DELAY VARIATION MODEL WITH TWO SERVICE QUEUES

Filip REZAC[1], *Miroslav VOZNAK*[1], *Frantisek HROMEK*[1]

[1]Department of Telecommunications, Faculty of Electrical Engineering and Computer Science, VSB – Technical University of Ostrava, 17. listopadu 15, 708 33 Ostrava, Czech Republic

filip.rezac@vsb.cz, miroslav.voznak@vsb.cz

Abstract. *Delay in VoIP technology is very unpleasant issue and therefore a voice packets prioritization must be ensured. To maintain the high call quality a maximum information delivery time from the sender to the recipient is set to 150 ms. This paper focuses on the design of a mathematical model of end-to-end delay of a VoIP connection, in particular on a delay variation. It describes all partial delay components and mechanisms, their generation, facilities and mathematical formulations. A new approach to the delay variation model is presented and its validation has been done by experimention.*

Keywords

VoIP, delay, jitter, M/D/2, queue, QoS.

1. Introduction

As we said before a delay is one of the main issues in packet-based networks which have impact on QoS. There are several components of the delay in an IP network which differ from each other in the way they originate. Each delay component has different impact on a voice packet delay. Delay components can be divided according to their origin [1], [2], [8], [11]:

- coder delay and Packetization delay in transmitter,
- queuing delay, Serialization delay and Propagation delay in transmission network,
- de-jitter delay, De-packetization delay and Decompression delay in receiver.

2. Delay Variation Model with Two Service Queues

Based on the analysis of the principle of service models with two priority queues, we can assume that delays in a higher priority service queue are shorter. If we opt for better system resources for the high-priority service queue, we are likely to experience poorer system resources for voice packets in the lower-priority service queue.

Delays are likely to get longer in a lower priority queue, see [3] and [12].

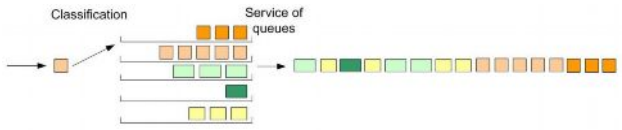

Fig. 1: Service model with two priority queues.

Even if service parameters get weaker, standard users are not likely to perceive the poorer quality provided the longer delay is offset by the size of the De-jitter buffer or where, given the imperfection of human hearing, the total length of the delay has a negligible impact. In order to be able to express the delay in the service element with priority queues, it is necessary to monitor solely queues that are used for the voice flow service.The method applied to transmit voice packets in priority queues corresponds with the *M/D/n/k* [9] model where *n* is the number of service queues and *k* is the size of the cache memory [4], [5].

In order to express the mathematical model which uses two service queues for transmission of voice streams, we can substitute the *M/D/n/k* model by the *M/D/2/k* model. In order to express the model we disregard the size of cache memory. This assumption enables us to replace the *M/D/2/k* model by the *M/D/2*model.The conditions for validating the *M/D/2* model are as follows:

- no interruption of the priority service process: Packets in the higher priority queue are served

before packets in the lower priority queue. When a packet with some priority arrives, the service is provided first,

- priority queues are served based on the FIFO (First In First Out) method,
- the arrival process corresponds to the Poisson distribution. Where every single stream matches the Poisson distribution, then the sum of such streams also matches the Poisson distribution,
- the service rate is a constant because the same codec and packets of the samesize are used,
- the arrival rate is also a constant since we assume a constant number of the flows with the same codec,
- the size of the priority queue's cache memory is infinite.

The system's utilisation can by express by the following formula:

$$\rho = \frac{\lambda}{\mu}, \tag{1}$$

where:

- ρ – system utilisation [-],
- λ – arrival rate [s^{-1}],
- μ – service rate [s^{-1}].

The stability condition $0 \leq \rho < 1$ needs to apply. The utilisation of a system with two priority queues can be express as follows:

$$\rho = \rho_1 + \rho_2, \tag{2}$$

where ρ i is the utilisation of the system queue. The system utilisation can be express as follows:

$$\rho = \frac{\lambda_1 + \lambda_2}{\mu}, \tag{3}$$

The arrival rate can be expressed by the following equation:

$$\lambda_i = M_i \frac{C_{BW}}{P_S}, \tag{4}$$

where:

- M_i – number of streams in queue i [-],
- C_{BW} – codec bandwidth [b/s],
- P_S – payload size [b].

