Analytic model of a delay variation valid for the RTP

Technical Report 16/2007

Miroslav Vozňák – František Hromek

CESNET z.s.p.o., Zikova 4, Praha, Czech Republic

miroslav.voznak@vsb.cz

29.11.2007

Keywords : VoIP, RTP, delay, jitter, Poisson, M/D/1/k

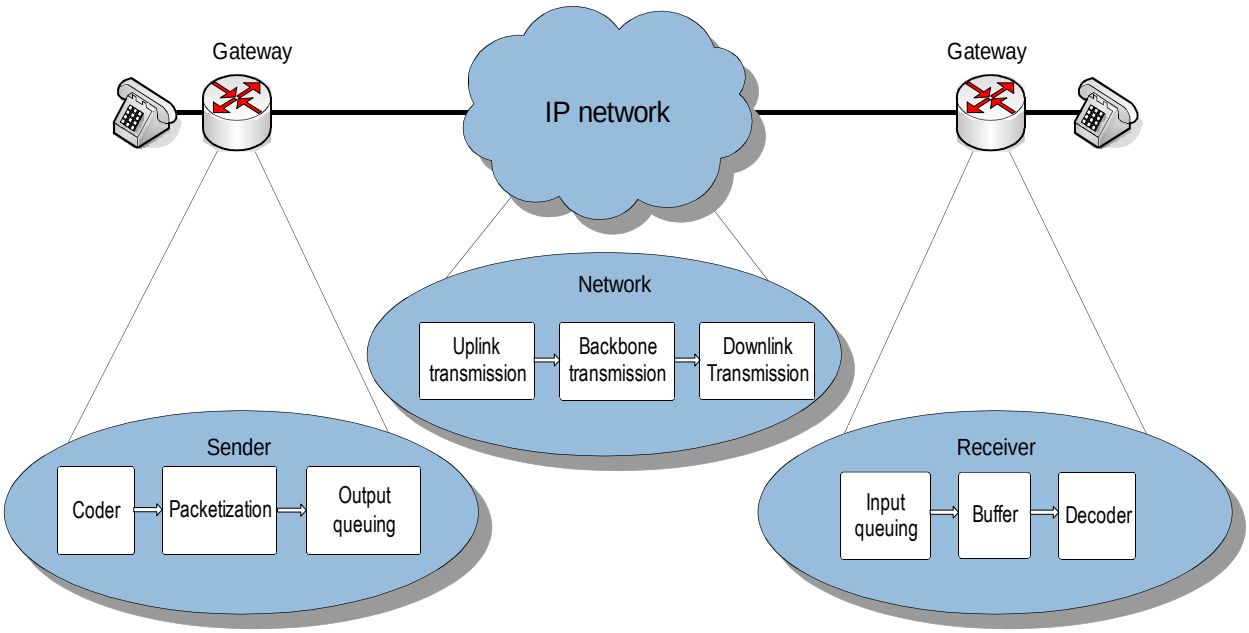
Abstract

This technical report focuses on the design of a mathematical model of end-to-end delay VoIP connection, in particular on a delay variation. It describes all partial delay components and mechanisms, its generation, facilities and its mathematical formulations. A new approach to the delay variation model is presented; its validation has been done by an experiment.

1 Introduction

A delay is one of the main issues in packet-based networks and as such, it poses one of the major threats to QoS mechanisms. The delay can have various causes including propagation, handling or processing. There are several types of delays in an IP network which differ from each other as to where they are created, mechanism of their creation or some other attributes. Each delay component influences the result voice packet delay in a different way. This paper provides a detailed description of individual delay components, and explains mechanisms of their creation. Subsequently, it focuses on the creation of a mathematical model of a VoIP end-to-end delay in the network. The delay components should be classified based on the place of their creation:

- 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.



- Fig. 1. Delay components.

2 Delay components

We can find two types of delay in the transmitter. The first is a *coder delay* and is affected by 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, e.g. for the ITU-T G.711 (PCM) codec it is 0.125 ms frame size without look-ahead and for the ITU-T G.729 (CS-ACELP (codec) it's the frame size value 10 ms and 5 ms look-ahead. The second type of delay in the transmitter is the *packetization delay.* This delay occurs when data blocks are encapsulated into packets and transmitted by the network. The packetization delay is set as multiples of the packetization period with which the particular codec operates and specifies how many data blocks are transmitted in one packet [1], [2].

