Mathematical Representation of VoIP Connection Delay

Ivan BAROŇÁK, Michal HALÁS

Dept. of Telecommunications, Slovak University of Technology, Ilkovičova 3, 812 19 Bratislava, Slovak Republic

baronak@ktl.elf.stuba.sk, halas@ktl.elf.stuba.sk

Abstract. *The main topic of this article is to define mathematical formulation of VoIP connection delay model. It handles about all partial delay components, the mechanism of their generation, facilities and their mathematical formulation. Thereafter based on mathematical formulation of all partial delay components, the final mathematical model of whole VoIP call delay is created. In conclusion of this article the results of the designed mathematical model are compared with the experimentally gained results.*

Keywords

VoIP, delay, mathematical model, M/D/1/∞.

1. Introduction

Using the IP protocol for the voice transmission its result quality is influenced by several negative factors. Most of them directly relate mostly with non-connection IP network orientation and with a zero support of the connection quality guarantee from networks [1]. Delay is a very important factor considerably influencing the result quality of VoIP (Voice over internet protocol) connection. There are several delay types in IP networks which differ from each other in the place and mechanism of their creation and other attributes. Each delay component influences the result voice packet delay in a different way.

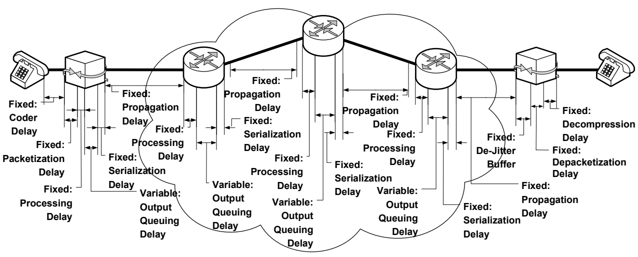

Fig. 1. Delay components.

It is very useful to be able to obtain the estimation of the mean delay in a VoIP network, already during the process of designing the network. It is necessary to be able to calculate the value of the final voice quality in the network. The first possibility how to obtain the value for the mean delay in the network is to realize a simulation of the data traffic in the network. This is very intensive technique. Another way is to create an analytical model of the mean delay in the VoIP network and based on it to calculate easily the value based on known parameters of the network and the data traffic inside it. This paper is focused exactly to define the analytical model of the whole voice transmissions' delay in the VoIP network.

2. Delay Components of VoIP Connection

It deals with the detail description of individual components, delay, explanation of creation mechanisms, their mathematical description and consequently it deals with the creation of a mathematical model of the VoIP connection delay in the network.

Individual delay components and places of their creation are characterized in Fig. 1. Delay components are:

1. Coder delay.
2. Packetization delay.
3. Output queuing delay.
4. Serialization delay.
5. Propagation delay.
6. De-jitter delay.
7. De-packetization delay.
8. Decompression delay.

Mathematical representations of the delay components are known in the technical community. Some of them are very easy to calculate, some of them are hardly defined for variable parameters of VoIP connection and some of them could be estimated based on suitable mathematical models.

The deal of this article is to define a simple analytical mathematical model of the whole voice connection delay in the VoIP network. Such a model is defined using known mathematical representations of each delay component and using the process of modeling the packet delay in switch using the bulk service M/D/1/∞. Using the designed model and knowing information of network parameters, we should be able to simply calculate the mean delay of voice connection in the VoIP network.

2.1. Coder Delay

Coder delay depends on a codec selection and a DSP (Digital signal processing) processor capacity. It is the time necessary for the DSP processor to compress PCM (Pulse-code modulation) frame of samples. Each coding algorithm specifies the size of PCM frame of samples that are coded together. It causes a delay, which is called "Frame size delay". This delay type depends on the frame size which is being coded and is exactly defined by the codec selection [2].

The second partial delay rising from the coding process is so-called algorithmic delay. The algorithmic delay rises in dependence on the algorithm operation mode of signal coding, considering the fact that some algorithms require to know *N+1*, *N+2*… sample to encode the *N* sample. Because of this reason it is necessary to keep individual samples back in the buffer store. This holdback results in the signal delay having an exactly defined delay value and depends on the algorithm, which is used by a particular coder. The sum of the above mentioned partial delays is marked as a coder delay. Individual delays for selected codecs are shown in Tab. 1.

