Authors' Information

Rossitza Iv. Goleva – Assistant-Professor; Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: rig@tu-sofia.bg

Mariya At. Goleva – Student; Department of Communication Networks, University of Bremen, Germany; e-mail: mgoleva@gmail.com

Dimitar K. Atamian - Assistant-Professor; Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: dka@tu-sofia.bg

Tashko Nikolov – Assistant-Professor, Ph.D. Department of Telecommunications, Technical University of Sofia, Bulgaria, Kl. Ohridski blvd. 8, Sofia. 1756, Bulgaria; e-mail: tan@tu-sofia.bg

Kostadin At. Golev – Developer; Bianor Ltd., Bulgaria; e-mail: kotseto@gmail.com

STUDY OF QUEUEING BEHAVIOUR IN IP BUFFERS

Seferin Mirtchev

Abstract: It is unquestioned that the importance of IP network will further increase and that it will serve as a platform for more and more services, requiring different types and degrees of service quality. Modern architectures and protocols are being standardized, which aims at guaranteeing the quality of service delivered to users. In this paper, we investigate the queueing behaviour found in IP output buffers. This queueing increases because multiple streams of packets with different length are being multiplexed together. We develop balance equations for the state of the system, from which we derive packet loss and delay results. To analyze these types of behaviour, we study the discrete-time version of the "classical" queue model M/M/1/k called Geo/Gx/1/k, where Gx denotes a different packet length distribution defined on a range between a minimum and maximum value.

Keywords: delay system, queueing analyses, discrete time queue, IP traffic modelling; packet size distribution.

ACM Classification Keywords: G.3 Probability and statistics: queueing theory, I.6.5 Model development

Introduction

The initial motivation for this paper is the necessity of traffic engineering in IP networks. Many analyses of Internet traffic behaviour require accurate knowledge of the traffic characteristics for purposes ranging from a management of the network quality of service to modelling the effect of new protocols on the existing traffic mix.

Modern architectures and protocols are being standardized, which aims at guaranteeing the quality of service delivered to users. The proper functioning of these protocols requires an increasingly detailed knowledge for statistical characteristics of IP packets. The amount of information flowing through the network also increases, and the challenge is to obtain the accurate information from a huge set of data packets.

The packet queueing in an IP router arises because multiple streams of packets from different input ports are being multiplexed together over the same output port. A key characteristic is that the packets have different length. The minimum header size in IPv4 is 20 octets, and in IPv6, it is 40 octets. The maximum packet size depends on the specific sub-networks technology: 1500 octets in Ethernet and 1000 octets are common in X.25 networks. The packet length distribution measured from the real traces exhibits the well-known multi-mode behaviour, with peaks for very short packets and for the different maximum transfer units in the network, with a dominating peak at 1500 bytes, due to the size of Ethernet frame. This specific packet length distribution has a direct impact on the service time and we need a different approach to the queueing analysis.

Discrete-time queueing systems have been a research topic for several decades now and there are many reference works on discrete-time queueing theory. Over the years, different methodologies have been developed to assess the performance of queueing systems. The two main analytical approaches are the matrix analytic method and the transform method for discrete and for continuous-time analyses. Many authors have considered the Geo/G/1 queueing system [Pitts, 2000], [Mirtchev, 2006], [Vicari, 1996], [Zang, 2001].

In [Atencia, 2005] is carried out a complete study of a discrete-time single-server queue with geometrical arrivals of both positive and negative customers. Negative arrivals are used as a control mechanism in many telecommunication and computer networks. [Atencia, 2006] is concerned with the study of a discrete-time single-server retrial queue with geometrical inter-arrival times and a phase-type service process. An iterative algorithm to calculate the stationary distribution of Markov chain is given.

[Salvador, 2004] is proposed a traffic model and a parameter fitting procedure that are capable of achieving accurate prediction of the queuing behaviour for IP traffic exhibiting long-range dependence. The modelling process is a discrete-time batch Markovian arrival process (dBMAP) that jointly characterizes the packet arrival process and the packet size distribution. In the proposed dBMAP, packet arrivals occur according to a discrete-time Markov modulated Poisson process (dMMPP) and each arrival is characterized by a packet size with a general distribution that may depend on the phase of the dMMPP.

[Cao, 2004] is presented an introduction to bandwidth estimation and a solution to the problem of the best-effort traffic for the case where the quality criteria specify negligible packet loss. The solution is a simple statistical model, which is built and validated using queueing theory and extensive empirical study.

It has been shown [Dan, 2005] that in the case of real-time communications, for which small buffers are used for delay reasons, short range dependence dominates the loss process and so the Markov-modulated Poisson process (MMPP) might be a reasonable source model. They have presented an exact mathematical model for the loss process of a MMPP+M/Ek/1/K queue and have concluded that the packet size distribution affects the packet loss process and thus the efficiency of forward error correction.