The estimation process is given by the reference (1):

$$T_{PD} = \frac{P_S}{C_{BW}} \quad [ms] \tag{1}$$

Where :

T_{PD} – packetization delay [ms]

P_S – payload size [b]

C_{BW} – codec bandwidth [kbit/s]

We can incur three types of delays in the receiver. The first type is the *de-jitter delay* which is closely related to the variable delay in the network when it is necessary to eliminate a variance of these variable components using supplementary buffer store in the receiver, this buffer is called a *playout* buffer. Its size is typically adjusted as a multiple of the packetization delay because of an optimization. If this value is adjusted statistically, then jitter buffer sizes are about 30-90 ms, a typical value is 60 ms. If the variable playout buffer is used, the size is adapted based on the real-time delay variation. In this case the typical maximum value is about 150 ms [3].

The second type is a *depacketization delay.* Its mechanism is very similar to that of the packetization delay mentioned above. The depacketization is a reverse packetization and therefore the size of depacketization delay of one block in the frame is in correlation with its packetization delay. In a real traffic the delay of each block within the frame of one packet occurs, always only for the value of the packetization delay. This is why we count with only one constant packetization delay value.

The third type is a *decompression delay*. The decompression delay, similarly to the coder delay depends on the compressing algorithm selection. On average, the decompression delay is approximately 10 % of the compressing codec delay for each voice block in the packet. But it is very dependent on the computing decoder operation and mainly on the number of voice blocks in one packet. This decompression delay might be defined by the following formula:

$$T_{DCD} = 0{,}1.N.T_{CD} \quad [ms] \tag{2}$$

Where:

T_{DCD} – decompression delay [ms]

N – number of the voice blocks in the packet

T_{CD} – coder delay [ms]

The last component of our classification is a delay in the transmission network. Again, there are three types of this delay. The first one depends on the transmission rate of the used interface and it is called as a *serialization delay*. The packet sending takes some time. This time depends on the transmission medium rate and on the size of packet. Relation (3) shows estimation of the time:

$$T_{SER} = \frac{P_S + H_L}{L_S} \quad [ms] \tag{3}$$

Where :

T_{SER} – serialization delay [ms]

L_S – line speed [kbit/s]

H_L – header length [b]

The second type of a delay originated in the transmission network is the *propagation delay*. This delay relates to the signal transmission, i.e. to its physical regularities of the propagation in the surroundings. This delay type depends on the used transmission technology, in particular on the distance over which the signal is transmitted. Today's networks are mostly built on single mode optical fibers. The light rate of spread in optical fiber is $v=2.07.\ 10^{-8}$ [ms^{-1}], from which the propagation delay should be defined using following formula (4).

$$T_{Prop} = \frac{L}{v} \quad [ms] \tag{4}$$

Where:

T_{Prop} – propagation delay [ms]

L – line length [km]

v – light rate of spread in optical fiber = $2.07\ .\ 10^{-8}$ [ms^{-1}]

The last type is the delay which occurs in active elements of the transmission network and relates to handling of RTP packets, in particular in the router queues. This delay is the most significant part of the jitter. A delay variation or a jitter is a metric that describes the level of disturbance of packet arrival times compared to the ideal arrival time. Such disturbances can be caused by queuing or by processing [1], [4].

3 Delay variation model

To describe queuing delay mechanisms for VoIP traffic is not simple. This topic is discussed in many publications and queuing theory provides solution to many issues. It involves mathematical analysis of processes including arrival at the input of a queue, waiting in the queue and the serving at the front of the queue and providing the appropriate performance parameters of the designed model.

It is proven that in certain circumstances the voice traffic can be modelled by a source signal the probabilistic random variable distribution of which matches Poisson's probability distribution. We can usually trace an influence of a jitter in the routers equipped with low-speed links. These routers often operate with PQ optimization (Priority Queuing). Priority queuing is mainly used for serving the voice flow and is based on a preferred packet sorting so that the selected packets are placed into priority queue [1].

Have a look at the figure 2 which shows several FIFO queues (at least two are necessary). Each queue has been assigned a different priority, there is a classifier making the decision in which of the queues to place the packet and a scheduler picking the packets starting with the higher priority queue, next with lower priority etc. Any packets in the high priority queue must be served first. When the queue is empty, the queue with lower priority can be served.

