_i^*}. \tag{15}$$

The sojourn time D_i of a random packet at node i is equal to the sum of the queuing time W_i and the packet service time. Using the results in (11) and (15), the average sojourn time $E[D_i]$ can be calculated as

$$E[D_i] = E[W_i] + \bar{d}_i = \frac{\gamma_i}{\theta_i^*} + \bar{d}_i. \tag{16}$$

For the wireless networks we investigated in this paper, the service time and CDF of the queuing delay of packets at node i, $1 \leqslant i \leqslant N$, are independent of the others. The sojourn time of a packet at each node is independent of others on the path. Let $W^{(H)}$ and $D^{(H)}$ denote the end-to-end queuing delay and end-to-end delay of a random H-hop path, respectively. According to (14) and (16), the CDF of $W^{(H)}$ and mean value of $D^{(H)}$ can be derived in (17) and (18), respectively.

$$F_{W^{(H)}}(x) = \int_0^x (f_1 * f_2 * \cdots * f_H)(t) \mathrm{d}t, \tag{17}$$

where $f_h(x), h \mathrm{in} [1, H]$ is the differential of $F_{hW}(x)$.

$$E[D^{(H)}] = \sum_{h=1}^{H} E[D_h] = \sum_{h=1}^{H} \left(\frac{\gamma_h}{\theta_h^*} + \bar{d}_h \right). \tag{18}$$

Let the set of all communication pairs in a network be denoted by F where $F = \{\langle i,j \rangle | i \text{ send packets to } j\}$. $\lambda_{\langle i,j \rangle}$ and $H_{\langle i,j \rangle}$ are the average packet arrival rate and path length of communication $\langle i,j \rangle$, respectively. Because the average end-to-end delay of a random path is given by in (18), the average end-to-end delay of the lossless network can be calculated as

$$E[D] = \frac{1}{\sum\limits_{\langle i,j \rangle \mathrm{in} F} \lambda_{\langle i,j \rangle}} \sum_{\langle i,j \rangle \mathrm{in} F} \lambda_{\langle i,j \rangle} E[D^{H_{\langle i,j \rangle}}]. \tag{19}$$

As mentioned in [15], it is difficult to calculate the EC function for a general channel model, while in some cases, the calculation is even impossible. Owning to the additional effect of MAC protocol $\langle P_{\mathrm{tr}}, P_{\mathrm{col}} \rangle$, the calculation of the EC function becomes more difficult and this difficulty results in the high complexity of calculating the residual effective capacity function. However, from (18) and (19), we find that both the calculation of the end-to-end delay of a chosen path and the network are based on two parameters: γ_i and θ_i^*. In this paper, we develop a simple algorithm to measure these two parameters and use the measured results to computer the end-to-end delay.

3.4 An algorithm for measuring parameters

Because the network is a stable system, each node in the network is also a stable queuing system. The average queuing delay of packets at node i is

$$E[W_i] = E[Q_i]\bar{d}_i + E[\tau_i], \tag{20}$$

where $E[\tau_i]$ is the average residual service time of the packet being served and $E[Q_i]$ is the average number of packets in the queue.

Substituting (20) into (15), θ_i^* can be calculated as

$$\theta_i^* = \frac{\gamma_i}{E[Q_i]\bar{d}_i + E[\tau_i]}. \tag{21}$$

From (21), we find θ_i^* can be calculated using the values of γ_i, $E[\tau_i]$, $E[Q_i]$ and λ_i. A sampling method is proposed to obtain these measured values.

At the n-th sampling epoch, we record the following values: (1) $S_{\rm in}$ indicates whether a packet is being served by node i, where

$$S_{\rm in} = \begin{cases} 1, & \text{a packet is in service,} \\ 0, & \text{no packet is in service,} \end{cases}$$

(2) $Q_{\rm in}$ is the number of packets in the queue, (3) $\tau_{\rm in}$ is the residual service time of the packet in service, and (4) $P_{\rm in}$ is the number of packets arrived in the previous sample interval. The mean values of $S_{\rm in}$, $Q_{\rm in}$, $\tau_{\rm in}$ and $P_{\rm in}$ are γ_i, $E[Q_i]$, $E[\tau_i]$ and λ_i, respectively.

