Ann Oper Res (2015) 233:171–180
DOI 10.1007/s10479-013-1472-7

Stochastic modeling for delay analysis of a VoIP network

Vandana Gupta · S Dharmaraja · Viswanathan Arunachalam

Published online: 8 October 2013

Abstract VoIP (“Voice Over Internet Protocol”) is the transmission of voice communication through the Internet via IP-based telephony networks. VoIP has become very popular over recent years due to the cost advantages for consumers and businesses compared to the traditional telephony networks. Since it is deployed on packet-based networks, one of the major Quality of Service (QoS) concerns of VoIP technology is the average end-to-end connection delay. The objective of this paper is to present a queuing model for obtaining the end-to-end delay of a VoIP connection. The paper first describes all the partial delay components, and their mathematical formulations. Subsequently, based on all the partial delay components, a queuing model for the end-to-end delay of a VoIP connection is presented. The proposed queueing model is analyzed using a generalized stochastic Petri net (GSPN) model. From the GSPN model, we obtain numerical results for the end-to-end delay, which are presented graphically. The results are in accordance with the expected behavior of delay in a VoIP network.

Keywords VoIP · Non-Markovian queue · Priority · Retrial · Generalized stochastic Petri net (GSPN) · End-to-end delay

1 Introduction

VoIP (Voice-over-IP) is the technology that enables people to use the Internet as the transmission medium for voice communications (Karapantazis and Pavlidou 2009). It refers to

V. Gupta
Department of Operational Research, University of Delhi, Delhi, India
e-mail: me.vandana.gupta@gmail.com

S Dharmaraja (✉)
Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, India
e-mail: dharmar@maths.iitd.ac.in

V. Arunachalam
Departamento de Estadistica, Universidad Nacional de Colombia, Bogota, Colombia
e-mail: varunachalam@unal.edu.co

the transmission of voice using IP technologies over packet switched networks, and consists of a set of facilities and protocols for managing the transmission of voice packets using IP. Internet Telephony is one of the typical applications of VoIP which has become very popular nowadays because of its cost efficiency. One of the main performance indicators that characterize the quality of voice communications over the Internet is the average end-to-end delay. It is one of the major issues in packet-based networks which have a direct impact on the QoS (Shim et al. 2003). Considering the advancement in technology over the years, traditional voice communication over the PSTN (Public switched telephone network) is characterized by its high quality. Hence, when it comes to VoIP, stern QoS constraints must be met in order to provide the same quality level.

VoIP connection delay is described as the amount of time it takes for speech to exit the speaker's mouth and reach the listener's ear. It is caused when voice (data) packets take more time than expected to reach their destination. This causes some disruption in the voice quality. There are a few analytical work available on VoIP delay. In Baronak and Halas (2007), a mathematical formulation of VoIP connection delay model has been proposed. The paper handles all the partial delay components, the mechanism of their generation, and their mathematical formulation. Thereafter based on the mathematical formulation of all partial delay components, the final mathematical model of the whole VoIP call delay is created. In Voznak and Hromek (2008), the authors focus on the design of a mathematical model of end-to-end delay of a VoIP connection, in particular on a delay variation. It also describes all partial delay components and its mathematical formulations. A new approach to the delay variation model is presented in this paper using M/D/1 queue, and the model is validated by an experiment. The technical report (Rezac et al. 2010) deals with the mathematical model of the end-to-end delay and delay variation in VoIP connections going through a two priority queue serving system. A comparative analysis of three queuing scenarios in VoIP, i.e., First-in-first-out queuing, Priority queuing and Weighted-Fair queuing is presented in Rashed and Kabir (2010).

In all the above mentioned literature, and to the best of our knowledge, a queuing representation for the total end-to-end VoIP connection delay has not been proposed so far. This motivated us to propose a queuing model for the end-to-end connection delay in a VoIP network. However we perform the analysis of the proposed queueing model using generalized stochastic Petri net (GSPN) modeling technique, and obtain the numerical results for the average end-to-end delay.

The rest of this paper is organized as follows. The complete description of VoIP delay is explained in Sect. 2. Section 3 presents the proposed queueing model for the end-to-end VoIP connection delay and the corresponding GSPN model. Numerical illustrations of the results obtained from the GSPN model are presented in Sect. 4. Finally concluding remarks are presented in Sect. 5.