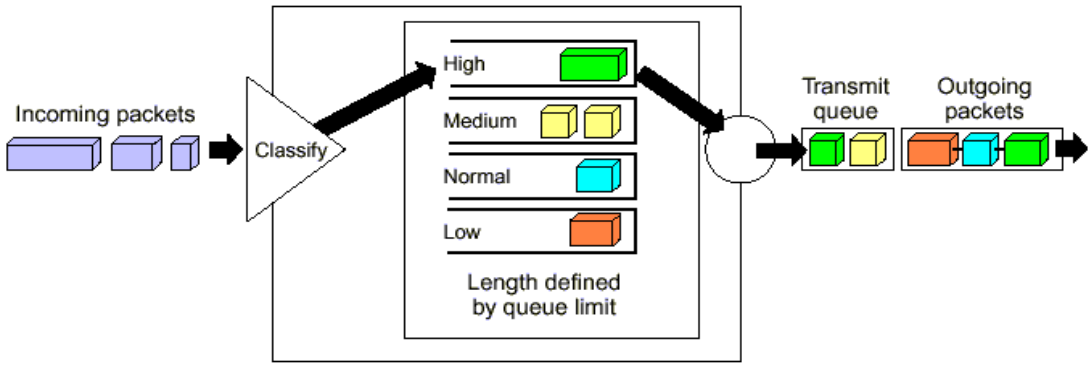


Fig. 2. Priority queuing.

If there is an effectively utilized packet fragmentation mechanism on the output of the line, it is possible to mitigate the influence of the serialization delay in data packets with a lower priority than that of the voice packets. In this case, for the modelling requirements of traffic loading and delay in router, it is sufficient to watch a delay only in the priority queue. Servicing requirement technique in the priority queue corresponds to the model of queuing system *M/D/1/k*, where k is size of buffer. The model notation used corresponds with Kendall's notation of queuing models [5].

In order to create an analytical model of the switching delay we can ignore the buffer size and count with a system of sufficient buffer size in which the loss of preferred packets doesn't occur. If this *M/D/1/k* model can be replaced by *M/D/1/∞* model, we are able to create an analytical expression of switch buffer store seizing. Consequently it is easier to gain an analytical model of the delay in the queue.

The conditions for validating the designed model:

- the arrival process is a Poisson process with an exponentially distributed random variable, we consider that every source of a stream corresponds to the Poisson distribution and therefore their sum also corresponds to Poisson distribution [4],
- $\lambda(t)$ is an arrival rate and this rate is a constant λ, it means we assume that only one type of the codec is used and there are M-sources,
- a service process in priority queue is FIFO (First In First Out),
- μ is a service rate and it is a constant because the same codec is used,
- we assume that the number of waiting positions in a priority queue is infinite.

We express the utilization of the system in equation (5) and for stability must be valid $0 \leq \rho < 1$:

$$\rho = \frac{\lambda}{\mu} \tag{5}$$

Where:

λ is an arrival rate [s^{-1}]

μ is a service rate [s^{-1}]

ρ is a system utilization

We can express the arrival rate by the following equation:

$$\lambda = \frac{C_{BW}}{P_S} \quad [s^{-1}] \tag{7}$$

and the service rate by the equation (8) below:

$$\mu = \frac{1}{T_{SER} + T_S} \quad [s^{-1}] \tag{8}$$

Where:

T_{SER} – serialization delay [s]

T_S – processing time (handling by processor) [s]

The probability that k-attempts will be waiting in the queue is:

$$p_k = (1-\rho)\sum_{j=1}^{k}(-1)^{k-j}(j\rho)^{k-j-1}\frac{(j\rho + k - j)e^{j\rho}}{(k-j)!} \quad \text{for } k \geq 2 \tag{9}$$

$$p_k = (1-\rho)(e^{\rho} - 1) \quad \text{for } k = 1 \tag{10}$$

$$p_k = (1-\rho) \quad \text{for } k = 0 \tag{11}$$

Equation (12) determines T [s] as a mean time which a request spends in the system and $\frac{1}{\mu}$ is the expected service time of one request.