Next, we substitute the sample average results into (21) to obtain θ_i^*. By substituting θ_i^* and $E[S_{\rm in}]$ into (14), (16), (18) and (19), the distribution of the queuing delay and the sojourn time at node i, and the average end-to-end delay of a random path and the network can be calculated, respectively.

4 Simulation results and applications

In this section, we provide various numerical examples to verify the correctness of our analysis and an example application to show the usefulness of our results. Owning to the complexity of calculating the EC function, we first implement a simple network scenario in MATLAB. Then, with the help of the OPNET modeler, we further verify our analytical model. Finally, we present some applications of our analytical results.

4.1 Numeric results

The network parameter settings in the MATLAB simulation platform are the same as those in the IEEE 802.11g standard. New traffic arrival processes for all nodes in the network are Poisson processes with the same parameters and a packet size of 9720 bits. The packet interval for background traffic is an exponential distribution with mean 0.005 s. We first compared the simulation results against the calculation results in a single-hop wireless network to verify that our analytical model provides an accurate estimation of the sojourn time of node. In this network, the values of $P_{i,{\rm tr}}$ and $P_{i,{\rm col}}$ satisfy $P_{i,{\rm tr}} = \frac{1}{N+1}$ and $P_{i,{\rm col}} = 1-(1-P_{i,{\rm tr}})^N$, where $N+1$ is the number of nodes in the network. The results are illustrated in Figure 2.

The results in Figure 2 show the relation between sojourn time, number of nodes, and traffic load because the end-to-end delay of a single-hop network is equal to the sojourn time at the nodes. Because there is no relay traffic in a single-hop network, the EB function can be directly calculated according to the new traffic information. In Figure 2, the dashed lines denote simulation results, while the solid lines are the calculations obtained using (16). Figure 2 confirms that our method can provide a good estimation of the sojourn time at nodes.

To verify our method in a multi-hop wireless network, we considered a 6-hop path with the values of the critical parameters in the MAC layer for each hop as given in Table 1. The other network parameters were the same as in the single-hop network. The average end-to-end delay with varying traffic arrival

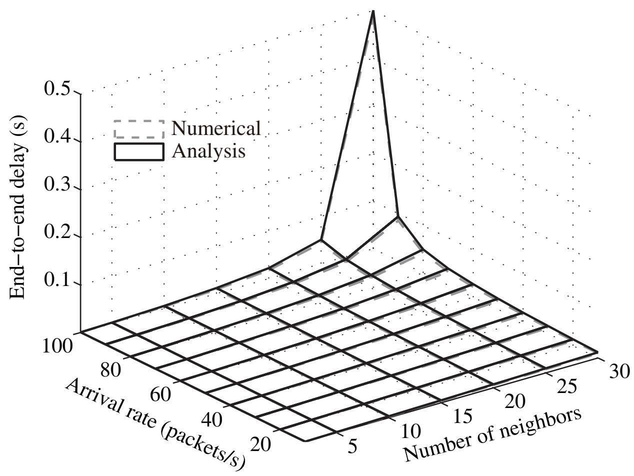

Figure 2 End-to-end delay of a single-hop network

Table 1 Simulation parameters

Item	Hop-1	Hop-2	Hop-3	Hop-4	Hop-5	Hop-6
$p_{\rm tr}$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{8}$	$\frac{1}{5}$
$p_{\rm col}$	$1-\left(1-\frac{1}{6}\right)^4$	$1-\left(1-\frac{1}{4}\right)^2$	$1-\left(1-\frac{1}{5}\right)^3$	$1-\left(1-\frac{1}{6}\right)^4$	$1-\left(1-\frac{1}{8}\right)^6$	$1-\left(1-\frac{1}{5}\right)^3$

rates is illustrated in Figure 3. Although we assumed an infinite buffer for each node in the analysis process, the results are also applicable in the case of a finite buffer. In the simulation, the buffer size for each node is 200 packets. When a node is under heavy load, many packets are dropped owning to overflow. Consequently, the actual arrival rate of the node slowly increases after a particular traffic arrival rate (for example, in Figure 3, the particular value is 160 packets/s). Accordingly, the increase in the average end-to-end delay tends to be flat.