2 VoIP delay

The two sources of delay in packet telephony are transit delay and jitter. Transit delay is the amount of time it takes for the signal to travel from the speaker, through all of the network elements, to the recipient. When a packet is delayed, listener will hear the voice later than he should. If the delay is constant, and not big, the conversation can be acceptable. But unfortunately, the delay is not always constant, and varies depending on some technical factors. This variation in delay is called jitter, which causes damage to voice quality. There are many causes of jitter: router congestion, parallel router operation, changes in physical pathways between the terminal clients, transmission issues, codec issues, and processor issues.

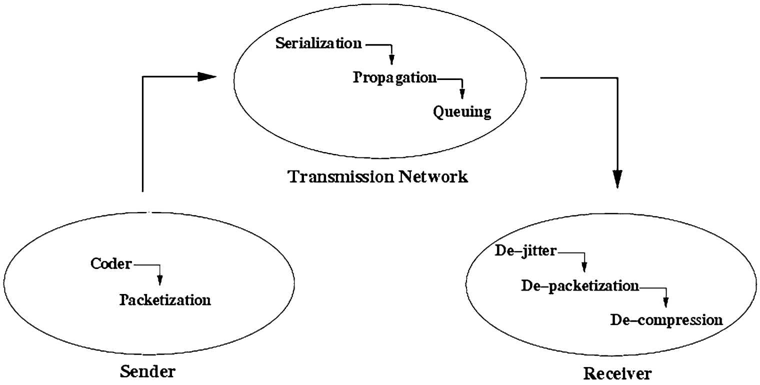


Fig. 1 Delay components and places of their origin

There are several components of transit delay in a VoIP network. These components differ from each other as to where they are generated, method of generation and some other characteristics. For a better understanding of the transit delay components, assume that whole VoIP network can be divided into three parts: Sender's end, Transmission network, and Receiver's end. The various delay components classified according to their origin is represented in Fig. 1, and are explained below.

- **Coder delay:** Coder delay depends on the used codec. It has two components: the frame size delay and the look-ahead delay. Their values are exactly defined for any particular coder.
- **Packetization delay:** The packetization delay rises during the process of data blocks encapsulation into packets, which are consequently transmitted by the network. It is set as multiples of the packetization period used by a particular codec and specifies how many data blocks are transmitted in one packet. The packetization delay (T_{PD}) is given as:

$$T_{PD} = \frac{P_S}{C_{BW}} \text{ ms}$$

where P_S is the payload size (b) and C_{BW} is the codec bandwidth (kbit/s).

- **Serialization delay:** Serialization delay depends on the transmission rate of the used interface. The transmission of packets takes some time which depends on the transmission medium rate and on the size of packet. The serialization delay (T_{Ser}) is given as:

$$T_{Ser} = \frac{P_S + H_L}{L_S} \text{ ms}$$

where H_L is the header length (b) and L_S is the line speed (kbit/s).

- **Propagation delay:** This delay relates to the physical environment of the propagation medium. It depends on the transmission technology used, in particular on the distance over which the signal is transmitted. Nowadays networks are mostly built on single mode optical fibers. The speed of light in optical fiber is $v = 2.07 \times 10^8$ (m/s). Therefore, the propagation delay (T_{Prop}) can be defined as:

$$T_{Prop} = \frac{L}{v} \text{ ms}$$

where L is the line length (km).

- **Queuing delay:** This delay occurs in active elements of the transmission network, in particular in the router queues. When packets are held in a queue because of congestion on an outbound interface, the result is queuing delay. This delay is the most significant part of the jitter. It is a variable delay.
- **De-jitter delay:** Because speech is a constant bit-rate service, the jitter from all the variable delays must be removed before the signal leaves the network. This is accomplished by a de-jitter buffer at the far-end (receiving) router/gateway. The de-jitter buffer transforms the variable delay into a fixed delay. The de-jitter buffers can be adaptive, but the maximum delay is fixed. Its size is typically adjusted as a multiple of the packetization delay.
- **De-packetization delay:** The de-packetization is a reverse packetization and therefore the size of de-packetization delay of one block in the frame is in correlation with its packetization delay.
- **De-compression delay:** The decompression delay, depends on the compressing algorithm selection. On an average, the decompression delay is approximately 10 % of the compressing codec delay for each voice block in the packet. This decompression delay (T_{DCD}) can be defined as:

$$T_{DCD} = 0.1 \times N \times T_{CD} \text{ ms}$$

where T_{CD} is coder delay (ms) and N is the number of voice blocks in the packet.