$$T = \frac{1}{\mu} + \frac{\rho}{2(1-\rho)\mu} \quad [s] \tag{12}$$

N in equation (13) stands for the mean number in the system

$$N = T.\lambda \tag{13}$$

We assume that there are M sources with Poisson distribution of inter-arrival times and that all RTP streams use the same codec. Then we can express the arrival rate as follows:

$$\lambda = M \cdot \frac{C_{BW}}{P_S} \quad [s^{-1}] \tag{14}$$

We know the transmission speed of the low-speed link and subsequently we can derive the equation for the calculation of the service rate in the system. We apply the relations (3) to the relation (8) and we obtain the following result:

$$\mu = \frac{L_S}{P_S + H_L + L_S.T_s} \quad [s^{-1}] \tag{15}$$

We apply the relations (14) and (15) to (6) and we obtain the following equation for the system utilization:

$$\rho = \frac{M.C_{BW}.(P_S + H_L + L_S.T_S)}{P_S.L_S} \qquad (16)$$

Equation (17) derived from equations (14), (15), (16) and (12) above expresses the mean service time

$$T = \frac{1}{2}.\frac{P_S + H_L + L_S.T_S}{L_S}.\frac{2.P_S.L_S - C_{BW}.M\left(P_S + H_L + L_S.T_S\right)}{P_S.L_S - C_{BW}.M\left(P_S + H_L + L_S.T_S\right)} \quad [s] \qquad (17)$$

Figure 3 illustrates the relation between the probability, number of calls and service time of the designed model (this graph is for G.729 codec and the 256 kbps serial link).