In this paper, we investigate the basic queueing behaviour of packets found in IP output buffers. This queueing is complicated because multiple streams of packets are being multiplexed together. The traffic is being generated from the packets of varying sizes that arrive for transmission on the link. The packets can queue up and loss if their size is bigger than the free positions of the buffer. The quality metrics for the best-effort traffic on the Internet are the packets loss and delay. To analyze these types of behaviour, we study the discrete-time version of the “classical” queue model M/M/1/k called Geo/Gx/1/k, where Gx denotes a different packet length distribution. We developed balance equations for the state of the system, from which we derived packets loss and delay.

Balance equations for the queue model Geo/Gx/1/k

Let us consider a single server finite queue delay system *Geo/Gx/1/k* with a geometric distributed inter-arrival time and different distributions of the packet length: truncated geometric, binomial, discrete uniform and discrete triangular. These packet length distributions are defined on a range between a minimum and maximum value.

We consider queueing phenomena in discrete-time queueing systems. That is, we assume a fundamental time unit (time slot), the time to transmit an octet (byte), T_b. Customers arrive in the queueing system under consideration during the consecutive slots, but they can only start service at the beginning of slots. That is, service of customers is synchronized with respect to slot boundaries. Further, customer service times are integer multiples of the slot length, which implies that customers leave the system at slot boundaries. During the consecutive slots, packets arrive in the system, are stored in a finite capacity queue and are served by a single server on a first in first out (FIFO) basis (fig.1).

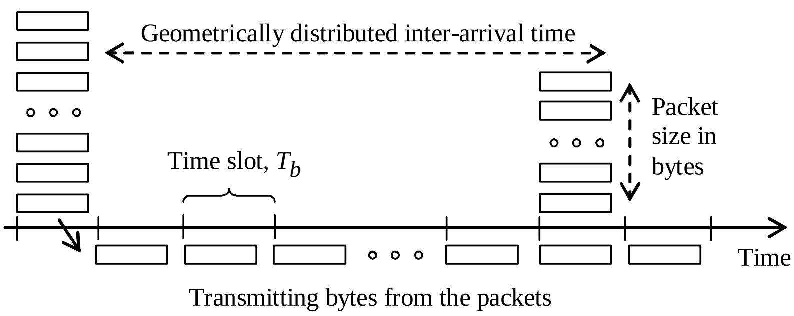

Fig.1. Timing of events in the Geo/Geo/1/k queueing system

We use a Bernoulli process for the packet arrivals, i.e. a geometrically distributed number of slots between arrivals. Let the probability that a packet arrives in an octet slot is *p*.

In this model, we assume a truncated geometric distribution at variable packet sizes with a minimum value m_1 and a maximum value m_2, as the first kind of distribution.

Let the probability that a packet completes service at the end of an octet slot is q. We define the probability that the packet size is n octets:

$$b_n = \frac{q(1-q)^{n-m_1}}{q\sum_{r=0}^{m_2-m_1}(1-q)^r}, \qquad m_1 \le n \le m_2 \quad . \tag{1}$$

The mean number of bytes in the packet by definition is

$$b = \sum_{i=m_1}^{m_2} i\, b_i \approx 1/q + m_1 \quad . \tag{2}$$

The second kind of a packet size distribution is binomial

$$b_n = \binom{m_2 - m_1}{n - m_1} q^{n-m_1}(1-q)^{m_2-n}, \qquad m_1 \le n \le m_2 \quad , \qquad b = m_1 + (m_2 - m_1)q \tag{3}$$

The third kind of a packet size distribution is discrete uniform

$$b_n = \frac{1}{m_2 - m_1 + 1} \quad \textit{for all values of } n, \qquad m_1 \le n \le m_2 \quad , \qquad b = (m_2 + m_1)/2 \tag{4}$$

The next kind of a packet size distribution is discrete triangular. When the mode is equal to the minimum value, we have linear decreasing distribution with the following probabilities that the packet size is n octets and the mean number of the bytes in the packet

$$b_n = \frac{m_2 - n + 1}{\sum_{r=m_1}^{m_2} m_2 - r + 1} \quad , \qquad m_1 \le n \le m_2 \quad , \qquad b = m_1 + (m_2 - m_1)/3 \tag{5}$$

When the mode is equal to the maximum value, we have linear increasing discrete triangular distribution

$$b_n = \frac{n - m_1 + 1}{\sum_{r=m_1}^{m_2} r - m_2 + 1} \quad , \qquad m_1 \le n \le m_2 \quad , \qquad b = m_1 + 2(m_2 - m_1)/3 \tag{6}$$