Hence, from the above discussion, it can be observed that there are two distinct parts of the end-to-end VoIP connection delay, a fixed part and a variable part. Queuing delay, which is the time spent in the queues, is the only variable part of the end-to-end delay and depends on current network load. All the other components of the end-to-end delay are fixed. Hence this paper focuses on the creation of a queuing model for the end-to-end connection delay in a VoIP network.

In a VoIP network, two types of arrivals can occur, real-time voice packets (which consists of both audio and video packets) and nonreal-time data packets. Queuing approach is one of the vital mechanisms in traffic management system. For this reason, it is important to implement a queuing discipline that governs the buffering mechanism of voice packets and data packets while they are waiting to be transmitted. Since delay in VoIP technology is a very unpleasant issue, voice packets prioritization must be ensured. Hence, we present here a queuing model with non-preemptive prioritization technique. The numerical analysis of the proposed queueing model is then performed using GSPN modeling technique.

3 Queueing model for end-to-end VoIP connection delay

We begin by describing the queuing model for the end-to-end delay in a VoIP network. We consider two types of arrivals in a VoIP network, voice packets and data packets. It is proven that in certain circumstances the arrival process of voice traffic as well as the data traffic can be modeled by a Poisson process (Voznak and Hromek 2008). We, therefore, assume that voice packets and data packets arrive independently according to Poisson processes with rates λ_1 and λ_2, respectively. Let $\lambda = \lambda_1 + \lambda_2$ be the total arrival rate. Let us designate type I customers to voice packets and type II customers to data packets. Now voice packets are necessary to be processed in preference of data packets to provide a better QoS. Consequently, we consider priority queuing mechanism giving higher priority to voice packets over data packets. Also, as mentioned earlier, the end-to-end delay in a VoIP network consists of a

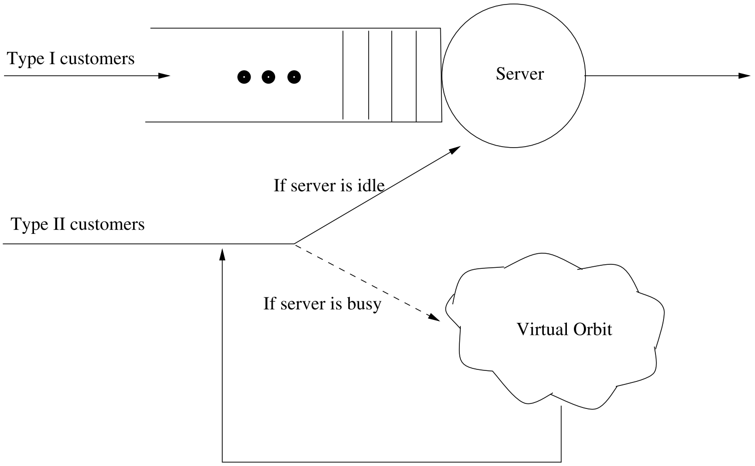


Fig. 2 Single server retrial queue with two types of customers

fixed part and a variable part. As a result, the total time spend in the system by voice packets or data packets can have any general distribution. Therefore, it can be considered that the service times of both types of customers follow different general distributions. However, for modeling simplification, we assume that the service times are independent and identically distributed and have the same distribution for both types of customers. We also assume that both the types of customers cannot leave the system without getting served.

In the following subsections, we present the queueing model to depict the end-to-end delay in a VoIP network, and its numerical analysis.