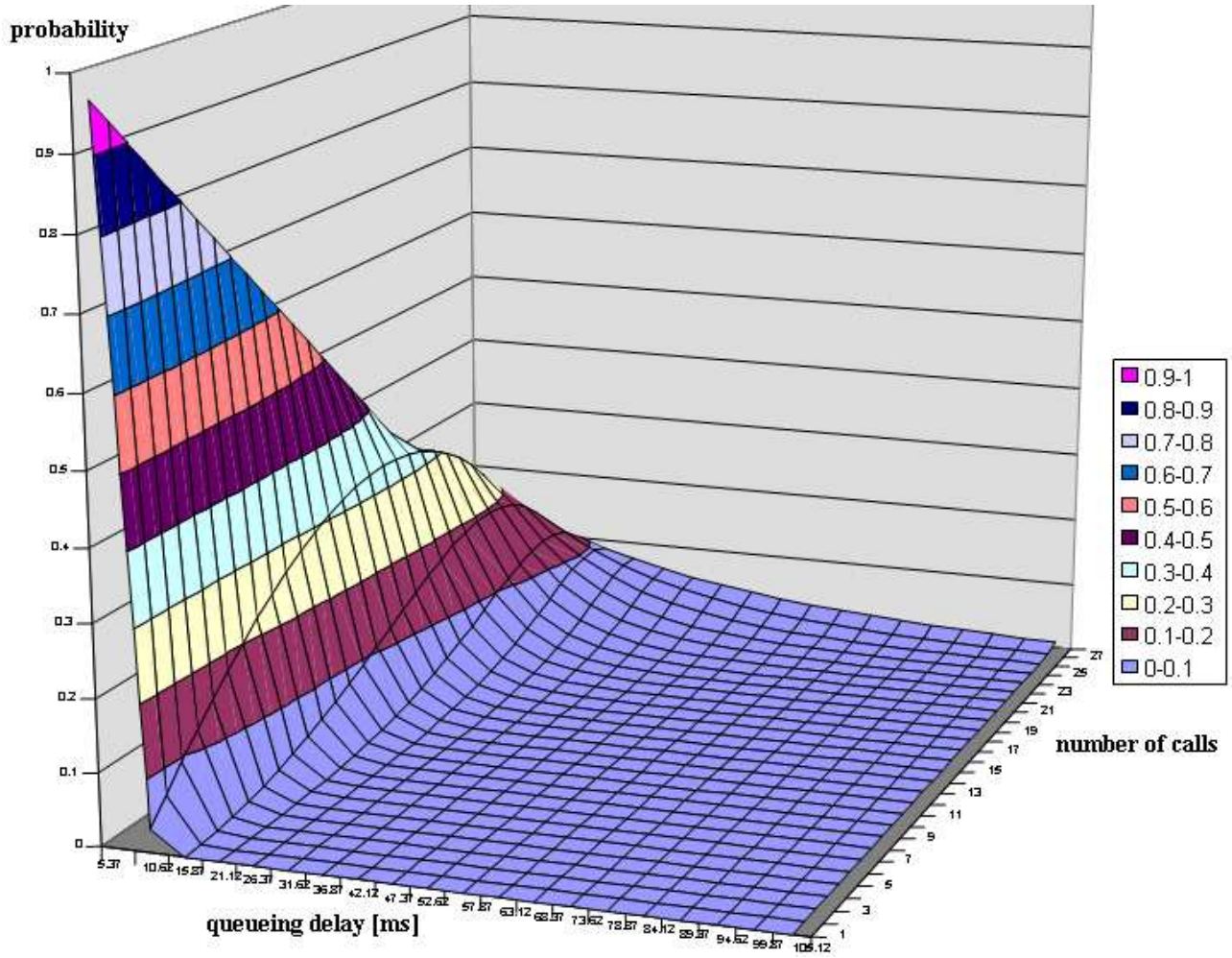


Fig.3. Relation between the probability, number of calls and service time

Likewise a relation for the probability that k-attempts will be waiting in the queue can be derived from equations (14), (15) and(16) applied to (9), (10) and (11) :

$$p_k = \left(1 - M \cdot C_{BW} \cdot \frac{P_S + H_L + L_S \cdot T_S}{L_S \cdot P_S}\right) \cdot \sum_{j=1}^{k} \left[\begin{array}{c} (-1)^{(k-j)} \cdot (j \cdot M \cdot C_{BW} \cdot \frac{P_S + H_L + L_S \cdot T_S}{P_S \cdot L_S})^{k-j-1} \cdot \\ (j \cdot M \cdot C_{BW} \cdot \frac{P_S + H_L + L_S \cdot T_S}{P_S \cdot L_S} + k - j) \cdot \frac{e^{j \cdot M \cdot C_{BW} \cdot \frac{P_S + H_L + L_S \cdot T_S}{P_S \cdot L_S}}}{(k-j)!} \end{array} \right]$$

for k≥2 (18)

$$pk = \left(1 - M.C_{BW} \cdot \frac{Ps + Hs + Ls.Ts}{Ls.Ps}\right) \cdot \left(e^{\left(M.C_{BW} \cdot \frac{P_S + H_L + L_S.T_S}{P_S.L_S}\right)} - 1\right)$$

for k=1 (19)

$$pk = \left(1 - M.C_{BW} \cdot \frac{Ps + Hs + Ls.Ts}{Ls.Ps}\right)$$

for k=0 (20)

The probability of waiting in the queue is expressed in the following relation (21):

$$p_{Tk} = p_k \cdot \frac{P_S + H_L + L_S.T_S}{L_S}$$ for k = < 0 , ∞ > (21)

4 Experiment

A test bed for the estimation of the designed model consisted of two routers interconnected by means of a serial interface with PPP Multilink. The VoIP calls were emulated by IxChariot tester which was used for endpoints and in a console mode for evaluation of the VoIP calls. IXIA IxChariot is a test tool for simulating VoIP traffic to predict device and system performance under various conditions. This tool was used for measuring and traffic simulation. The tests were performed between pairs of network - connected computers. IxChariot endpoints created the RTP streams between pairs and the results were sent to the console and analyzed. Figure 4 illustrates the situation.