The service rate can by express by the following equation:

$$\mu = \frac{1}{T_{SER} + T_S}, \tag{5}$$

where:

- T_{SER} – serialization delay [s],
- T_S – processing time [s].

The mean service time of the process in a higher priority queue can be express as follows, see [10] and [12]:

$$\bar{T}_1 = \frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho_1)}. \tag{6}$$

Similarly, the mean service time of the process in a lower priority queue can be express as follows:

$$\bar{T}_2 = \frac{1}{\mu} + \frac{\rho}{2\mu(1-\rho)(1-\rho_1)}. \tag{7}$$

The relation between the mean service time and the system utilisation of the queue is shown in the Fig. 2.

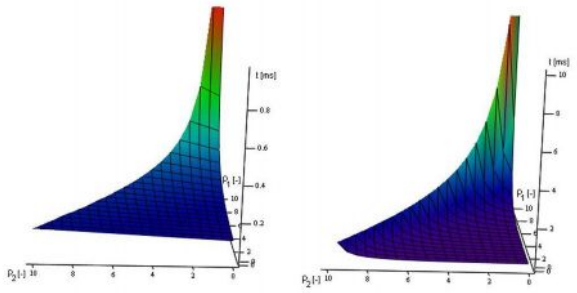

Fig. 2: Relation between the mean service time in a higher and lower priority queue and system utilisation of the queue.

In the system "without interruption" the mean service time is generally the sum of the service time, time of the remaining services, time it takes voice packets included in the same or a higher-priority queue to be transmitted and time needed to transmit voice packets of a higher priority that came while the packet was waiting to be processed by the system, see [6], [7] and [10].

A key parameter is the time it takes to process a service element. This parameter needs to be determined individually for each service element. It is determined by hardware (processor, motherboard and network card, etc.) and software (operating system, kernel, etc.) used. The only option to determine the processing time is based on knowledge of the behaviour characteristic of the element in the increasing load, see [8] and [12].

Assuming we know both the line speed and the processing time, we can express the service rate by the following equation:

$$\mu = \frac{L_S}{P_S + H_L + L_S T_S}, \tag{8}$$

where:

- H_L – header length [b],

- L_S – line speed [b/s].

The utilisation of the system queue can by express as follows:

$$\rho_i = \frac{M_i C_{BW}(P_S + H_S + L_S T_S)}{P_S L_S}. \quad (9)$$

The system utilisation can be expressed as follows:

$$\rho = \frac{C_{BW}(M_1 + M_2)(P_S + H_S + L_S T_S)}{P_S L_S}. \quad (10)$$

The mean service time of the process in a higher priority queue can be express by the following formula:

$$T_1 = \frac{1}{2}\frac{P_S + H_S + L_S T_S}{L_S} \cdot \frac{2P_S L_S - C_{BW}(M_1 - M_2)(P_S + H_S + L_S T_S)}{P_S L_S - C_{BW} M_1 (P_S + H_S + L_S T_S)}. \quad (11)$$

The mean service time of the process in a lower priority queue can be expressed by the following formula:

$$T_2 = \frac{1}{2}\frac{P_S + H_S + L_S T_S}{L_S} \cdot \frac{2(P_S L_S)^2 - C_{BW} P_S L_S (M + 2M_1)(P_S + H_S + L_S T_S) + }{(P_S L_S)^2 - C_{BW} P_S L_S (M + M_1)(P_S + H_S + L_S T_S) + } \quad (12)$$
$$\frac{+ 2C_{BW}^2 M M_1 (P_S + H_S + L_S T_S)^2}{+ C_{BW}^2 M M_1 (P_S + H_S + L_S T_S)^2}.$$

End-to-end delay can be expressed by substituting the model designed for single service queue. The end-to-end delay in a lower priority queue can by expressas follows, see [8], [11] and [12]:

$$T_{1c} = (1 + 0{,}1M) T_{CD} + \frac{P_S}{C_{BW}} + \frac{1}{v}\sum_{i=1}^{n} L_i + T_{DJD} + \sum_{i=2}^{n} T_{1i}, \quad (13)$$

where:

- T_{1c} – end-to-End delay [s],
- N – number of voice blocks in a packet [-],
- T_{CD} – total delay of the codec [s],
- L_i – length of line i [m],
- v – speed of signal transmission in the environment [m/s],
- T_{DJD} – de-jitter delay [s],
- T_{1i} – mean service time i service element i [s].