3.1 Infinite capacity M/G/1 retrial queuing system with two types of customers and non-preemptive priority

We consider the scenario where a voice packet can wait only for either a data packet that is already in service, or for other voice packets ahead of it (cisco.com 2006). That is, a non-preemptive priority queuing mechanism is applied to the incoming voice and data traffic. Moreover, the data packets, if on arrival find the server busy, can be buffered from where a reattempt is made after a random amount of time seeking service. Hence to model the above mentioned situation, we consider an M/G/1 priority retrial queuing system with two types of customers, voice packets (type I) and data packets (type II). Type I customers have a higher priority over type II customers. If a type II customer finds the server idle upon arrival, it immediately goes for service. On the other hand, if a type II customer finds the server busy upon arrival, it enters a virtual orbit with the intention of looking for service again after a random amount of time. The retrial time (the time interval between two consecutive attempts made by a customer in the virtual orbit) is exponentially distributed with mean $1/\alpha$, and is independent of all previous retrial times and all other stochastic processes in the system. Type I customers are queued in a priority queue of infinite capacity after blocking. As soon as the server is free, one of the customers, if any, in the priority queue is served. Therefore, the customers in the virtual orbit will be served only when there are no customers in the priority queue. The above discussed queuing model is depicted in Fig. 2. Note that the analysis of the proposed queueing model is performed via a GSPN formulation which is presented in the next subsection.

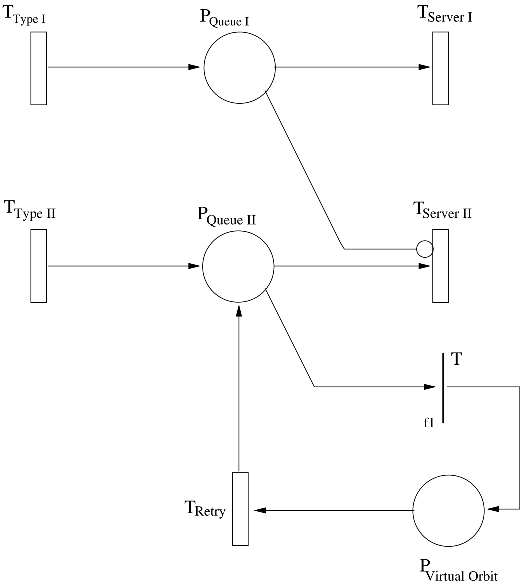


Fig. 3 GSPN model for the proposed queueing model

3.2 GSPN formulation for the queueing model

Generalized stochastic Petri nets (GSPN) (Trivedi 2001; Chiola et al. 1993; Jayaparvathy et al. 2007) are performance analysis tools based on the graphical system representation, and are extensions to Petri nets (PN). Unlike PNs, in GSPNs some transitions are timed, while others are immediate. With timed transitions random firing delays are associated which are exponentially distributed, whereas the firing of immediate transitions takes place in zero time, with priority over timed transitions. The GSPN formulation for the proposed infinite capacity retrial queuing system with two types of customers and non-preemptive priority is presented in Fig. 3. For a brief introduction on GSPN modeling, readers can refer to Trivedi (2001). Note that the GSPN model assumes exponentially distributed firing times for all the timed transitions, even though the service times of both type I and type II customers follow some general distribution. This assumption is made to facilitate GSPN modeling. However, to model the proposed non Markovian M/G/1 retrial queuing system using GSPN modeling, we pursue the following approach. The corresponding GSPN formulation provides the mean end-to-end delay suffered by the first packet (i.e., the packet at the head of line (HOL)). To compute the mean end-to-end delay of the subsequent packets, we model each station as an M/G/l queue, with the mean service time to be the mean delay suffered by the HOL packet (Jayaparvathy et al. 2007).

As shown in the figure, the transitions T_{TypeI} and T_{TypeII} represent the arrival of voice packets and data packets, respectively. When transition T_{TypeI} (T_{TypeII}) fires, one token is deposited in the place P_{QueueI} ($P_{QueueII}$). The mean firing time of T_{TypeI} is the mean inter arrival time of voice packets (i.e., $1/\lambda_1$). Similarly, T_{TypeII} has a mean firing time $1/\lambda_2$). The arriving voice packets form a priority queue, and are buffered in the place P_{QueueI}. As long as there are tokens in P_{QueueI}, the transition $T_{ServerI}$ is enabled. The mean firing time of $T_{ServerI}$ is the mean service time of voice packets (which is taken to be exponential for modeling simplification for the HoL packet).

