Vladimir MATIĆ, Aleksandar LEBL, Dragan MITIĆ, Miroslav DUKIĆ

Institute for Telecommunications and Electronics IRITEL A.D. BELGRADE, Faculty of Electrical Engineering Belgrade

Estimation of Post Dialling Delay in Telephone Networks

Abstract. *In this paper we present that Post Dialling Delay in one telephone network depends on the service discipline in particular network nodes. It is indicated that deviation of PDD from its mean value decreases as the number of nodes included in one established connection increases. The consequence of this fact is that the condition of serving 95% of connections is more stringent in the connections with small number of network nodes included in the connection, while the condition dealing with the mean value of PDD is more stringent in the connections with great number of nodes.*

Streszczenie. *W artykule analizowane jest opóźnienie połączenia w sieciach telefonicznych w zależności od liczby węzłów i liczby połączeń. (**Określenie opóźnienia połączenia w sieciach telefonicznych**)*

Keywords: Telephone signalling, Post dialling delay, telephone networks
Słowa kluczowe: sieci telefoniczne, opóźnienie połaczenia.

Introduction

The speed of telephone network reaction is, besides voice quality, the main indicator of Quality of Service (QoS). The speed of network reaction is expressed by the time while the activities, such as connection realization to the called side, transferring the signal of called subscriber answer, connection disconnect, etc., are accomplished. The speed of these activities is measured by the time interval needed for their realization, and this time intervals are called Post Dialing Delay, Answer Signal Delay, Call Release Delay, etc. Among these indicators of function speed, it seems that the most important one is Post Dialing Delay (PDD), which is also called Post Selection Delay. It is the key indicator of speed of the connection realization for two reasons. This phase is technically the most complicated one (1), and in this phase the user expects the positive answer, i.e. he is prone to abandon from the attempt to establish the connection (2). The greatest values of this time are suggested in Recommendation [1]. These values are determined for ISDN network, but they are also applied in mixed, [2], and packet networks [3]. In this paper we try to indicate the possibility to estimate PDD in depending on the type of service in network nodes.

Definition and recommendations

For all types of networks, PDD can be defined as the time interval from the end of sending user signalling address information till the beginning of the called side response, Fig. 1. This definition is detailed for some networks, where PDD is defined as the time interval between specific signalling messages (from the message SETUP till the message ALERTING in ISDN) or methods (from the method INVITE till the answer 180 RINGING in SIP signalling, [4]). We adopt that PDD in SIP signalling ends by the preliminary message 180 RINGING, rather than the preliminary message 100 TRYING, as stated in [4], or the final message 200 OK, as stated in [5].

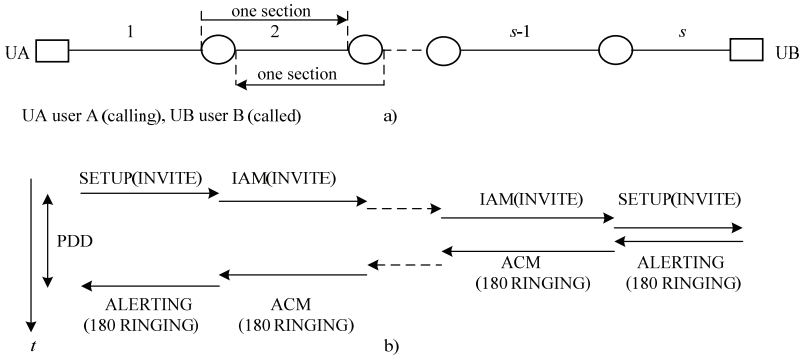

Fig.1. Presentation of the PDD time interval

In order to achieve the good service for the user, the recommendations prescribe the values of PDD. These recommendations determine: the longest mean time of PDD (t_{mPDD}) and the longest PDD duration for which 95% of connections will receive the answer from the called side (t_{95PDD}).