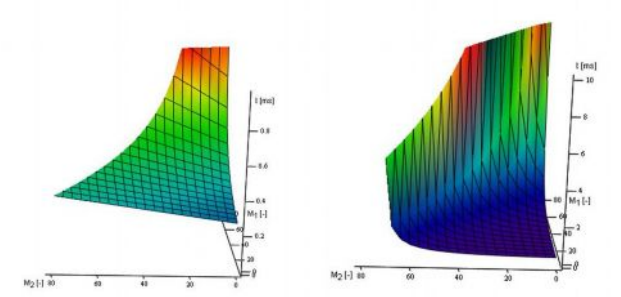

Fig. 3: Relation between the mean service time in a higher and lower priority queue and the number of calls.

End-to-end delay in a higher priority queue can by expressed by the following formula [12]:

$$T_{2c} = (1 + 0{,}1M) T_{CD} + \frac{P_S}{C_{BW}} + \frac{1}{v}\sum_{i=1}^{n} L_i + T_{DJD} + \sum_{i=2}^{n} T_{2i}. \quad (14)$$

3. Experiment

The workplace in which we carried out the estimation of the proposed model consisted of a service element (PC1) with *Traffic Control*, three performance endpoints and a console workstation. VoIP calls were emulated by IxChariot Performance endpoints and the IxChariot Console was used to assess VoIP calls.Experiments were out under different conditions. IxChariot endpoints generate voice streams between PC2 and PC3 and between PC4 and PC3.

Linux distribution *OpenSuse 10.3*, with the implemented support of the QoS was used as the operating system in the core element. Two queues to process voice streams and one queue to process the rest of the traffic have been defined.

The structure of the experimental workplace is illustrated in Fig. 4. The configuration of the *Traffic Control* in the core element is described below. Three priority queues were defined.

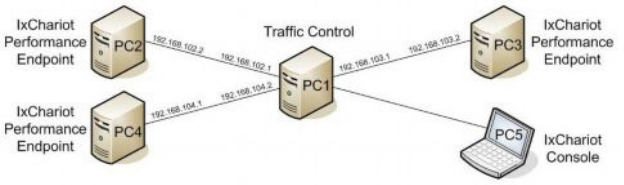

Fig. 4: Experimental testbed.

```
tc qdisc add dev eth1 root handle 1:0 prio
tc filter add dev eth1 parent 1:0 prio 1
protocol ip u32 match ip tos 0x28 0xff flowid
1:1
tc filter add dev eth1 parent 1:0 prio 2
protocol ip u32 match ip tos 0x48 0xff flowid
1:2
tc filter add dev eth1 parent 1:0 prio 3
```

```
protocol ip u32 match ip tos 0x00 0xff flowid
1:3
tc qdisc add dev eth2 root handle 1:0 prio
tc filter add dev eth2 parent 1:0 prio 1
protocol ip u32 match ip tos 0x28 0xff flowid
1:1
tc filter add dev eth2 parent 1:0 prio 2
protocol ip u32 match ip tos 0x48 0xff flowid
1:2
tc filter add dev eth2 parent 1:0 prio 3
protocol ip u32 match ip tos 0x00 0xff flowid
1:3
tc qdisc add dev eth3 root handle 1:0 prio
tc filter add dev eth3 parent 1:0 prio 1
protocol ip u32 match ip tos 0x28 0xff flowid
1:1
tc filter add dev eth3 parent 1:0 prio 2
protocol ip u32 match ip tos 0x48 0xff flowid
1:2
tc filter add dev eth3 parent 1:0 prio 3
protocol ip u32 match ip tos 0x00 0xff flowid
1:3
```

The Network Interface Cards were configured using *Ethtool*. The network address was configured using standard Linux commands. An example NIC configuration for PC2 is described below.

```
ethtool –s eth1 speed 10 duplex full autoneg off
ifconfig eth1 192.168.102.2 netmask
255.255.255.0
route add default gw 192.168.102.1
```

The number of voice streams was the same. TOS 0x28 values were used in voice streams between PC2 and PC3. TOS 0x48 values were used in RTP streams between PC4 and PC3.