Tokens in the place $P_{QueueII}$ represents the arriving data packets. The arriving data packets are however not buffered for service. As mentioned earlier, if on arrival of a data packet the server is idle, then it is served, otherwise it goes into a virtual orbit. When the place P_{QueueI} is empty, it indicates that the server is idle. Hence, when there is a token in $P_{QueueII}$ but at the same time there is no token in P_{QueueI} (this is taken care by the inhibitor arc from P_{QueueI} to $T_{ServerII}$ in the GSPN), then the transition $T_{ServerII}$ is enabled. The firing of $T_{serverII}$ indicates that a data packet is served. On the other hand, when there is a token in P_{QueueI}, it indicates that the server is busy, and when the server is busy, the arriving data packet goes into the virtual orbit. Hence, when there is a token at both the places P_{QueueI} and $P_{QueueII}$, then the immediate transition T is enabled. This is taken care of by a guard function $f1$ on the immediate transition T which restricts its firing. Because of the guard function $f1$, T will fire only if there is at least one token in P_{QueueI}. The firing of T deposits a token in the place $P_{VirtualOrbit}$. It is to be noted that there is a single server in the system. However, for modeling simplification, we have used two timed transitions $T_{ServerI}$ and $T_{ServerII}$ to represent the service times of voice packets and data packets, respectively, though they have same service time distribution with same parameter value.

Now from the virtual orbit, the data packets keep on retrying for service after an exponential time. This is taken care by the timed transition T_{Retry} which has a mean time of $1/\alpha$. The firing of T_{Retry} deposits a token back to $P_{QueueII}$ from where the data packets retry for service. And the process continues. In the next subsection, we will see that how the end-to-end delay is numerically obtained.

3.3 Numerical analysis of end-to-end delay

The GSPN formulation discussed in the previous subsection provides the mean end-to-end delay (or the mean delay) suffered by the packet at the HoL. To compute the mean end-to-end delay of the subsequent packets, we model each station as an M/G/1 queue, with the mean service time to be the mean delay suffered by the HoL packet. We get the mean delay at the GSPN level as follows: The mean delay of the HoL packet at each station, $\bar{D}_{HoL}$, is the sum of the mean packet holding time and the mean service time undergone by the HoL packet. This can be obtained as follows:

For a place P and for a transition T, let us donate $\sharp(P)$ as the average number of tokens in the place P and η_T as the average throughput of the transition T, which is defined as the average rate at which tokens are deposited by the transition T in its output places. Following these notations, $\bar{D}_{HoL}$ of voice packets is given by

$$\bar{D}_{HoL(voice)} = \frac{\sharp(P_{QueueI})}{\eta_{T_{TypeI}}} + \frac{1}{\mu} \tag{1}$$

where $\frac{1}{\mu}$ is the mean service time undergone by the HoL packet.

Similarly, $\bar{D}_{HoL}$ of data packets is given by

$$\bar{D}_{HoL(data)} = \frac{\sharp(P_{QueueII})}{\eta_{T_{TypeII}}} + \frac{\sharp(P_{VirtualOrbit})}{\eta_{T_{Retry}}} + \frac{1}{\mu}. \tag{2}$$

The rest of the buffer is modeled as an M/G/1 queue with mean service time to be $\bar{D}_{HoL}$. The mean voice packet delay, $\bar{D}_{voice}$ can then be obtained by applying the *Pollackzek-Kinchine* mean value formula (Castaneda et al. 2012) as

$$\bar{D}_{voice} = \bar{D}_{HoL(voice)}\left[1 + \frac{\rho}{2(1-\rho)}\left(1 + C^2_{D_{voice}}\right)\right] \tag{3}$$

where $\rho = \lambda\bar{D}_{HoL(voice)}$. If the delay of the HoL voice packet is represented by the random variable D_{voice}, then

$$C^2_{D_{voice}} = \frac{E(D^2_{voice})}{\bar{D}^2_{HoL(voice)}} \tag{4}$$

where

$$E\left(D^2_{voice}\right) = 2\left(\frac{\sharp(P_{QueueI})}{\eta_{T_{TypeI}}}\right)^2.$$

Similarly, mean data packet delay $\bar{D}_{data}$ can be obtained as

$$\bar{D}_{data} = \bar{D}_{HoL(data)}\left[1 + \frac{\rho}{2(1-\rho)}\left(1 + C^2_{D_{data}}\right)\right] \tag{5}$$

where $\rho = \lambda\bar{D}_{HoL(data)}$. If the delay of the HoL data packet is represented by the random variable D_{data}, then

$$C^2_{D_{data}} = \frac{E(D^2_{data})}{\bar{D}^2_{HoL(data)}} \tag{6}$$

where

$$E\left(D^2_{data}\right) = 2\left(\frac{\sharp(P_{QueueI})}{\eta_{T_{TypeI}}}\right)^2.$$

