Research of the Process of Securing of QoS in Integral Networks Based on IP

Vassil M. Kadrev[1] and Nikolai V. Penev[2]

Abstract – **The construction of integral communication networks based on IP is especially important because of the high degree of using the lines and nodes of the network compared with TDM and ATM. In order the IP network to play the role of an integral network there must be special mechanisms for guaranteeing of QoS in it which in the other networks is achieved by establishing a connection or contracting in advance the speed of the established channel. This problem in particular is the object of the research.**

Keywords – **IP, QoS, Traffic**

I. Formulation of the Problem

The construction of integral communication networks is especially important because of the higher degree of utilization of the lines and the nodes in the network of the summed subscriber traffic. The quality of services QoS is determined by two factors: losses and delay, standardized for each type of traffic.

Information losses because of errors in the physical transfer environment, the processors and the memories, as well as of procedure errors in the prioritising protocols can be reduced to a minimum with the development of the technologies and the methods for verification of protocols [1], [2], [3]. The second factor, however, the traffic delay, must be always considered, especially in integral IP network with different services (multimedia, voice, data) and corresponding norms for delay for each of them.

The keeping of QoS in the IP network is realized by its prioritising protocols performing their functions in different levels of OSI (802.1p at layer 2, TOS and DiffSurf at layer 3, TCP and UDP at layer 4, as well as prioritising of layer 7). The process of securing of QoS is based on many parametres, assigned to the messages from different layers - the MAC address and 802.1p-labels assigned to the frames of layer 2; IP addresses, TOS labels and Diff Surf labels, assigned to the packets of layer 3; numbers of ports in headers of TCP and UDP datagrams of layer 4; user identity, user location and time of day in layer 7.

The prioritising of the messages of different layers guarantees a minimal delay end to end for the traffic with high priority. By the listed parametres different disciplines of prioritising which can be realized by the listed above protocols of each layer can be supported. For example they are: 1) a discipline with static priorities; 2) a discipline with dynamic priorities depending on the waiting time of the message (frame, packet, datagram) in the buffer of the network node, increasing with the waiting time in the buffer lineally with higher slope with higher static priority; 3) a discipline with dynamic priority with an increase of the priority with the decrease of the difference between the delay norm and the accumulated by that time delay and with considering of the remaining route and the current stay of the network, etc.

Every discipline is characterized by the following parametres: the value of the traffic delay with different priority, interruption or not of the low-priority order by a high-priority one, the value of the signalisation traffic for realizing of the discipline, the value of the processor time for doing the procedure by the network node.

Discipline 1 characterized with small signal traffic (only the number of the static priority is transmitted) and short processor time due to its simplicity (a check of the static priority and buffering in the corresponding queue). Discipline 2 does not increase the signal traffic but the processor time of node increases. Discipline 3 increases the signal traffic (the number of the static priority, the moment of generating the message and a message for synchronization of the network nodes are transmitted) and the processor time of the node because of the many additional parametres: number of retranslation sections to the receiver, existence of bottlenecks on the route in the direction of the receiver etc.

Because the priority discipline is introduced to guarantee QoS for traffics with different priority it must be quantitatively assessed by the delay for the different services.

II. Quantitative Ratios And Results

The present investigation does not concern discipline 3, which is about the whole structure of the network and its current state.

Discipline 1 can be described by using of a priority system for mass servicing (which interprets the switch in the network node) with static priorities, with or without interruption of the low-priority order.

Because the priority discipline with interruption of the low-priority order complicates the processing of the messages in the network nodes it is natural to carry out an investigation for the simpler discipline without interruption and only if it does not satisfy the requirements for QoS then an investigation with interruption of low-priority order must be carry out, as is the case in discipline 2.

Common for all listed mechanisms for control of IP traffic is the multi-priority traffic of the arriving requests for services. For an IP network with integrated services including

[1] Vassil M. Kadrev is with the Higher School of Transport Engineering, Geo Milev 158, 1754 Sofia, Bulgaria, E-mail: kadrev@vtu.acad.bg
[2] Nikolai V. Penev is with the War Academy, E. Georgiev 82, Sofia, Bulgaria, E-mail: penevnv@yahoo.com

subscribers and transit network nodes the delay of the packets end to end (for each service) is defined as T_w:

$$T_w = 2.t_{SL} + \sum_{i=1}^{d}\left(t_w + t_s + t_T\right) - t_T\,, \tag{1}$$

where:

- t_{SL} – average time for the transfer of the packet by the subscriber line;
- d – average number of transits;
- t_w – average time for waiting of the packet in the queues;
- t_s – average time for processing of the packet in one router;
- t_T – average time for transfer of the packet by the trunk.