The recommendations are defined for local, transit and international connections. In [1] and [2] it is stated that t_{mPDD}≤3s, 5s and 8s, and t_{95PDD}≤6s, 8s and 11s for local, transit and international connections (respectively) and normal traffic load (load A). For the increased traffic load (load B) the values t_{mPDD}≤4.5s, 7.5s and 12s, and t_{95PDD}≤9s, 12s and 16.5s for local, transit and international connections (respectively) are recommended. In [1] it is stated that typical local, transit and international connection passes through 1-4, 5-7 and 8-10 network nodes, respectively. It can be noticed that the ratio of the greatest recommended values t_{95PDD} and t_{mPDD} is t_{95PDD}/t_{mPDD}=2 (for local connections), t_{95PDD}/t_{mPDD}=1.6 (for transit connections), t_{95PDD}/t_{mPDD}=1.375 (for international connections).

What elements constitute PDD

Interval PDD can be presented as the sum of $2 \cdot s$ time intervals, concerning one network node and information transfer to the next network node. These time intervals exist for the transfer from the calling to the called user, and vice versa, Fig. 1. Actions performed in one node and on one link (i.e. on one section) in literature are called subcall or call segment, [6]. The time needed for execution of one call segment is random variable, which depends on the type of

operation in network node and on link. In some nodes it is necessary to access the base. It can be said that the probability distribution of one call segment duration time depends on the service discipline in the node and on the number of service elements in section. In the next section we shall see what main distributions could exist.

Servicing in one section

1. Servicing by one server

Let us suppose that the time interval, which address signalling message spends in network node and on the link towards the next network node, is caused by only one bottleneck, i.e. by waiting time on processor service of signalling message or by the waiting on sending over link. The time interval of signalling message processing (service time, t_s) is either constant, or is negative exponentially distributed. (Precisely speaking, the time interval spent in the node is the sum of service time and the waiting time. In the periods of heavy load, which are especially interesting for us, service time can be neglected in relation to waiting time). We shall consider service according to FIFO discipline. In the first case (constant service time), the queueing system is M/D/1, and in the second case (negative exponentially distributed service time), queueing system is M/M/1. It means that all serving resources except one are over-dimensioned, i.e. they provide service without significant waiting. The waiting is caused by offered traffic A on only one server and can be presented by random variable $\boldsymbol{T_D}$ or $\boldsymbol{T_M}$. As it is well known and explained in [7], the mean waiting time, t_{mD}, and the probability of waiting longer than t, $P(T_D>t)$, is in the case of M/D/1 system:

(1) $$t_{mD} = \frac{t_s \cdot A}{2 \cdot (1 - A)}$$)

(2) $$P(T_D > t) = 1 - (1 - A) \cdot \sum_{i=0}^{l} \frac{(-A \cdot (\frac{t}{t_s} - i))^i}{i!} \cdot e^{A \cdot (\frac{t}{t_s} - i)}$$

where l is the integer part of the ratio t/t_s.

In the case of M/M/1 system, the mean waiting time t_{mM} and the probability of waiting $P(T_M>t)$ are:

(3) $$t_{mM} = \frac{t_s \cdot A}{(1 - A)}$$

(4) $$P(T_M > t) = A \cdot e^{-(1-A) \cdot \frac{t}{t_s}}$$

The characteristics of serving by one server are:

- probability distribution function of waiting time is exponential;
- mean waiting time is shorter in the case of M/D/1 model than in the case of M/M/1 model, but model M/M/1 appears more often in practical systems;
- the ratio t_{95}/t_m for both models is about 3, and, according to [7], the more stringent criterion, which is recommended in [8] and [9], is the one dealing with t_{95}.

Fig. 2. presents cumulative distribution function in these two systems for traffic load A=0.9. On the x axis the service time t_s is used as the unit of time.

Therefore, if the delay in one network node is the consequence of the service in only one element (processor, link), then exponential distribution is valid for cumulative distribution function. The ratio of t_{95} to t_m is about 3, but the absolute values for constant service time are significantly lower. Unfortunately, constant service time is rare in the practice. We can note that, in this case, is more difficult to satisfy the criterion t_{95} than the criterion t_m, [8].