Coder	Type	Rate [kbps]	Packetization period [ms]	Frame size [ms]	Algorithmic delay [ms]	Codec delay [ms]
G.711	PCM	64	20	0,125	0	0,125
G.723.1	MPC-MLQ	5,33	30	30	7,5	37,5
G.723.1	ACELP	6,4	30	30	7,5	37,5
G.726	ADPCM	32	20	10	0	10
G.728	LD-CELP	16	30	0,625	0	0,625
G.729A	CS-ACELP	8	20	10	5	15

Tab. 1. Coder delay: PCM – Pulse-Coded Modulation; MPC-MLQ – Multipulse LPC with Maximum Likelihood Quantization, ACELP – Algebraic Code Excited Linear Prediction, ADPCM – Adaptive Differential Pulse-Code Modulation, LD-CELP – Low-Delay Code Excited Linear Prediction, CS-ACELP – Conjugate Structure Algebraic Code Excited Linear Prediction.

2.2. Packetization Delay

The packetization delay rises during the process of data blocks encapsulation into packets, which are consequently transmitted by the network. By reason of that the IP network operates with a variable packet size, it is possible to define which packet size we want the system to generate. Thanks to this process we can affect the amount of data blocks from the coder to be transmitted together in one IP packet [2], [3].

As a matter of course, if we want to transmit several blocks at once, it is necessary to keep them back in the buffer store till the coder generates the last data block for the particular packet, afterwards it is possible to encapsulate these data and start the transmission.

The packetization delay is adjusted in multiples of the packetization period of the selected codec. Hence it specifies how many data blocks are transmitted in one packet.

It is normal that the smaller packets we generate the fewer data blocks are transmitted at once and the smaller delay is added to the process. In the case that we create smaller size packets, it is necessary to transmit more IP packets at the same time, than in the case when bigger size packets are created.

A problem occurs with the transmit area efficiency utilization. Whereas each packet includes its header and tail, which has to be transmitted several times during a bigger packet amount transmission, in some cases the effective data flow can be almost triple of the original data coder flow. It follows the necessity to find appropriate compromise between delays and transmission efficiency.

From the given specification it is valid that the amount of transmitted data in RTP packet has to be minimum 20 B and maximum 160 B. The estimation process is given by the reference (1):

$$T_P = \frac{8.P_S}{C_{BW}} \qquad [ms] \tag{1}$$

where T_P is packetization delay [ms], P_S denotes payload size [B], and C_{BW} is Codec Bandwidth [kbit/s].

In a real traffic, as it was mentioned above, the packetization delay is adjusted directly, and in multiples of the packetization period of a given codec. The period for the codecs G.711 a G.729 is 10 ms, for the codec G.726 20 ms and for the codecs G.728 a G.723.1 it is 30 ms.

2.3. Change of Bandwidth

Each protocol and each transmit technology used in IP networks adds its information in form of an additional header and tail to useful data. The data contain specific information which is required by a particular technology or protocol. According to which transmitting technology and protocol is used, the amount of additional information varies, except the useful data, necessary to be transmitted, [4], [6].

During the whole routing of IP packet transmission it is necessary to use several transmit technologies and protocols. That time, during the transmission between particular technologies, useless headers are eliminated and headers requisite for transmission in a new surrounding are added.

Of course, the transmission of additional information requires a sufficient transmission area. According to the amount of additional information in the header it is necessary to estimate how high transmission bandwidth will be requisite in each segment of the transmission route.

From available information about the size of the required transmission codec area and its sampling interval which is in principle the same as packetization delay, following the reference (2), we are able to estimate the IP packet data sector size.

$$P_S = \frac{C_{BW}.T_{SI}}{8} \qquad [B] \tag{2}$$

where P_S is payload size [B], T_{SI} is voice sample interval = T_P packetization delay [ms], C_{BW} denotes codec bandwidth [kbit/s], and for RTP it must be valid 20 B≤ P_S ≤ 160 B.

With this estimation we gain an evidence of the size of the data field in the packet. If we add to it a header and tail size, with the help of (3) and (4) references we are able to estimate the required real transmission area of a particular data flow.

$$T_{BW} = \frac{H_L.8}{T_{SI}} + C_{BW} \quad [kbit/s] \tag{3}$$

where T_{BW} is total bandwidth [kbit/s], H_L denotes header and tail length [B], and T_{SI} is voice sample interval [ms].

After modification

$$T_{BW} = \frac{H_L.C_{BW}}{P_S} + C_{BW} \quad [kbit/s]. \tag{4}$$

Each technology and protocol has a specified header and tail of an exactly defined size, see Tab. 2.

Technology / protocol	Header and tail size [B]
Ethernet	14 B
Frame Relay	4 B
PPP	6 B
IP + UDP + RTP	40 B
IP + UDP + cRTP	2 B

Tab. 2. Header size.

From the table it results that the combination of IP+UDP+RPT protocols, which is standardly used to transmit a voice signal in a IP network, adds 40 B header to each packet. The size of this header could in certain cases reach almost double size of a data field. This situation occurs mainly using the codecs with very small transmission requirements. When it is necessary to choose a very low payload size because of restriction of the packetization delay, in these cases the efficiency of transmission area utilization is very low. Whereas the majority of data in a header changes very little, it is possible to use a technology of header compression cRTP. This technology allows a header compression IP+UDP+cRTP the way, that a bulk of a time only 2 B of additional information are transmitted.