Thus we have a batch arrival process with geometrically distributed inter-arrival times. That is, the number of slots that separate consecutive slots where there are customer arrivals, constitute a series of independent and identically geometric distributed random variables. The probability no octets arriving in a time slot is

$$a_0 = 1 - p \quad . \tag{7}$$

The probability that n octets arriving in a time slot is

$$a_n = p\, b_n, \quad m_1 \le n \le m_2 \quad . \tag{8}$$

The mean packet service time is the octet transmission time multiplied by the mean number of octets

$$\tau = T_b \sum_{i=m_1}^{m_2} i\, b_i = T_b\, b, \ s \quad . \tag{9}$$

The mean arrival rate is

$$\lambda = p/T_b, \quad packets/s \quad . \tag{10}$$

Therefore, the offered traffic is given by

$$A = \lambda\tau = p \sum_{i=m_1}^{m_2} i\, b_i, \ erl \quad . \tag{11}$$

We define the state probability P_i of being of state ***i***, as the probability that there are ***i*** octets in the system at the end of any time slot. For the system to contain ***i*** bytes at the end of any time slots it could have contained any of *0,1,2,…..,i+1* at the end of the previous slot. State ***i*** can be reached from any of the states *0* up to ***i*** by a precise number of arrivals. To move from ***I*** *+1* to ***i*** requires that there are no arrivals.

We can write the first equation by considering all the ways in which it is possible to reach the empty state

$$P_0 = P_0 a_0 + P_1 a_0 \quad . \tag{12}$$

Similarly, we find a formula for the next state probabilities by writing the balance equations

$$P_i = P_{i+1} a_0 , \qquad 1 \le i \le m_1 - 1 \quad . \tag{13}$$

We continue with this process when the packet arrives in a time slot with length between m_1 and m_2.bytes

$$\begin{aligned}
&P_{m_1} = (P_0 + P_1) a_{m_1} + P_{m_1+1} a_0 \\
&P_{m_1+1} = (P_0 + P_1) a_{m_1+1} + P_2 a_{m_1} + P_{m_1+2} a_0 \\
&\qquad o \;\; o \;\; o \\
&P_{m_2} = (P_0 + P_1) a_{m_2} + P_2 a_{m_2-1} + \ldots + P_{m_2-m_1+1} a_{m_1} + P_{m_2+1} a_0 \\
&P_{m_2+1} = P_2 a_{m_2} + P_3 a_{m_2-1} + \ldots + P_{m_2-m_1+2} a_{m_1} + P_{m_1+2} a_0 \\
&\qquad o \;\; o \;\; o \\
&P_{k-1} = P_{k-m_2} a_{m_2} + P_{k-m_2+1} a_{m_2-1} + \ldots + P_{k-m_1} a_{m_1} + P_k a_0 \\
&P_k = P_{k-m_2+1} a_{m_2} + P_{k-m_2+2} a_{m_2-1} + \ldots + P_{k-m_1+1} a_{m_1} + P_{k+1} a_0
\end{aligned} \tag{14}$$

Then using the fact that all the state probabilities must sum to 1

$$\sum_{i=0}^{k+1} P_i = 1 \quad , \tag{15}$$

We can solve the system equations (12), (13), (14) and 15 and calculate the state probabilities.

Performance Measures

The carried traffic is equivalent to the probability that the system is busy

$$A_o = 1 - P_0 , \; erl \quad . \tag{16}$$

The packet congestion probability is the ratio of lost traffic (offered minus carried traffic) to offered traffic

$$B = \left(A - A_o\right) / A \quad . \tag{17}$$

The mean number of bytes and packets present in the system in steady state by definition is

$$L_b = \sum_{j=1}^{k+1} j P_j , \; bytes; \qquad L_p = L_b / b , \; packets \quad . \tag{18}$$

From the Little formula, we have the normalized mean system time of the bytes (time is measured in time slots)

$$\frac{W_b}{T_b} = \frac{L_b}{T_b \lambda b} = \frac{L_b}{A} \quad . \tag{19}$$

Numerical Results

In this section, we give numerical results obtained by a Pascal program on a personal computer. The described methods were tested on a computer over a wide range of arguments.

Figures 2 and 3 show the stationary probability distribution in a single server queue *Geo/Gx/1/k* with 0.8 and 0.7 erl offered traffic respectively, 1000 waiting positions, 30 bytes minimum packet length, 80 bytes maximum packet length and different packet length distributions: discrete uniform, truncated geometric, binomial, discrete triangular decreasing and discrete triangular increasing. We can see that the probability distributions are almost linear decreasing in logarithmic scale and the influence of the packet length distribution kind on the stationary probability is negligible even though in case of discrete triangular increasing packet length distribution.