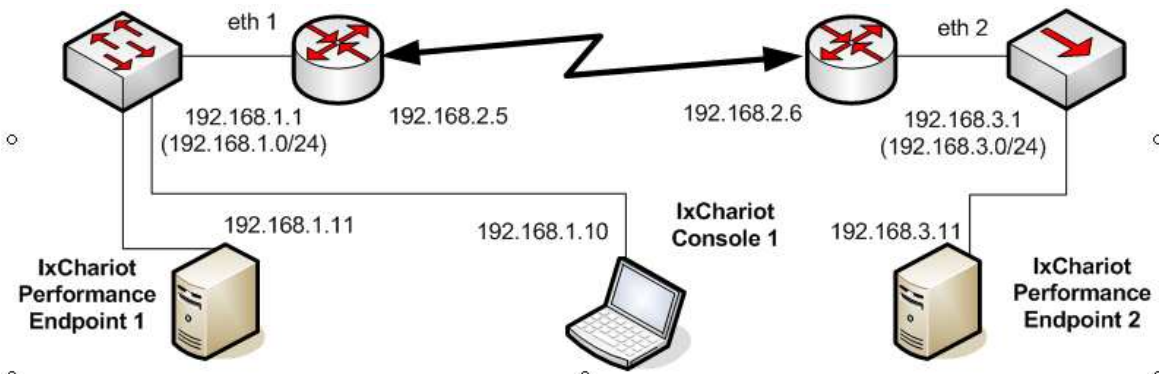


Fig. 4. Scheme of the topology used in the experiment.

The configuration of the serial interface is described below. The *bandwith* value determines the bandwidth that will be used by the interface in range from 128 to 2048 Kbps.

```
interface Serial1/0
description Seriove rozhrani- DTE
bandwidth 2048
no ip address
encapsulation ppp
load-interval 30
no fair-queue
ppp multilink
ppp multilink group 1
```

The queuing mechanism used in this case was priority queuing. The highest priority queue was reserved for RTP packets with the lowest destination port 16384 and the highest port number 32767. The IP RTP Priority command shown in the following example of configuration was used to provide the highest priority to RTP packets. The last parameter is the maximum bandwidth allowed for this queue.

```
interface Multilink1
ip address 192.168.2.5 255.255.255.0
ppp multilink
ppp multilink fragment delay 20
ppp multilink interleave
ppp multilink group 1
max-reserved-bandwidth 100
ip rtp priority 16384 16383 2000
```

Other parameters such as type of codec, timing and number of the RTP sessions also had to be specified directly in the IxChariott tool. The tests ran in an environment with and without a traffic saturation which was done by a UDP generator. The tests were automatically performed by a batch file which was created for this purpose. The files stated below were used to initialise tests and the results were exported to HTML files. These files define the conditions for the performance of the tests and are executed by the following

commands:

```
runtst 1024-711-01-20-1.tst
fmttst.exe 1024-711-01-20-1.tst 1024-711-01-20-1.tst.txt -c
fmttst.exe 1024-711-01-20-1.tst 1024-711-01-20-1.tst.html -h
sleep 30
```

The first line refers to the *runtst* file which runs a test that is passed as a parameter. The second line refers to the *fmttst* file which exports the results to a .txt file while the third line exports the results to an .html file. The line *sleep 30* was inserted there because of errors in the initialization of the endpoints. Once the tests have been finished, we have identified several parameters. Figure 5 below shows an example of the final result.