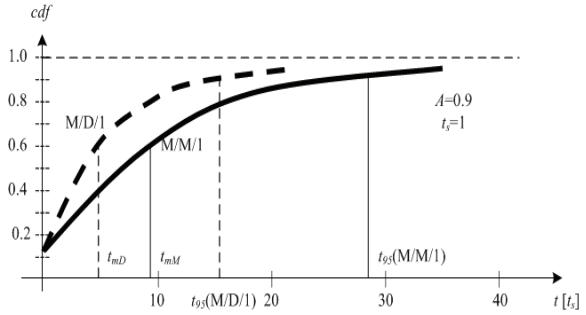

Fig. 2. Cumulative distribution function of service time for M/D/1 and M/M/1 system when A=0.9

If link is the main reason of waiting, then the service time is proportional to the duration of the packets, which are sent over the link. If the signalling processor is the element, which causes waiting, then the dispersion of service time for particular signalling messages can be greater than the dispersion of the duration of particular messages.

2. Multiphase servicing in one section

Let us consider one network node and one link, i.e. one section in the case that congestion can happen on several elements: signalling processor, data base, sending buffer. In this case, the time, which signalling information spends in the considered node and link is the sum of service time and waiting time in each element of service in that node (link). In general circumstances, it is very complicate to calculate this time. That's why two assumptions are adopted. The first one is that the service of the same kind is performed in each service system of one section, and the second one is that the time spent on one section (node + link) is the random variable, which is the sum of k random variables representing the time spent for particular phases in the node and on the link, where k is the number of service phases on one section (k=2, 3, 4…). Let us suppose that in each service phase of one section the time spent in that phase is random variable with exponential distribution of the time duration and mean value $1/\lambda$. In each phase a lot of different, independent of each other, requests are serviced. That's why the distribution of total time $\boldsymbol{T_E}$, spent on one section with k phases, is well known Erlang-k distribution, expressed by the equation

(5) $$P_E(T_E > t) = \sum_{i=0}^{k-1} e^{-\lambda \cdot t} \cdot \frac{(\lambda \cdot t)^i}{i!}$$

The mean value t_{mE} of random variable T_E for which holds Erlang-k distribution is k/λ.

For this distribution, there exists one value t_{95E} that holds

$$1 - P_E(T_E > t_{95E}) = P_E(T_E \le t_{95E}) \ge 0,95$$

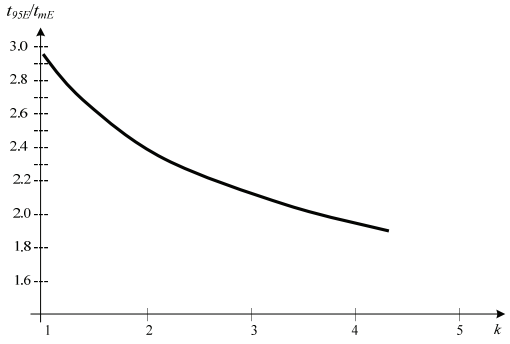

Fig. 3. Ratio $\boldsymbol{t_{95E}/t_{mE}}$ in the function of number of service phases on one section

Fig. 3. presents the ratio t_{95E}/t_{mE} for the total time spent on one section as the function of the number of service phases.

Comparing the cases 1. (one phase) and 2. (multiphase) when the mean time spent on one section is same, it can be said that the service in few phases is more favourable, because the dispersion, i.e. deviation of the time spent on one section from its mean value is less than in case of service on one channel.

Total duration of time interval PDD

Time interval PDD is, obviously, random variable, that is the sum of mutually independent random values of call segment durations. It must be indicated that PDD is constituted of the call segments of the messages sent forward (e.g. SETUP, IAM (Initial Address Message), INVITE) and backward (e.g. ALERT, ACM (Address Complete Message), 180 RINGING), i.e. $2 \cdot s$ segments, Fig. 1. (However, PDD is mainly constituted of sending messages forward, because these messages are more complicate and require the decision about forwarding and acceptance in each node). Therefore, the characteristics of PDD will be the same as the characteristics of random variable, which is the sum of $2 \cdot s$ independent random variables. As is well known, the mean time PDD will be equal to the sum of all mean durations of particular call segments, i.e.

$$t_{mPDD} = \sum_{i=1}^{2 \cdot s} t_{mi} \tag{6}$$

What will be the duration t_{95} of PDD, i.e. t_{95PDD}? In general situation the calculation of this value will be very complicate. But, as we know that

$$\sigma^2{}_{PDD} = \sum_{i=1}^{2 \cdot s} \sigma_i{}^2 \tag{7}$$

and starting from the fact that is for each distribution the time interval t_{95} proportional (~) to the standard deviation σ ($t_{95} \sim \sigma$), we can conclude that t_{95PDD} is proportional to the square root of the number of sections, i.e.