Next option how to decrease VoIP transmission requirements is the utilization of VAD (Voice Activity Detection) technology in a transmitter. The technology tries to identify places in a call, when the customer doesn't speak, therefore it isn't necessary to transmit whole voice data. This technology is suitable to be replenished in the receiver by CNG (Comfort Noise Generator) technology, which during the call, in the time when no signal is being transmitted, locally generates a noise in the receiver, what causes the communication comfort increase. VAD decreases the transmission area at average 35 %.

2.4. Serialization Delay

Serialization delay is the next delay component, which depends on the transmission bandwidth [2], [7]. The packet sending takes some time. This time depends on the transmission medium rate and on the individual packet size. Of course, this delay negatively influences the result voice delay. For its estimation the relation (5) is valid:

$$T_{SER} = \frac{F_S}{L_S} \quad [ms] \tag{5}$$

where T_{SER} is serialization delay [ms], F_S frame size is a total packet size [b], and L_S denotes line speed [kbit/s].

$$F_S = H_L + P_S \quad [b] \tag{6}$$

where P_S is payload size [b], and H_L denotes header and tail length [b].

Serialization delay is negatively manifested mainly on low rate lines, during the transmission of big packets. Then it can even reach values of several tenths of milliseconds.

2.5. Network Switching Delay

Several data and voice flows meet in the network switch and can direct different directions. A situation may occur when several voice and data flows have to be directed only at one output line. In such a situation it is important to decide in which order packets are sent on the output line [2].

With a chosen way of packet processing it is possible to affect a result delay in a notable constant. Voice packets are necessary to be processed in preference of other data traffic and this way to secure a higher connection quality. Voice packets identification is possible to be done with the help of interserv or diffserv mechanisms. Particular processing in the switch is mainly managed with the help of PQ (Priority queuing) method, eventually PQ/WFQ (Priority queuing / Weighted fair queuing) or with some other method.

2.6. Traffic Source Modeling

The description of a VoIP traffic by a suitable mathematical model isn't a simple thing to do. This topic is discussed in many publications [8], [9], [10], [11]. It is proven that in certain circumstances the voice traffic can be modeled by a source signal, which probabilistic random variable distribution matches Poisson's probability distribution [12], [13], [14]. Using this model it is possible to simulate time probability of voice packet access, during which time in a certain interval of traffic parameters it offers sufficient exact data.

Because of a simple analytical expression and implementation into the switch model, we have decided to use this level of abstraction. Assumptions, where the modeling, with the help of Poisson's probability distribution, conduces to gain relevant information, are as follows:

- For codecs where the sampling period is T_S=20 ms (G.711, G.729, G.726 24/32 kbit/s), the Poisson's model is suitable for the network loading max 60%,

- For codecs where the sampling period is T_S=30 ms (G.723.1 5,3/6,3 kbit/s, G.728), the Poisson's model is suitable for the network loading max 70%.

If you meet all listed requirements, we can await that in spite of this form of abstraction, from the analytical switch model, we can get information representing real situation.

2.7. Analytical Expression of Switching Delay

Switch traffic with PQ optimalization is based on a preferred packet servicing in a primary voice PQ queue [17], [18], [19]. In the case when there is an effectively utilized packet fragmentation mechanism on the output switch line, it is possible to ignore the influence of serialization delay of data packets with a lower priority than voice packets have. In this case, for the modeling requirements of traffic loading and delay in switch, it is sufficient to watch a delay only in a priority queue. Servicing requirement technique in a priority queue responds to the model of bulk service M/D/1/k, where k is a buffer size.

In order to create an analytical model of switching delay we can ignore the buffer size and muse upon a system with a sufficient buffer size, where the loss of preferred packets doesn't occur.

Providing the M/D/1/k model can be replaced by M/D/1/∞ one, thanks to what we are able to create the analytical expression of switch buffer store seizing. Consequently it is very easy to gain an analytical model of the packet delay in a switch [21], [22], [23].

For the above mentioned model of a bulk service it is valid that the switch from one stage is possible only to the directly closest one, what responds to the model of access or the sending of one packet in a switch.

This model is suitable only if it meets the following conditions:

- The period between accesses of particular requirements responds to the Poisson's probability access distribution,
  - if we start from an assumption that only a voice traffic from M sources enters the model of switch, consequently their mergence has Poisson's probability distribution,
- $\lambda(t)$ – failure-probability density of requirement's rise on the service is constant,
  - if we reflect upon the delay model, which is entered by M access voice flows, not rising or waning that time, this requirement is fulfilled,
- Servicing technique follows FIFO rule (First In First Out),
  - switch operation mode in PQ queue satisfies FIFO rule,
- Service time is a constant value,
  - if we assume that M input voice flows use the same compressing codec and generate the same length packets, the transmission time of each packet on the output line is constant, consequently the packet service time is constant, too.