We used the OPNET Modeler to conduct simulations to further verify our method. We tested our results on a random topology network of size 1 km × 1 km. There were 50 nodes in the network, with each node generating the same size 8192 bits packets; the destination of a packet was randomly selected as one of the nodes. The wireless channel was a Rayleigh fading channel and with 20 MHz bandwidth. All the nodes were immobile, and there was no Doppler shift. Here, the analytical results were obtained based on the algorithm proposed in Subsection 3.4, using a sampling interval of 1 ms and calculating the mean value per ten thousand sample values.

Because we consider the influence of traffic burstiness, channel time-varying and the access protocol on the end-to-end delay, we verified our analytical model in different situations. First, under the same MAC protocol, we compared the results of different traffic models with varying arrival rates. Then, the results for different MAC protocols and multiple average SNR values were considered. Three traffic models can be created by setting different values for MMPP2: 1) Poisson process ($\lambda_1 = \lambda_2$); 2) ON-OFF model ($\lambda_2 = 0$); and 3) constant interval time. As the probability access protocol was considered in the MATLAB simulations, in the OPNET simulation, two types of MAC protocols were used: an ideal MAC protocol and CSMA/CA.

Comparisons of the results for these three traffic models with different arrival rates are displayed in Figures 4 and 5. The results shown in these figures are the average end-to-end delay of the networks with the aforementioned MAC protocols, where "ana" and "sim" represent the analytical and simulation results, respectively. The results of the cases with different average SNR values are shown in Figures 6 and 7. It can be seen from Figures 4 to 7 that our method not only correctly estimates the delay in the cases with a constant arrival process, but also provides accurate estimations for other traffic models with burstiness.