$$t_{95PDD} \sim \sqrt{2 \cdot s} \tag{8}$$

When the number of sections increases, the ratio t_{95PDD}/t_{mPDD} decreases, because it is

$$\frac{t_{95PDD}}{t_{mPDD}} \sim \frac{\sqrt{2 \cdot s}}{2 \cdot s} = \frac{1}{\sqrt{2 \cdot s}}, s > 1 \tag{9}$$

It can be concluded that the deviation of time interval PDD from its mean value is smaller if the number of segments is greater. If in each network section the processing of signalling information is subject of multistep service (subsection 2.), this effect is even more expressed.

Conclusions

PDD is the time interval calculated from sending address information towards called user till receiving the answer from the called user side. The answers from the particular points of the network can't be considered.

PDD is defined by its mean value t_{mPDD} and the time t_{95PDD} while the answer is received from the called side in 95% occurrences. The recommendations dealing with PDD also consider these two time intervals. It can be concluded that the mean value t_{mPDD} increases proportionally to the number of sections in the network, which signalling message passes. Comparing to the increasing of mean value of PDD, the deviation of PDD from its mean value increases slower as the number of sections increases.

This increasing of PDD deviation is greatest for local connections (2 to 8 sections), and smallest for international connections (16 to 20 sections).

The increasing of PDD deviation from its mean value is additionally reduced if in network sections multistep (multiphase) service is performed.

In the connections with small number of sections it is more difficult to satisfy the recommendations dealing with the greatest allowed value of t_{95PDD}. In opposite, in the case of great number of sections, the recommendations dealing with the greatest value of t_{mPDD} become more stringent.

REFERENCES

[1] ITU-T: Recommendation E.721 - Network grade of service parameters and target values for circuit-switched services in the evolving ISDN, May 1999

[2] ITU-T: Recommendation E.671 - Post-selection delay in PSTN/ISDN networks using Internet telephony for a portion of the connection, March 2000

[3] Arjona, A., Westphal, C., Yla-Jaaski, A., Kristensson, M. and Manner, J.: Towards High Quality VoIP in 3G Networks an Empirical Approach, Int. J. of Communications, Network and System Sciences, 1(2008), No 4, pp 350-361

[4] Eyers, T. and Schulzrinne, H.: Predicting Internet Telephony Call Setup Delay, Proc. IP Telephony Workshop, April 2000

[5] Pack, S. and Lee, H.: Call Setup Latency Analysis in SIP-Based Voice over WLANs, IEEE Communications Letters, 12(2008), No 2, pp. 103-105

[6] Lin, H., Seth, T., Broscius, A. and Huitema, C.: VoIP Signaling Performance Requirements and Expectations, Internet draft, IETF, October 1999

[7] Markov, Ž. and Manević, I.: Determination of More Stringent Criterion for Common Control Unit of Digital Telephone Exchange, Int. J. Electron. Commun. (AEU) 52(1998), No 2, pp 101-103

[8] ITU-T: Recommendation Q.543 - Digital exchange design objectives, March 1993

[9] ITU-T: Recommendation Q.725 - Signaling System No7 – signaling performance in the telephone application, March 1993

***Authors**: mr Vladimir Matić, dipl. Ing., IRITEL A.D., Batajnički put 23, 11080 Belgrade, Serbia, (phone 381-11-3073485; e-mail: vmatic@iritel.com; dr Aleksandar Lebl dipl.ing., IRITEL A.D., Batajnički put 23, 11080 Belgrade, Serbia, (phone 381-11-3073422; e-mail: lebl@iritel.com; dr Dragan Mitić dipl.ing., IRITEL A.D., Batajnički put 23, 11080 Belgrade, Serbia, phone 381-11-3073420; e-mail: mita@iritel.com; prof. dr Miroslav Dukić dipl.ing., Faculty of Electrical Engineering, Bulevar Kralja Aleksandra 73, 11000 Belgrade, Serbia, phone 381-11-3370106; e-mail: dukic@etf.rs.*