Following the given assumptions we can estimate the system loading, hence the output line loading by voice packets, with the help of the reference

$$\rho = \frac{\lambda}{\mu} \tag{7}$$

where λ is intensity of requirements´ access [s^{-1}], μ is service rate of requirements [s^{-1}], ρ denotes system loading. For the system's stability it must be valid that $0 \le \rho < 1$.

Providing that we count only with one traffic source with a constant bit rate C_{BW} and a constant data packet size P_S, the intensity of requirements´ access can be expressed as follows

$$\lambda = \frac{C_{BW}}{P_S} \quad [s^{-1}]. \tag{8}$$

For the service rate of requirements it is valid

$$\mu = \frac{1}{T_{SER} + T_S} \quad [s^{-1}] \tag{9}$$

where T_{SER} is serial delay of the output line [s], and T_S denotes packet processing time in the switch [s].

Then for the probability that there are k requirements waiting in the system, the following reference is valid:

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

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

$$p_k = (1-\rho) \qquad \text{for } k = 0$$

For the mean time the requirement spends in the system, the following reference is valid

$$T = \frac{1}{\mu} + \frac{\rho}{2(1-\rho)\mu} \quad [s] \tag{11}$$

where $1/\mu$ is a service time of one requirement.

Then, for the average number of requirements in N system, the following reference is valid

$$N = T.\lambda\,. \tag{12}$$

Providing that the switch is entered only by M voice flows and with the utilization of Poisson's distribution, where by the mergence of several voice flows with Poisson's probability distribution we gain a data flow, as well as with the Poisson's probability distribution, which λ parameter satisfies the sum of λ_i parameters of individual voice flows, then we can state the access intensity as follows

$$\lambda = \sum_{i=1}^{M} \frac{C_{BWi}}{P_{Si}} \quad [s^{-1}]. \tag{13}$$

Providing that all voice flows entering the switch use the same codec and are transmitted by the same transmission surroundings, the reference (13) can be simplified as follows

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

If we know the output line transfer rate and packet processing time for the particular switch, it is possible to define the service rate of the particular system after institution of references (5) and (6) to the relation (9) as follows

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

After institution of references (14) and (15) to the reference (7) we get the following formulation for the system overloading

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

After institution of references (14), (15) and (16) to the reference (11) we get the following formulation of the mean service time

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

Likewise after institution of (14), (15) and (16) to the reference (10) it is valid for the probability that there are exactly k requirements in the system

for $k \geq 2$

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

for $k=1$

$$p_k = \left(1 - M.C_{BW} \cdot \frac{P_S + H_S + L_S.T_S}{L_S.P_S}\right) \cdot \left(e^{\left(M.C_{BW} \cdot \frac{P_S + H_L + L_S.T_S}{P_S.L_S} \right)} - 1 \right), \tag{18}$$

for $k=0$

$$p_k = \left(1 - M.C_{BW} \cdot \frac{P_S + H_S + L_S.T_S}{L_S.P_S}\right).$$

Thereof if we express the service time with the help of the reference (15) for the service rate, we can get the following reference for the delay probability

$$p_{Tk} = p_k \cdot \frac{P_S + H_L + L_S.T_S}{L_S} \quad \text{for } k = < 0, \infty > \tag{19}$$

The relation (17) specifies the average packet delay time in the switch, with the relation (19), which specifies the probability distribution of the mentioned delay, and together they generate the mathematical model of the delay in the switch. Following this model we can estimate the average packet delay in the switch for the selected input configuration parameters.

It is important to be aware of the fact that the mentioned model of the packet delay in the switch counts also with the serial output line delay. It follows that, in the result delay gained with the help of this model, there is a serial delay on its output line inclusive.

For instance the reference between the number of calls and average delay on 256 kbit/s line with PPP protocol is shown in Fig. 2.

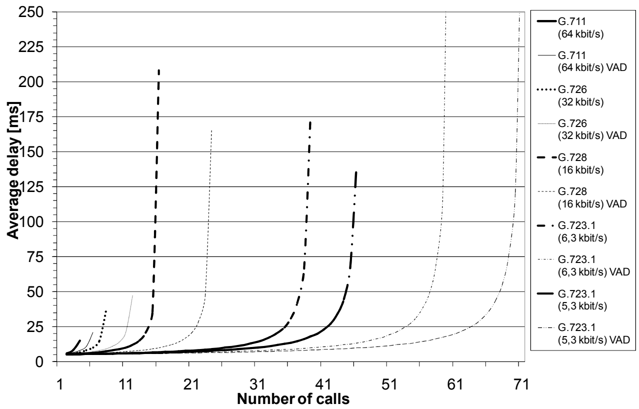

Fig. 2. Average delay dependency on the number of calls for 256 kbit/s line.