The components of T_w: t_{SL} и t_T, which depend on technologies can be taken as constants and t_w and d remain to influence T_w.

In the present work the studied mechanism of servicing by which in accordance with the strict requirements for work in real time a guaranteed frequency band, small delay of the packets end to end and lack of loss of packets as a result of their arranging in queues are provided is modelled. Because of that, in servicing of a common traffic flow every router in the network (which requires separate functions in the network control for this) must distribute the frequency band and the necessary buffer space in accordance with the priority of the arriving packets from different services. Because a great number of traffic flows on lines with high capacity arrive at the network node, it can be assumed that the input traffic for the network node is Poissonian [2]. The separate router is a system for a mass servicing with waiting and a priority servicing discipline. Under these conditions T_w for each service can be determined. It is known that the teletraffic system M/D/1 is characterized by average waiting time twice as short as the system M/M/1. The investigation will be carry out with using of the teletraffic system M/M/1 on the condition that the results from the determination of the average waiting time for a system for mass servicing M/M/1 are the upper limit for the average waiting time for servicing system M/D/1 as it is in practice. It is possible to find out correlation between QoS and speed of lines, the packet size and the number of transit sections.

The determining of the results about the influence of the priority servicing on the QoS for one servicing device can be used also for the network as a whole.

The position of every packet in the queue of the servicing device will be a variable in time because a packet of higher priority may enter the queue. The priority system may be, on the one hand, dependent on whether the priority is absolute (fixed) or on a given function, and the other hand, whether the processing of the serviced packets is interrupted at the moment of arrival of another packet with a higher priority and is then restored (from the moment of interruption).

For a priority servicing system (with P priorities, where $p = 1, 2, 3, \dots P$) for the delay time of the messages from every priority (with $0 \le \rho < 1$, i.e. lack of losses) [4]:

$$T_{w\,p} = \begin{cases} \dfrac{\dfrac{\rho_p}{\mu_p} + \sum\limits_{i=p+1}^{P}\rho_i\cdot\left(\dfrac{1}{\mu_p} + \dfrac{1}{\mu_i}\right) + \sum\limits_{i=p+1}^{P}\rho_i T_i}{1 - \sum\limits_{i=p}^{P}\rho_i}, & p \ge j\,, \\ \infty, & p < j \end{cases} \tag{2}$$

where: j is a minimal positive integer number for which (relating of the arriving traffic ρ) : $\sum\limits_{i=j}^{P}\rho_i < 1$; μ – intensity of servicing.

We accept that traffic with different proportions of the following kinds of services is processed [1]: multimedia, voice, interactive data, downloads.

The average servicing time $1/\mu$ depends on the average length of the packets l and the average transfer speed on the line c. The average value ρ of the arriving traffic depends on the average speed of arriving of the packets λ and the average processing time $1/\mu = const$. We will consider that the system is in stationary mode and that for the arriving traffic $\sum\limits_{i=1}^{4}\rho_i = \rho < 1$.

The priority of servicing increases with the increase of p. We accept the speeds of transfer on connection line c_T are: 64 *kbps*, 128 *kbps*, 2 *Mbps*, 8 *Mbps*, 32 *Mbps*, 155 *Mbps*, 622 *Mbps*, 2400 *Mbps*; the average number of transit section is $d = 15$ or 30; the processing time of the packet in a network node t_s and the transfer time on a subscriber line are negligible constants.

The parametres of the arriving traffic and of the processing [1] are shown in Table I.

The average admissible waiting time in the queue of one processing device is the difference between the average admissible delay for each service and the total average transmission time end to end with d sections.

The results obtained for the delay values depending on the input traffic ρ, the number of the hopes d, the length of the packets l and at concrete proportion of the input traffic from the separated services with priority processing (Table II) are shown on Figs 1, 2, 3 and 4.

TABLE I
TYPE OF TRAFFIC SOURCES AND PARAMETERS OF SERVICING

Type of traffic sources	Multimedia	Voice	Interactive data	Download
Parameters of servicing	4	3	2	1
Priority of servicing, *p*	4	3	2	1
Norm of delay, *ms*	t_4	t_3	t_2	t_1
Average admissible time for one hope, *ms*	t_4/d	t_3/d	t_2/d	t_1/d
Average length of PDU2, *kbit*	l_{24}	l_{23}	l_{22}	l_{21}
Average time for transfer by trunk, *ms*	t_{t4}	t_{t3}	t_{t2}	t_{t1}
Average time for transfer by subscriber line, *ms*	t_{s4}	t_{s3}	t_{s2}	t_{s1}