4 Numerical illustration and observations

In this section, we present numerical illustration of the average end-to-end VoIP connection delay obtained from the proposed queuing model. We assume that the service time of both the type of customers follow deterministic distribution (D) with mean service time equal to a unit time. Ensuring the stability of the system, we assume the following parameter values for the purpose of numerical illustration: retrial rate of type II customer $\alpha = 0.3$; λ_1 varies from 0.01 to 0.08; and λ_2 varies from 0.025 to 0.200. To get the numerical results from the GSPN model corresponding to the proposed queuing model, we make use of the software package SHARPE (Sahner et al. 1996).

Using these parameters values, we obtain the following graphical results. Figure 4 plots the average number of type I customers in the system for varying values of λ_1. It is observed that the average number of type I customers increases with the increasing arrival rate, as expected. Figure 5 plots the average end-to-end delay of type I customers for varying values of λ_1. It is observed from the above graph that the number of type I customers increases with increasing arrival rate, and thus the average end-to-end delay of the type I customers also increases. Figure 6 plots the average number of type II customers in the system for varying values of λ_2. It can be seen that it exhibits the same behavior, i.e., the average number of type II customers increases with the increasing arrival rate. Consequently, with the increase in the arrival rate of type II customers, the average end-to-end delay of the same also increases. This is shown in Fig. 7.

Fig. 4 Average number of voice packets vs arrival rate of voice packets (λ_1)

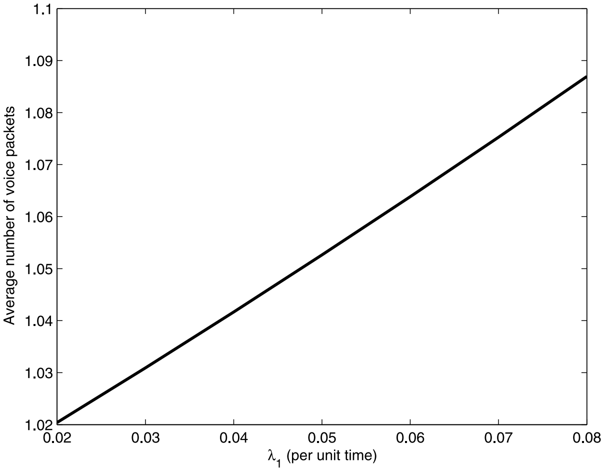


Fig. 5 Average delay suffered by voice packets vs arrival rate of voice packets (λ_1)

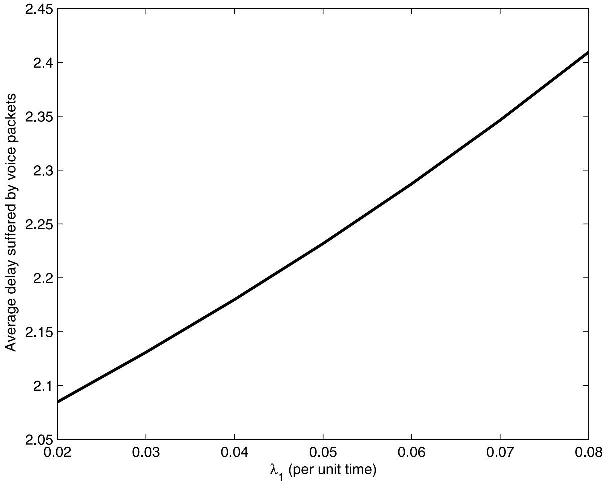
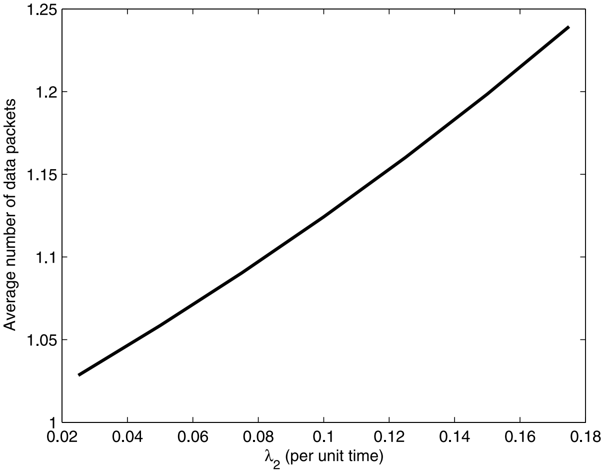