2.8. Propagation Delay

The transmission delay relates with signal transmission physical regularities in certain physical surroundings [25], whether we deal with the electric signal rate in a line wire, or light in an optical fiber. This delay type is very dependent on the used transmission technology and mainly on the distance through which the signal is transmitted.

Dealing with distances from units to tenths of kilometers, delay values are trivial. Therefore it is enough to reflect on the presence of this delay type only in big spinal networks. These network types are presently built exclusively on the basis of optic signal transmission.

Diametrically different situation happened when the voice transmission is done using the wide range wireless connection, e.g. satellite connection. In this case the transmission delay should gain great values, but this is very specific situation and therefore this article doesn't handle about this case. The whole article handles only about the wired VoIP networks.

Light, as well as any electromagnetic wave, is characterized by its rate of spread in particular surroundings, which is closely related with the refractive index for particular transmission surroundings. Using the reference (20) we can estimate the light rate of spread in an optical fiber

$$v = \frac{c}{n} = 2,07.10^8 \quad [\mathrm{ms}^{-1}] \tag{20}$$

where $c = 3.10^8$ [ms^{-1}] is speed of light in vacuum, $n = 1,45$ is refractive index for silica glass for the wavelength $\lambda = 1\mu$m, and v denotes light rate of spread in the optical fiber [ms^{-1}]. If this rate is instituted to the reference (21) we get the total value of the transmission delay for the particular line transmission.

$$T_{PD} = \frac{1000L}{v} \quad [ms] \tag{21}$$

where T_{PD} is propagation delay [ms], and L is length [m].

From the above mentioned references it follows, that the value of the transmission delay is approximately 4,83 μs on 1 km of the optical line. The introduced value, in relation to other delays influencing voice transmission, used in short distances, is almost trivial. The only cases when this delay type can considerably influence the general delay are very long transcontinental lines, when this delay can reach values about several tenths of milliseconds.

2.9. De-jitter Delay

The de-jitter delay is closely related with the variable delay in the network, when it is necessary to eliminate changes of these variable components with the help of supplementary buffer store, so-called jitter buffer [4], [2]. On one hand the jitter buffer size directly influences the possibility to align bigger delay scatter, on the other hand it increases the statistical part of the delay. It follows that it is necessary to set an optimal buffer size so that it is able to eliminate an ultimate part of divergences in packet accesses.

On the other hand it is necessary to secure that the buffer store will not bring a too big delay into the transmission network. Its size is typically adjusted in the range from 1.5 to 2 multiple of the sum of all variable delay components. In most cases, when this value is adjusted statistically, jitter buffer sizes are about 30-60 ms. If the jitter buffer size changes dynamically, they are adjusted typically as its maximum values about 100-150 ms. Though, more exact adjustment is possible to be done only by a good situation analyze in the network. In the delay model the size of this buffer directly figures as the next additive delay component.

2.10. Depacketization Delay

The depacketization delay, with its mechanism of generation, is very similar to the packetization delay [2], [3]. In the case when one packet contains several data blocks it is necessary to keep all blocks belonging to the particular packet back during the signal packetization, till the last data block is generated. Then it is possible to send the whole packet by the network. A similar situation occurs during the depacketization of data blocks. In the moment of the packet access we receive several data blocks which were transmitted through it. The first block can be immediately moved forward to decompression, however next blocks have to be kept back in the buffer store, so that we can send it to the decoder in exact time periods, as they were generated by the switch. The first data block doesn't have to be kept back at all. The last data block has to be kept back for at longest, exactly for the value of packetization delay.

A situation occurs that the first block was kept back for the value of the packetization delay during the packetization and during the depacketization it isn't kept back at all. Vice-versa, the last block isn't kept back during the packetization, but during the depacketization it is kept back for the value of the packetization delay.

Following the initiated fact, 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. On that account during the estimation of the last delay we count only with one constant packetization delay value.

2.11. Decompression Delay

The decompression delay, likewise the coder delay, is dependent on the compressing algorithm selection [24]. At average the decompression delay is approximately 10% of the compressing codec delay. But it is very dependent on the computing decoder operation speed.

This delay type can be mathematically shown as follows

$$T_{DCD} = 0,1.T_{CD} \quad [\mathrm{ms}] \tag{22}$$

where T_{DCD} is decompression delay [ms], and T_{CD} denotes coder delay [ms].

3. Resultant Delay Model

Using the above mentioned information the model of the resultant delay in the network can be expressed as follows.