Average servicing time, $1/\mu = const$	$\frac{1}{\mu_4}=\frac{l_4}{c_4}$	$\frac{1}{\mu_3}=\frac{l_3}{c_3}$	$\frac{1}{\mu_2}=\frac{l_2}{c_2}$	$\frac{1}{\mu_1}=\frac{l_1}{c_1}$
Total average transfer time end to end, *ms*	$2t_{s4}+(d-1)t_{t4}$	$2t_{s3}+(d-1)t_{t3}$	$2t_{s2}+(d-1)t_{t2}$	$2t_{s1}+(d-1)t_{t1}$
Average admissible waiting time in the queue of a node, *ms*	$t_{w4}=t_4-1/\mu_4$	$t_{w3}=t_3-1/\mu_3$	$t_{w2}=t_2-1/\mu_2$	$t_{w1}=t_1-1/\mu_1$

TABLE II

VALUES OF OUTPUT DATA FOR DETERMINING OF THE SERVICING DELAYS

	Multimedia 4	Voice 3	Interactive data 2	Download 1
Percentage of the arriving traffic, %	20	10	35	35
Proportion of the input traffic ρ	0,2 – Figs. 1 and 3; 0,9 – Figs. 2 and 4.			
Number of hops	15 – Figs. 1 and 3; 30 – Figs. 2 and 4.			
Length of the packets, *bits*	8000 – Figs. 1 and 2; 32000 – Figs. 3 and 4.			
Waiting time, *s*	w4	w3	w2	w1

III. CONCLUSION

From the results obtained it can be seen that with the given parametres of the lines, devices, and the traffic (the size and the proportions of the arriving traffic - the loading of the router and the size of the packets) the delay which the separate services receive is within the norms for a wide range of change for the values of these parametres. It is so in using of lines with speed above 2 *Mbps* and quality servicing of the arriving traffic for all kinds of services.

By increasing the load from 0,2 to 0,9 it is impossible to achieve quality servicing of low-priority traffic (on Figs 2 and 4 it is missing because it is not within the norms). Similarly with the increase of the length of the packets from 8000 to 32000 bits for quality servicing of the integrated traffic using of faster lines is necessary - 32 *Mbps* (Figs. 3 and 4).

As a general effect of the increase of the loading of the network nodes and the length of the packets, the decrease of the reserve of admissible delay in the queues for the traffic of the separated services in accordance with the norms for delay end to end can be noted.

REFERENCES

[1]. Parvanova N. Broadband ISDN. NIIS-CENTI, Sofia, 1997.

[2]. Alcatel Telecommunications Review, No. 2, 1999.

[3]. Boyanov K. and all Computers networks and Internet. Sofia, CLPOI-BAS, 1998.

[4]. Kleinrock L. Communication Nets (stochastic message flow and delay). Mcgraw-Hill.

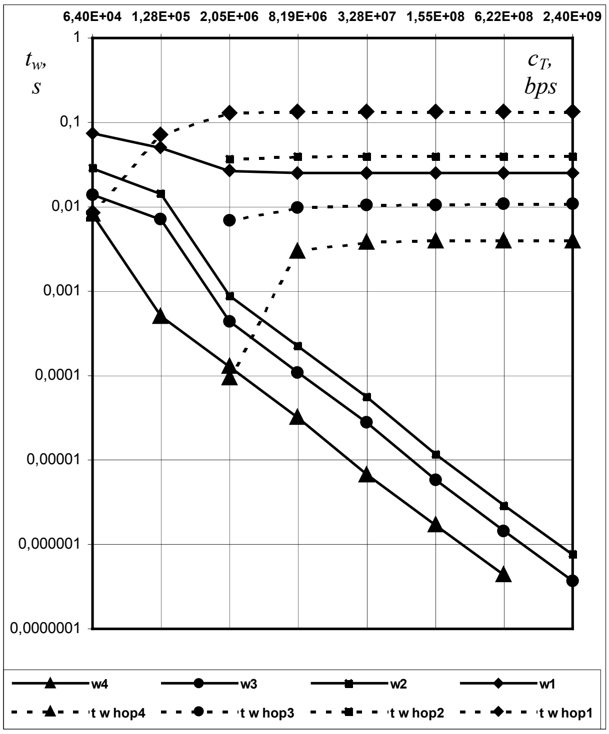

Fig. 1. Values of the processing delays and the delay norm for the different services for ρ = 0,2, number of hops = 15, length of the packets = 8000.