Fig. 6 Average number of data packets vs arrival rate of data packets (λ_2)



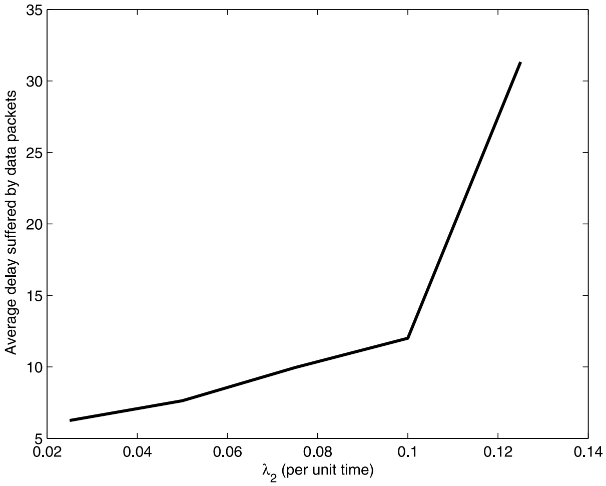


Fig. 7 Average delay suffered by data packets vs arrival rate of data packets (λ_2)

5 Conclusion

In this paper, we present a queuing model for the end-to-end delay of a VoIP connection. This queuing model is suitable for the approximation of voice traffic and data traffic from sources with Poisson probability distribution. All the partial delay components in a VoIP network are explained. Thereafter we present a non-Markovian M/G/1 queuing model to analyze the end-to-end VoIP connection delay. The main contribution of this paper is that we make use of an equivalent GSPN model to get the analytical results for the proposed queueing model.

References

Baronak, I., & Halas, M. (2007). Mathematical representation of VoIP connection delay. *Radioengineering*, *16*, 77–85.

Castaneda, L. B., Arunachalam, V., & Dharmaraja, S. (2012). *Introduction to probability and stochastic processes with applications*. New Jersey: Wiley.

Chiola, G., Marsan, M. A., Balbo, G., & Conte, G. (1993). Generalized stochastic Petri nets: a definition at the net level and its implications. *IEEE Transaction on Software Engineering*, *19*(2), 89–107.

http://www.cisco.com/en/US/tech/tk652/tk698/technologies_white_paper09186a00800a8993.shtml.

Jayaparvathy, R., Anand, S., Dharmaraja, S., & Srikanth, S. (2007). Performance analysis of IEEE 802.11 DCF with stochastic reward nets. *International Journal of Communication Systems*, *20*(3), 273–296.

Karapantazis, S., & Pavlidou, F. N. (2009). VoIP: a comprehensive survey on a promising technology. *Computer Networks*, *53*, 2050–2090.

Rashed, M. M. G., & Kabir, M. (2010). A comparative study of different queuing techniques in VoIP, video conferencing and file transfer. *Dafodil International University Journal of Science and technology*, *5*(1), 37–47.

Rezac, F., Voznak, M., & Hromek, F. (2010). Delay variation model with two service queues. *Information and Communication Technologies and Services*, *8*(1), 24–29.

Sahner, R. A., Trivedi, K. S., & Puliafito, A. (1996). *Performance and reliability analysis of computer systems: an example based approach using the SHARPE software package*. Massachusetts: Kluwer Academic.

Shim, C., Xie, L., Zhang, B., & Sloane, C. J. (2003). *How delay and packet loss impact voice quality in VoIP*. White paper.

Trivedi, K. S. (2001). *Probability and statistics with reliability, queueing, and computer science applications* (2nd ed.). New York: Wiley.

Voznak, M., & Hromek, F. (2008). *Analytic model of a delay variation valid for the RTP*. *Networking studies II: selected technical reports* (pp. 103–113). Praha: CESNET. ISBN: 978-80254-2151-2