Let's count with a network structure, which is illustrated in Fig. 3. The two endpoints operate in a network, which are connected with 3 switches and 4 different lines. Only a voice communication is progressing in the network. Because of the simplification, let us assume that all voice flows are using the same compressing algorithm with the same adjustment.

The endpoints are specified by the used codec and the size of the data packet part, which they generate. Each line is specified by its transfer rate, used transfer protocol, transfer technology and its length. Switches, which are located in the network, execute M_i current connections directed towards the same output line and are specified by their process delay. There is a buffer store in a switch designed for jitter compensation that is specified by its size.

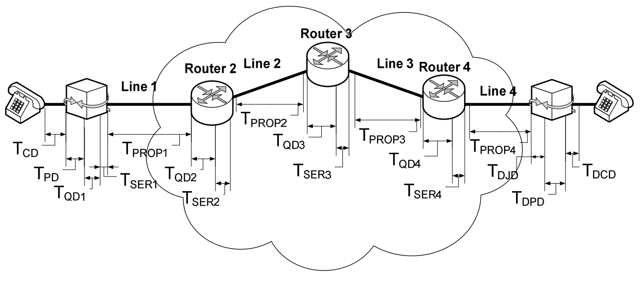

Fig. 3. Individual delay components.

Following the previous information we can express the general delay for this case as follows

$$T = T_{CD} + T_{PD} + T_{SER1} + T_{PROP1} + T_{QD2} + T_{SER2} + T_{PROP2} +$$
$$+ T_{QD3} + T_{SER3} + T_{PROP3} + T_{QD4} + T_{SER4} + T_{PROP4} + T_{DJD} + T_{DCD}$$

$$T \ = \ T_{CD} + T_{PD} + \sum_{i=1}^{4} T_{SERi} + \sum_{i=1}^{4} T_{PROP1} + \sum_{i=2}^{4} T_{QD2} + T_{DJD} + T_{DCD} \quad [\mathrm{ms}] \tag{23}$$

Consequently after institution we get a mathematical model of the delay in the network

$$T \ = \ (1{,}1.\mathrm{N}).T_{CD} + \frac{P_S}{C_{BW}} + \frac{1000.\sum_{i=1}^{4} L_i}{v} + T_{DJD} +$$
$$+ \frac{1}{2} \cdot \sum_{i=2}^{4} \left[\frac{P_S + H_{Li} + L_{Si}.T_{Si}}{L_{Si}} \cdot \frac{2.P_S.L_{Si} - C_{BW}.M_i\left(P_S + H_{Li} + L_{Si}.T_{Si}\right)}{P_S.L_{Si} - C_{BW}.M_i\left(P_S + H_{Li} + L_{Si}.T_{Si}\right)} \right] \quad [\mathrm{ms}] \tag{24}$$

where T_{CD} is general coder delay [ms] from Tab. 1, P_S is payload size, the selected data block size which is transmitted in one packet [b], C_{BW} denotes codec bandwidth – the codec transfer rate [kbit/s] from Tab. 1, H_{Li} is header and tail length – the size of the packet header and tail for the line i [b] from Tab. 2, L_{Si} is line speed i – the transmit rate of the line [kbit/s], L_i denotes line length i – the transmit line length [km], v is rate of propagation of light in the optical fibre [ms^{-1}], T_{DJD} is dejitter delay – the selected dejitter buffer size [ms], M_i denotes the number of current calls in the switch directed towards the output line i, and T_{Si} is the packet processing delay in the switch i [ms].

The reference (24) is a mathematical expression of the delay model in the IP network of the defined structure. This model can be generalized to any of well-known network topology and with knowledge of the traffic in a network; consequently we can estimate the average voice connection delay.

4. Verification of Experimental Results

In order to verify the results gained with the help of the designed delay model, several experimental measurements have been done. As the optimal process how to realize all mentioned measurements one process was selected, where on a known network structure a sufficient number of voice connections is generated and parameters of each of them are observed. We have decided to use the product IxChariot from Ixia company. The presented system enables not only to generate the traffic of the exactly defined parameters, but consequently to evaluate it in detail. Inter alia it supports the creation of several VoIP voice channels, following exactly specified parameters. IxChariot surroundings also enable us to observe individual connection parameters such as delay, packet loss, jitter, MOS (Mean opinion score) and R factor.

During the experiment voice flows were generated with the help of codecs G.711, G.723.1 and G.729. Rate adjustments of PPP line were changing in values 128, 256, 512, 1024, 2048 kbit/s and also the values of framing for the individual codecs were changing. For the codecs G.711 and G.729 tests for the values of the framing period 20, 40. 60 ms were done and for the codec G.723.1 for the values 30 and 60 ms. For each line rate transmission tests of one, two, three, etc. current voice channels were done step by step. The maximum number of current voice connection was adjusted regarding PPP line rate the way that there won't occur any line overloading and consequently any unasked packet loss.