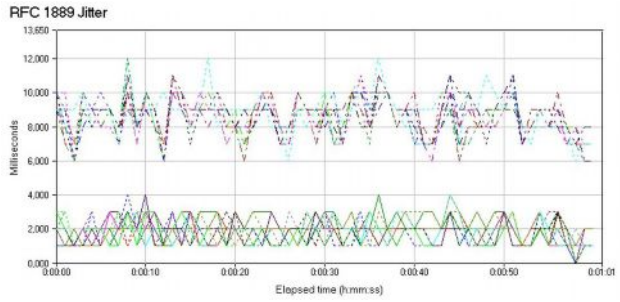

Fig. 5: Example of test results.

Each RTP stream used a different communication port. For your experiment, we used G. 711a and 20 ms as a delay between the datagram.

The relation between the mean service time in a higher and lower priority queue and equally distributed load in the queues is shown in Fig. 6.

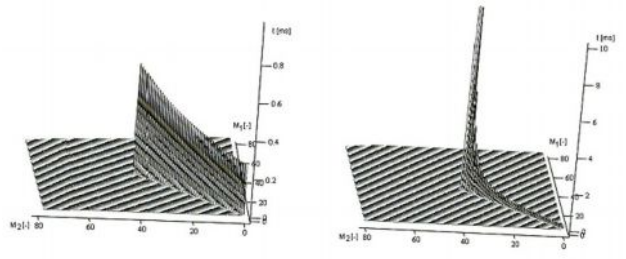

Fig. 6: Relation between the mean service time in a higher and lower priority queue and equally distributed load in the queues.

The mathematical model uses values characteristic for the G.711 codec. The length of transmission lines was set to 50 meters and the De-jitter buffer size was set to 1 ms.

The accuracy of the model depends on what T_S parameter is chosen. In order to compare the conformity of real and theoretical values, T_S = 0,11 ms was applied.

A comparison of theoretical values and the results of the experiment is shown in the diagrams below.

In experimental workplace we have been theoretically able to run approximately 110 calls. In real terms we performed only 75 calls without other influences that we were not able to reflect in the model, such as an unpredictable processing of the call and loss of the information.

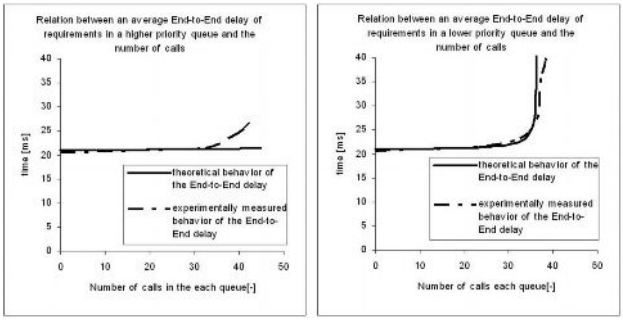

Fig. 7: Relation between an average end-to-end delay of requirements in a higher and lower priority queue and the number of calls.

Relative errors of the model for each of the queues are shown in Fig. 8.

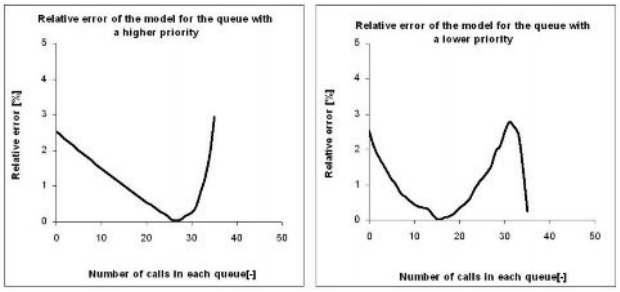

Fig. 8: Relative error of the model for the queue with a higher and lower priority.

4. Conclusion

The proposed mathematical model is suitable for the approximation of voice traffic which consists of sources with the Poisson's probability distribution. However, as the load increases, the mathematical model does not return exact information. The measurements have shown that the mathematical model strongly depends on selection of T_S value. T_S has proved to be significant between 50 % and 70 % of the line load, due of the emergence of the processing delay. Because of the use of the processing time in the mathematical model, we are able to get data with accuracy below ± 3 % up to the 70 % of line load. Furthermore (over 70 % of line load), the tests did not reproduce due to the unpredictable behaviour of call processing and loss.