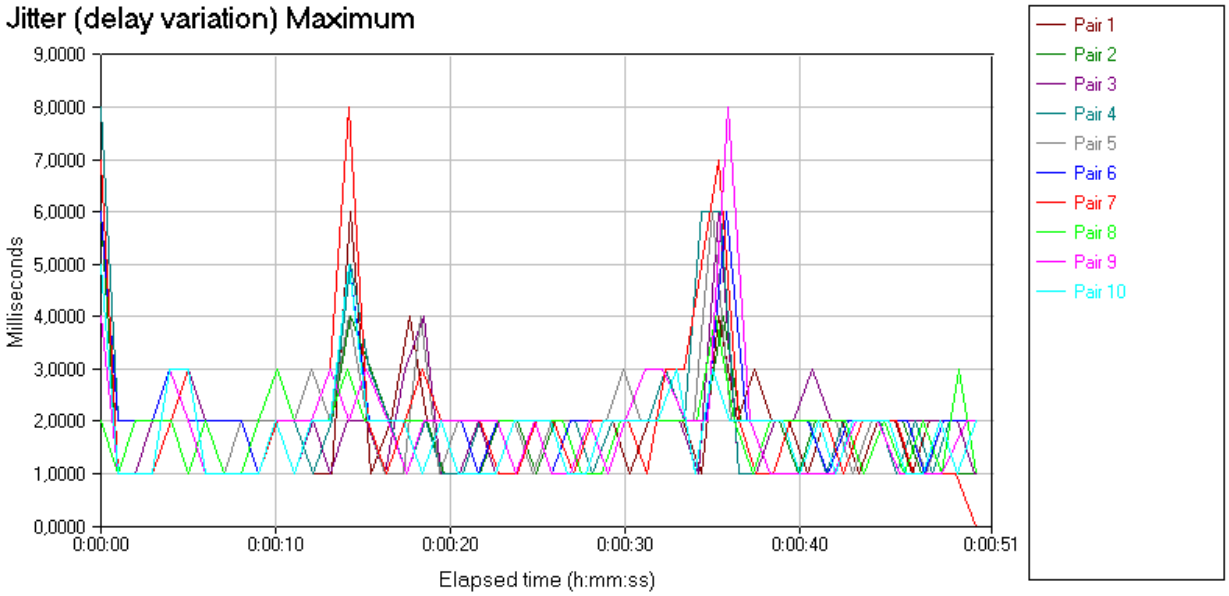


Figure 5. Example of the results.

The results were classified in the groups as follows: MOS, R_factor, jitter and one way delay. With this we could determine average values for all measured results.

It is important to say that the measures were made at different speed (128, 256, 512, 1024 and 2048 kbps) both in an environment without saturation and with saturation. The duration of the test was set to 1 minute during which all observed parameters were recorded at one-second intervals. Every test was repeated five times in order to eliminate any aberrations. We have obtained results of more than 5000 measurements.

5 Conclusion

The designed mathematical model works with a voice traffic approximation supported by a traffic source with Poisson's probability distribution. The described way does not exactly imitate real characteristics of voice traffic, in particular a certain tendency to form clusters. Therefore it was assumed that with the increasing line load the mathematical model will not return absolutely exact information. The measurements showed that in most cases the designed mathematical model returns data with ± 6% accuracy up to the 80 % line load. With the increasing number of simultaneous calls and with the decreasing line load the accuracy of gained data increases. Even though individual voice flows do not match the model of signal source with the Poisson's probability distribution, their sum approximates to this model, in particular with the growing number of calls.

Where 10 simultaneous calls do not load the output line by more than 40 %, the exactness of the model reaches ± 1.5%. As most of designed VoIP networks operate with a much higher number of simultaneous connections, we can assume that the model will return sufficiently exact assessment of an average delay in the network.

We would like to thank our colleagues Michal Halas from the Slovak Technical University for collaboration and Eduardo Rocha, an Erasmus student of the Universidade de Aveiro (now PhD student at the same university) who spent a lot of time in the VoIP laboratory and performed thousands of measurements [6], [7] and [8].

6 References:

[1] PETERS, J., DAVIDSON, J. *Voice over IP Fundamentals.* Indianopolis: Cisco Press, 2000. ISBN 1-57870-168-6.

[2] VOZŇÁK, M.*Výpočet šířky pásma pro hovor v sítích s protokolem IP.* IV. Seminář EaTT 2001, Ostrava VŠB-TU, ISBN 80-248-0031-4

[3] FUJIMOTO, K. , ATA, S., MURATA, M. *Adaptive Playout Buffer Algorithm for Enhancing Perceived Quality of Streaming Applications.* Osaka University, 200.

[4] VOZŇÁK, M.*Voice over IP and Jitter Avoidance on Low Speed Links.* International Conference Research in Telecommunication Technology RTT2002, Zilina, 2002, ISBN 80-227-1934-X.

[5] GROSS, D., HARRIS, C. *Fundamentals of queuing theory.* New York: JW&S, 1998. ISBN 0-471-17083-6.

[6] VOZŇÁK, M., HALÁS,M. *Model end-to-end zpoždění pro VoIP.* Ve sborníku konference Širokopásmové sítě a jejich aplikace, str. 128-131, UP Olomouc, ISBN 978-80-244-1687-8, 29-30.5.2007.

[7] HALÁS, M., KYRBASHOV, B., VOZŇÁK, M. *Factors influencing voice quality in VoIP technology.* In: 9th International Conference on Informatics‘ 2007, pp. 32-35, Bratislava, ISBN 978-80-969243-7-0, 21-22.June 2007.

[8] VOZŇÁK, M., ROCHA, E., KYRBASHOV, B. *End-to-end delay of VoIP.* In: International Conference RTT 2007, Žilina: University of Žilina, 2007, pp. 466-469, ISBN 978-80-8070-735-4, September 2007.