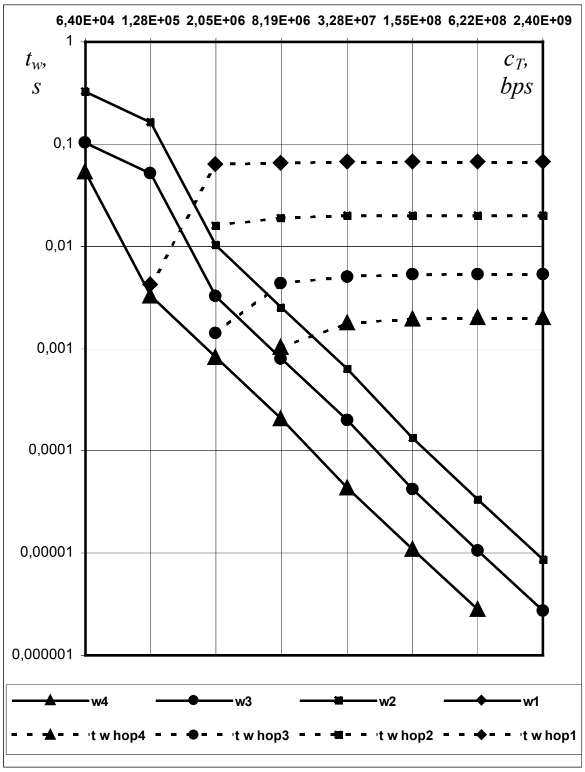

Fig. 2. Values of the processing delays and the delay norm for the different services for ρ = 0,9, number of hops = 30, length of the packets = 8000.

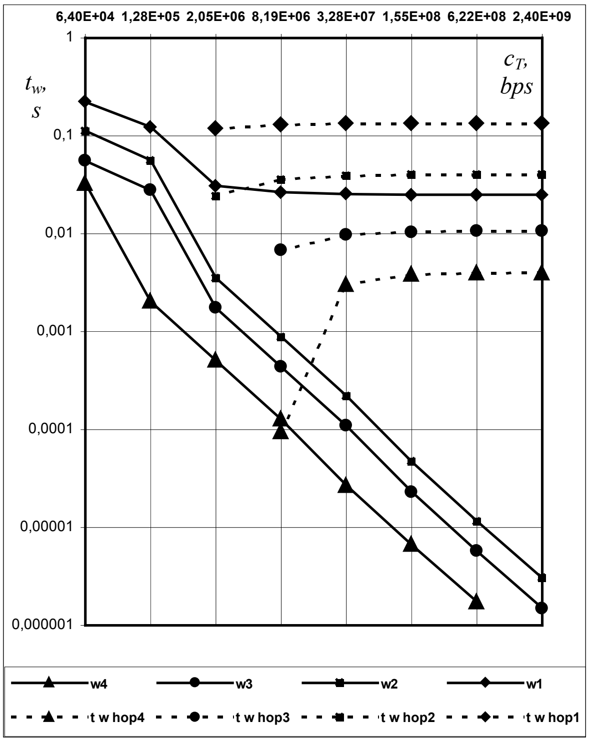

Fig. 3. Values of the processing delays and the delay norm for the different services for ρ = 0,2, number of hops = 15, length of the packets = 32000.

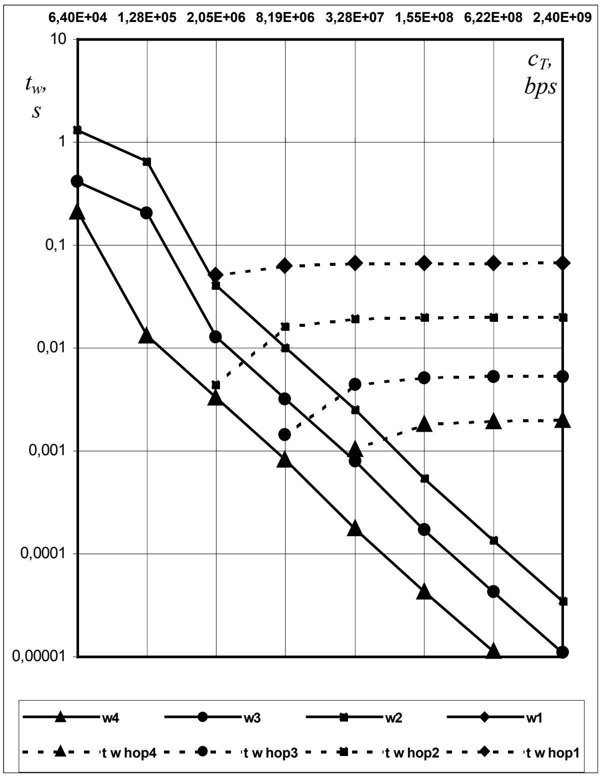

Fig. 4. Values of the processing delays and the delay norm for the different services for ρ = 0,9, number of hops = 30, length of the packets = 32000.