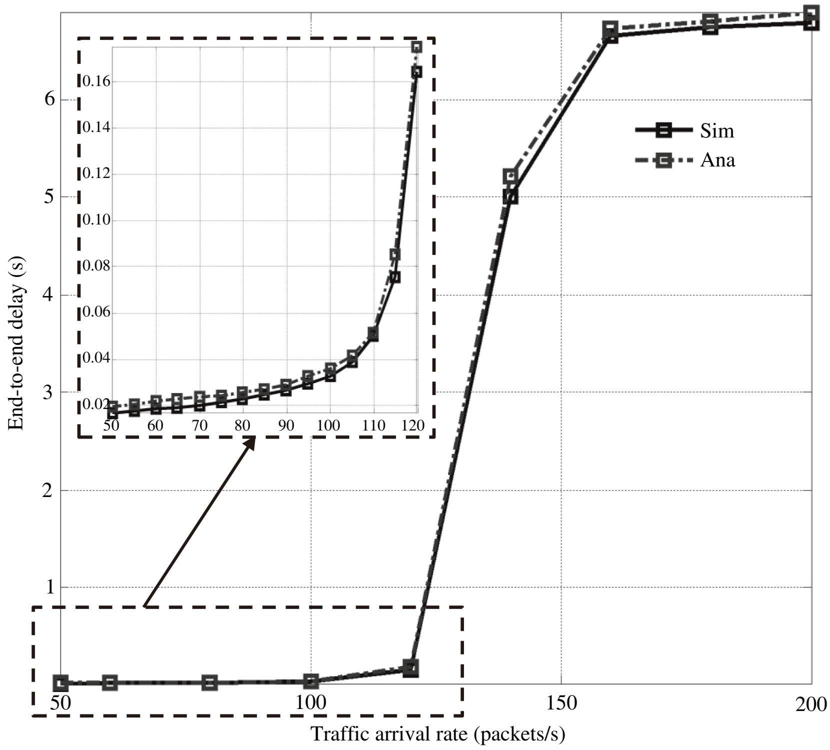

Figure 3   End-to-end delay of a multi-hop network

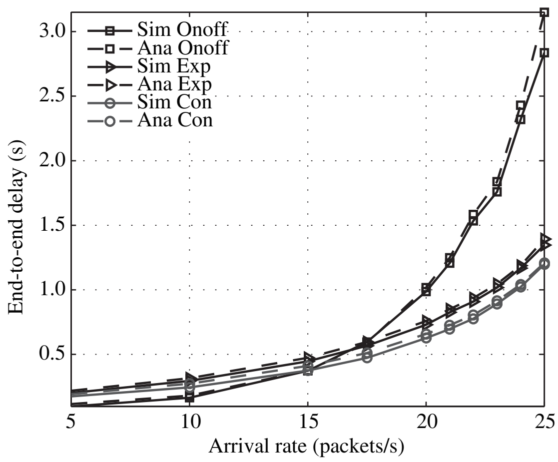

Figure 4   Results for CSMA protocol

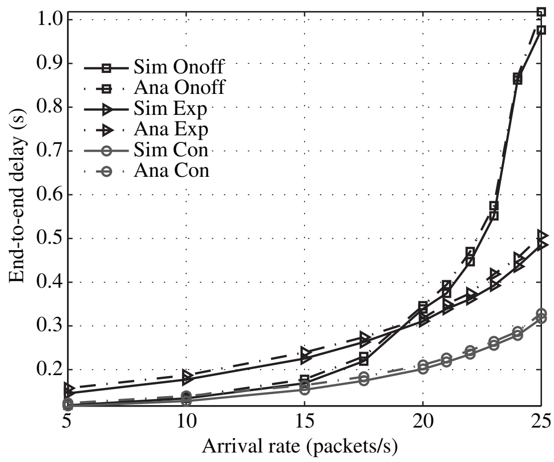

Figure 5   Results for ideal MAC protocol

4.2 Applications

The work discussed in Section 4 has provided methods to derive the end-to-end delay for a random chosen path and network, while the simulation results and numerical calculations in Subsection 4.1 verify the accuracy of our analysis. After obtaining the end-to-end delay performance, different algorithms can be designed to guarantee QoS from different aspects. One application is introduced bellow as an example of using the results given in Section 3 as a metric to ensure QoS provisioning.

Admission control schemes are used to grant/deny new traffic arriving at the network by considering the network load condition or the values of the QoS metrics. For real-time traffic, end-to-end delay is the QoS metric for admission control, because some available routing protocols, such as dynamic source routing (DSR) can ensure that the source node has information of the whole path between the source and destination nodes. While the network is functioning in a stable state, a new traffic flow with an end-to-end delay requirement arrives. The source node checks its routing table, if there is no path information