For each line rate and each defined number of current connections several tests for individual framing values were performed. Within one test only voice flows of one codec were generated, with the same framing adjustments for all voice channels. Duration of all tests was adjusted on 1 minute, during which time all observed parameters were recorded in one-second intervals. On the score of divergence minimalization a sequence of 5 independent measuring with identical adjustments was always done.

All gained information from individual tests was resumed, statistically evaluated and consequently confronted with the estimated values according to the designed mathematical model of delay. The aim was to verify the estimated values and to define the interval of operating parameters, in which the mentioned model offers responsible results. Fig. 4 expresses percent divergence of the estimated values from the measured ones in dependence on the line load for the PPP line rate 1024 kbit/s.

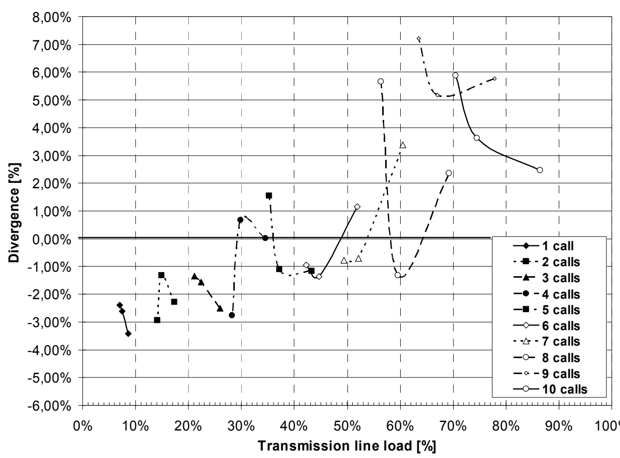

Fig. 4. Divergence estimation dependence on load for 1024 kbit/s line and G.711 codec.

The mode of dependence on the line load was chosen upon the assumption that with the increasing line load a certain form of clusters in voice traffic will be manifested, though it wasn't allowed in the model.

5. Conclusion

There was a voice traffic approximation used in the designed mathematical model, with the help of traffic source, which had Poisson's probability distribution. The mentioned way doesn't represent real properties of voice traffic exactly, particularly it's certain cluster character. Therefore it was assumed that with the increasing line load the given mathematical model won't provide absolutely exact information. There was an assumption that gained results of the mathematical model will differ from the real traffic in the line load more than 70%. [12]

Measurements showed that in the majority of cases the designed mathematical model provides data with closeness ± 6% up to the line load 80%. With the increasing number of current calls and with the decreasing line load the closeness of gained data increases. Even though individual voice flows don't match the model of signal source with the Poisson's probability distribution, their sum approaches this model specifically with the growing number of calls. In cases when 10 current calls don't load the output line more than 40%, the exactness of the model approaches values of ± 1.5%. Forasmuch as in the majority of designed VoIP networks it is counted with much bigger number of current connections, there is an assumption that the given model will provide sufficiently exact assessment of the average delay in the network.

Acknowledgement

This work is a part of AV 1002/2004, VEGA 1/0156/03 and VEGA 1/3118/06 projects.

References

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

[2] *Understanding Delay in Packet Voice Networks*. CISCO, http://www.cisco.com, Mar. 2005, 20 p.

[3] The Essential Report on IP Telephony. *World Telecomunication Development Conference WTDC-02*, Istanbul, Turkey, 2003, 191 p.

[4] *IP Voice Bandwidth and Latency*. http://www.techeez.com, Mar. 2004, 8 p.

[5] ŠIMÁK, B., KOŠKA, K. Processing of audio and video signals transferred over ATM networks. In *Proceedings of Workshop 2001*. Prague: CTU, 2001, vol. A, p. 314. ISBN 80-01-02335-4.

[6] *ITU-T Recommendation E.492*, Traffic Reference Period. Geneva, Feb. 1996.

[7] MINOLY, D., MINOLI, E. *Delivering Voice over IP Networks*. Second Edition, Willey Publishing Inc., ISBN 0-471-38606-5, Sep. 2002, 512 p.

[8] SRIRAM, K., WHITT, W. Arrival processes in packets multiplexers for voice and data. *IEEE Journal on Selected Areas in Communications*, 1986, vol. SAC-4, no.6, pp. 833-846.

[9] FISHER, M., MASI, D. Voice packet arrival models and their affect on packet performance. In *Applied Telecommunications Symposium*. April 3-7, 2005. San Diego, Calif., 2005, 8 p

[10] ANDERSSON, A. Capacity study of statistical multiplexing for IP telephony. *SIC Technical Report ISRN*, 2000, 40 p.