T_S is a key parameter and it marks a time it takes to process a service element. This parameter needs to be determined individually for each service element. It is determined by hardware (processor, motherboard and network card, etc.) and software (operating system, kernel, etc.) used. The only option to determine the processing time is based on knowledge of the behaviour characteristic of the element in the increasing load.

Up to the 70 % of line load, the maximum deviation between the theoretical model and real values was 0,75 ms. The delay incurred in the queuing element with delay below 1ms can not be considered as sufficiently precise since the absolute measurement error of the method using IxChariot equals 1ms. As regards the end-to-end delay, the relative error measured during the experiment is less than 3 % when compared to the theoretical values obtained through the application of the mathematical model.

Even though individual voice connections do not match the model of a signal source with the Poisson's probability distribution, the sum of a greater number of voice connections returns average values that are closer to values returned by the proposed model. If we apply this model to describe VoIP networks that process a greater number of simultaneous voice connections, we can assume that the proposed model will return sufficiently exact assessment of an average delay in the network.

Acknowledgements

This project has been supported by a research intent "Optical Network of National Research and Its New Applications" (MSM 6383917201). The report includes the result achieved by authors in a research activity Multimedia transmissions and collaborative environment, http://www.ces.net/project/15/.

References

[1] VOZŇÁK, M. Voice over IP and Jitter Avoidance on Low Speed Links. In: *Proceedings RTT2002*. Žilina, 2002. ISBN 80-227-1934-X.

[2] HARDY, W. *VoIP Service Quality.* New York: McGraw-Hill, 2003. ISBN 0-07-141076-7.

[3] HALÁS, M.; KYRBASHOV, B.; VOZŇÁK, M. Factors influencing voicequality in VoIP technology. In: *9th International Conference on Informatics*. Bratislava, 2007, p. 32-35. ISBN 978-80-969243-7-0.

[4] VOZŇÁK, M.; HROMEK, F. Optimalization of VoIP service queues. In: *Proceedings RTT2008, Vyhne, Slovakia*. 2008. Bratislava: Slovak University of Technology. ISBN 978-80-227-2939-0.

[5] GROSS, D.; HARRIS, C. *Fundamentals of Queuing Theory.* New York: JohnWiley & Sons, 1998. ISBN 0-471-17083-6.

[6] BERGIDA, S.; SHAVITT, Y. Analysis of shared memory priority queues with two discard levels. In: *IEEE Israel Conference*, 2006, p. 42-46. ISBN 1-4244-0230-1.

[7] MARIANOV, V.; SERRA, D. *Location Models for Airline Hubs Behaving as M/D/c Queues*. Available at WWW: <http://dx.doi.org/10.1.1.92.4247>.

[8] VOZŇÁK, M.; HROMEK, F. *Analytic model of a delay variation valid for the RTP*. In LHOTKA, L.; SATRAPA, P. *Networking Studies II: Selected Technical Reports*. Praha: CESNET, 2008, p.103-113. ISBN 978-80 254-2151-2. Available online.

[9] HAMPL, P. *Kendallova klasifikace obsluhových systémů*. Access Server, 2005. ISSN 1214-9675. Available at WWW: <http://access.feld.cvut.cz>.

[10] ZÍTEK, F. *Ztracený čas: Elementy teorie hromadné obsluhy*. Praha: Academia, 1969.

[11] HALÁS, M. *Optimalizácia hlasovej prevádzky s ohľadom na kvalitu hovoru v sieťach s technológiou VoIP*. Doctoral Thesis, 2006.

[12] HROMEK, F. *Optimalizace obsluhy RTP front*. Doctoral Thesis, Ostrava, 2008.

About Authors

Filip REZAC was born in 1985. In 2007, he received a Bachelor title in VSB-TU Ostrava, Faculty of Electronics and Computer Science, Department of Informatics. Two years later he received the MoS title focused on mobile technology in the same workplace. Currently in the doctoral study he focuses on Voice over IP technology, Network Security and Call Quality in VoIP.

Miroslav VOZNAK was born in 1971. He has been studying Telecommunications engineering at VSB-TU Ostrava and he hold a doctorate in Telecommunications. He received Ph.D. degree in 2002 at the Faculty of Electrical Engineering and Computer Science where he works as an associate professor nowadays. He gives the lectures at the Department of Telecommunications of VSB – TU in Ostrava.