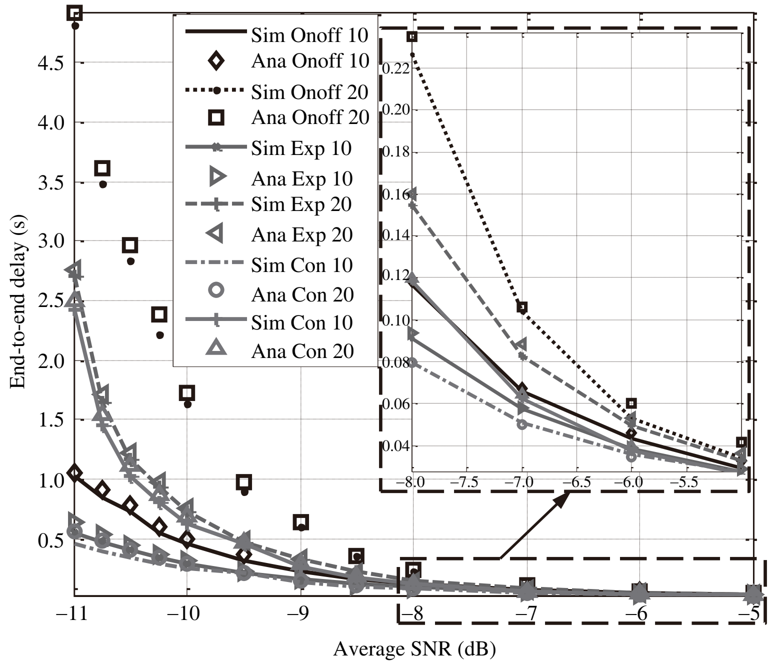

Figure 6 End-to-end delay with varying SNR under the CSMA protocol

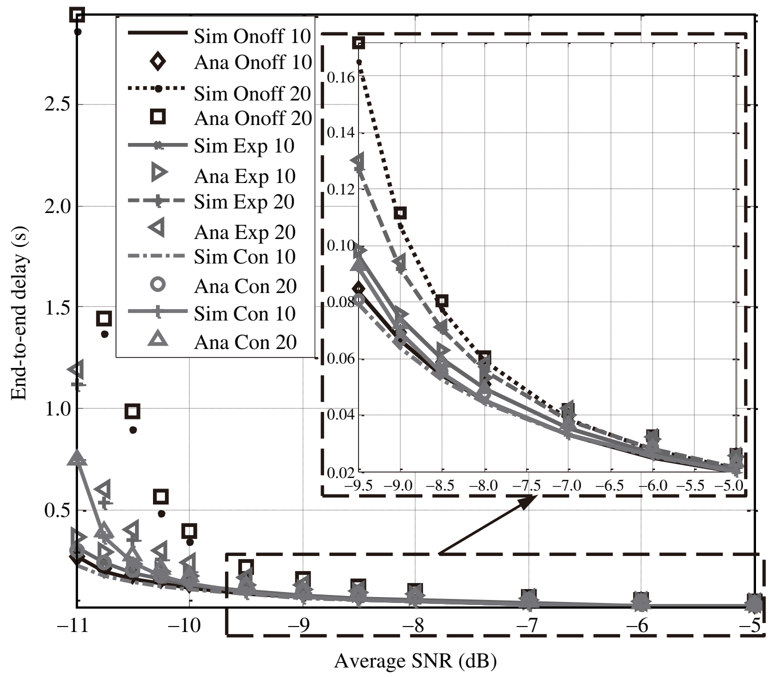

Figure 7 End-to-end delay with varying SNR under the ideal MAC protocol

or the path has expired, the source node uses a source routing protocol to find the path and collect end-to-end delay information of the path. Based on the delay information of an existing route path or new path between the source and destination, the source node determines whether this new arrival can be supported without violating the ongoing traffic. If the answer is yes, the new flow is allowed access; otherwise, it is denied.

5 Conclusion

In this paper, we extended the application range of EB and EC theory and exploited them in dealing with the general case. Based on this extension, we proposed an accurate method for analyzing the end-to-end delay performance of multi-hop wireless networks and developed a simple algorithm to obtain the values of critical parameters for the analytical method. Numerical results indicate that this method provides a good estimation of the average end-to-end delay. Finally, we discussed an application of our results, namely, an efficient admission control for real-time traffic in multi-hop wireless networks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61172079, 61231008, 61201141), National Basic Research Program of China (973) (Grant No. 2009CB320404), the 111 Project (Grant No. B08038), National S&T Major Project (Grant No. 2010ZX03003-001) and Shaanxi Province Science and Technology Research and Development Program (Grant No. 2011KJXX-40)