[11] ALFA, A., LI, W. A homogeneous PCS network with Markov call arrival process and phase type cell residence time. *Wireless Networks* 8, 2002, p. 597-605.

[12] PAXSON, V., FLOYD, S. Wide area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking,* 1995, vol. 3, pp. 226-244.

[13] JIANG, W., SCHULZRINNE, H. Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation. In *Proc. IEEE Int. Conf. Computer Commun. Networks.* Las Vegas (NV), 2000, p. 82-87.

[14] DEMICHELIS, K. Measurement of packet delay variation on real traffic – network monitoring. *IETF Draft*, PAM 2000, Hamilton, New Zealand, 2000.

[15] FILKA, M., SLAVÍČEK, K., PETRENKO, M., DOSTÁL, O. Transmission capacities for evolution and image medical information. In *International Scientific Conference - COFAX 2004*. COFAX - Telekomunikácie 2004. Bratislava (Slovakia), 2004, p. 125 - 128, ISBN 80-967019-6-7.

[16] MEDVECKÝ, M. ORGOŇ, M. Bezpečnosť IP sietí (Safety of IP networks). *Časopis TELEKOMUNIKACE*, 2005, vol. XLII , no.3, p. 13–18 (in Slovak).

[17] CAO, J., CLEVELAND, W., LIN, D., SUN, D. *Internet Traffic Tends Toward Poisson and Independent as the Load Increases*. Lucent Technologies, Apr. 2002, 18 p.

[18] GEORGOULAS, S., PAVLOU, G., FLEGKAS, P., TRIMINTZIOS, P. Buffer and bandwidth management for the expedited forwarding traffic class in differentiated services networks. In *Proceedings of LCS*, 2002, 4 p.

[19] WILLINGER, W., PAXSON, V. Where mathematics meets the internet. *Notices of the AMS*, 1998, vol. 45, pp 961-970.

[20] CHROMÝ, E. Optimizing processes in convergent networks. In *RTT Conference focused on Research in Telecommunication Technology, RTT 2006*. Nove Mesto na Morave (Czech Republic), Sept. 11-13, 2006, pp. 295-299, ISBN 80-214-3243-8.

[21] BU, T., LIU, Y., TOWSLEY, D. On the TCP-friendliness of VoIP traffic. In *IEEE Infocom 2006*, 12 p.

[22] BI, Q., CHEN, P., YANG, Y., ZHANG, Q. An analysis of VoIP service using 1xEV-DO Revision A System, 2006. *IEEE Journal on Selected Areas in Communications*, 2006, vol.24, no.1, p.36-45.

[23] VOZŇÁK, M., ZUKAL, D. Assessment of VoIP quality. In *6th International Conference RTT 2005*, Hradec nad Moravici (Czech Republic), 2005, p.641-645, ISBN 80-248-0897-8.

[24] HASSAN, M., NAYANDORO, A., ATIQUZZAMAN, M. Internet telephony: Services, technical challenges, and products. *IEEE Communications Magazine*, April 2000, pp. 96-103.

[25] VOZŇÁK, M. *Optimalizace hlasového provozu s ohledem na kvalitu hovoru v sítích s technológií VoIP (Optimization of voice transmission with respect to the quality of calls in VoIP networks)*. Ostrava, 2002, 112 p. (in Czech).

[26] GALAJDA, P., MARCHEVSKÝ, S., KOCÚR, D., BENČO, S. Mobile Learning at the Technical University of Košice. In *Proceedings of The 9th IEEE International Conference on Intelligent Engineering Systems 2005* (INES'2005). Cruising on Mediterranean Sea, September 16-19, 2005, pp. 233-237.

About Authors...

Ivan BAROŇÁK, prof. Ing., PhD. was born in Žilina, Slovakia, in July 1955. He achieved the electronic engineering degree at the Slovak Technical University Bratislava in 1980. Since 1981 he has been a lecturer at the Department of Telecommunications, STU Bratislava. In 1992 he submitted PhD work in the field of Terminal telephone equipment. In 1995 he became an associate professor for the subject applied information. Nowadays he works as an associate professor at the Department of Telecommunications of FEI STU in Bratislava. Scientifically, professionally and pedagogically he focuses on problems of digital switching systems, ATM, Telecommunication management (TMN), Next Generation Networks, problem of optimal modeling of private telecommunication networks and services.

Michal HALÁS, Ing., PhD. was born in Bratislava, Slovakia in December 1978. He graduated from the Slovak University of Technology and received the electronic engineering degree in 2003. Since this year he has started postgraduate study at the Department of Telecommunications STU Bratislava. In 2006 he submitted PhD work from the field of Delay in VoIP networks. Nowadays he works as a lecturer at the Department of Telecommunications, FEI STU in Bratislava.