References

1 Zhao H T, Wang S, Wei J B, et al. Model-based approach for available bandwidth prediction in multi-hop wireless networks. Sci China Inf Sci, 2011, 54: 1916–1927
2 Chai R, Wang X J, Chen Q B, Svensson T. Utility-based bandwidth allocation algorithm for heterogeneous wireless networks. Sci China Inf Sci, 2013, 56: 022313
3 Xiang L, Ge X H, Wang C X, et al. Energy efficiency evaluation of cellular networks based on spatial distributions of traffic load and power consumption. IEEE Trans Wirel Commun, 2013, 12: 961–973
4 Li H K, Cheng Y, Zhou C, et al. Minimizing end-to-end delay: a novel routing metric for multi-radio wireless mesh networks. In: Proceedings of IEEE INFOCOM, Rio de Janeiro, 2009. 46–54
5 Ge X H, Huang K, Wang C X, et al. Capacity analysis of a multi-cell multi-antenna cooperative cellular network with co-channel interference. IEEE Trans Wirel Commun, 2011, 10: 3298–3309
6 Yu S M, Kim S L. End-to-end delay in wireless random networks. IEEE Commun Lett, 2010, 14: 109–111
7 Xie M, Haenggi M. Towards an end-to-end delay analysis of wire-less multihop networks. Ad Hoc Netw, 2009, 7: 849–861
8 Gupta G A, Shroff N B. Delay analysis and optimality of scheduling policies for multihop wireless networks. IEEE/ACM Trans Netw, 2011, 19: 129–141
9 Bisnik N, Abouzeid A. Queuing network models for delay analysis of multihop wireless ad hoc networks. Ad Hoc Netw, 2009, 7: 79–97
10 Yeung S N, Lehoczky J. End-to-end delay analysis for real-time networks. In: Proceedings of IEEE RTSS Processing, London, 2001. 209–309
11 Liebeherr J, Yashar G F , Burchard A. On the impact of link scheduling on end-to-end delay in large networks. IEEE J Sel Areas Commun, 2011, 29: 1009–1020
12 Le L, Hossain E. A analytical model for ARQ cooperative diversity in multi-hop wireless network. IEEE Trans Wirel Commun, 2008, 7: 1786–1791
13 Almut B, Liebeherr J, Patek S D. A min-plus calculus for end-to-end statistical service guarantees. IEEE Trans Inf Theory, 2006, 52: 4105–4114
14 Chen Y, Yang Y, Darwazeh I. A cross-layer analytical model of end-to-end delay performance for wireless multi-hop environments. In: Proceedings of IEEE Global Telecommunications Conference, Miami, 2010. 1–6
15 Wu D P, Negi R. Effective capacity: A wireless link model for support of quality of service. IEEE Trans Wirel Commun, 2003, 2: 630–643
16 Jiao W G, Sheng M, Zhang Y, et al. On end-to-end delay of multi-hop wireless networks. In: Proceedings of IEEE 77th Vehicular Technology Conference, Dresden, 2013. 1–5
17 Jiang Y M, Liu Y. Stochastic Network Calculus. Berlin: Springer, 2008
18 Elwalid A I, Mitra D. Effective bandwidth of general markovian traffic source and admission control of high speed networks. IEEE/ACM Trans Netw, 1993, 1: 329–343
19 Chang C S, Thomas J A. Effective bandwidth in high-speed digital networks. IEEE J Sel Areas Commun, 1995, 13: 1091–1100
20 Xu D, Feng Z Y, Zhang P. Effective capacity of delay quality-of-service constrained spectrum sharing cognitive radio with outdated channel feedback. Sci China Inf Sci, 2013, 56: 062311
21 Leung K K, Massey W A, Whitt W. Traffic models for wireless communication networks. IEEE J Sel Areas Commun, 1994, 12: 1353–1364
22 Giacomazzi P. Closed-form analysis of end-to-end network delay with markov-modulated poisson and fluid traffic. Comput Commun, 2009, 32: 640–648
23 Felemban E, Ekici E. Single hop IEEE 802.11 DCF analysis revisited accurate modeling of channel access delay and throughput for saturated and unsaturated traffic cases. IEEE Trans Wirel Commun, 2011, 10: 3256–3266
24 Dai L. Stability and delay analysis of buffered aloha networks. IEEE Trans Wirel Commun, 2012, 11: 2707–2719