Figures 4 and 5 illustrate the dependence on the packet congestion probability from the queue length when the offered traffic is 0.7 erl, the range of packet length is from 30 to 80 bytes and different packet length distributions. When the queue length is big the packet congestion probability is almost linear decreasing in logarithmic scale.

The packet length distribution in defined rage is not so essential. The main reason for this behaviour is the fact that the packet length is limited.

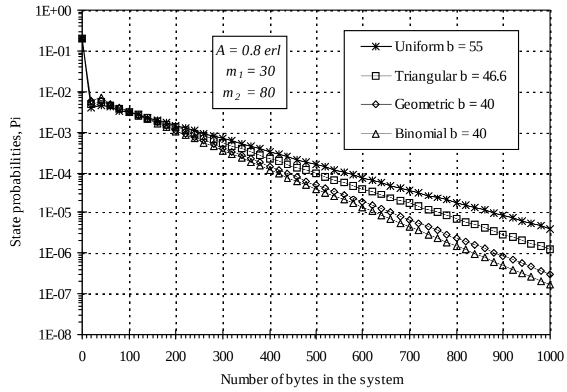

Fig.2. Graph of the state probability distributions for a finite queue with Geometric, Binomial, Uniformly and Triangular decreasing packet length distribution

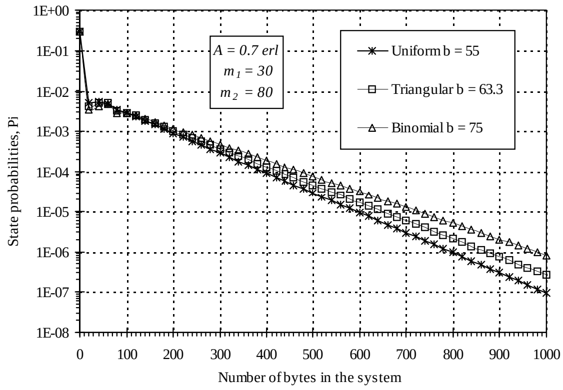

Fig.3. Graph of the state probability distributions for a finite queue with Binomial, Uniformly and Triangular increasing packet length distribution

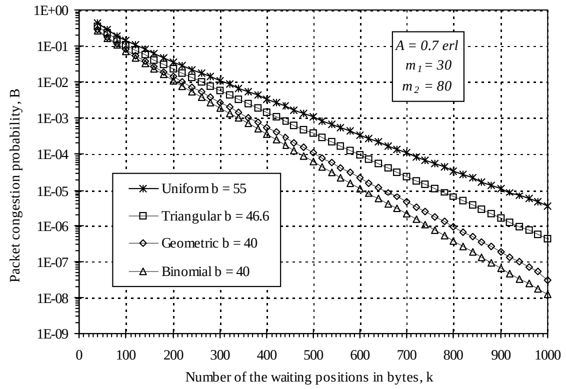

Fig.4. Packet congestion probability in the Geo/Gx/1/k with different packet length distributions and mean packet lengths between the minimum and the average value

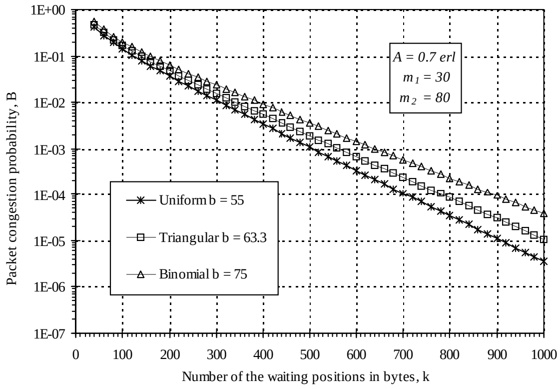

Fig.5. Packet congestion probability in the Geo/Gx/1/k with different packet length distributions and mean packet lengths between the average and the maximum value

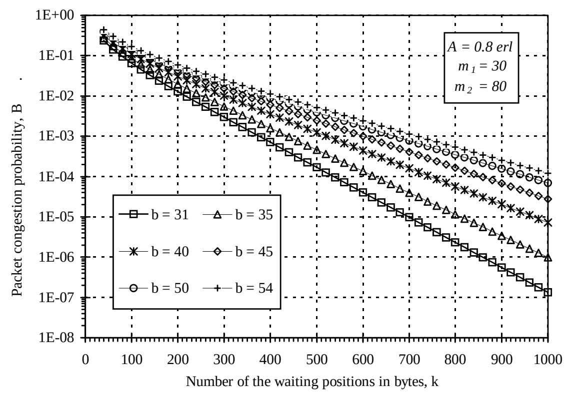

Fig.6. Packet congestion probability in discrete time single server queue with a truncated geometric packet length distribution and different mean packet lengths between the minimum and the average value

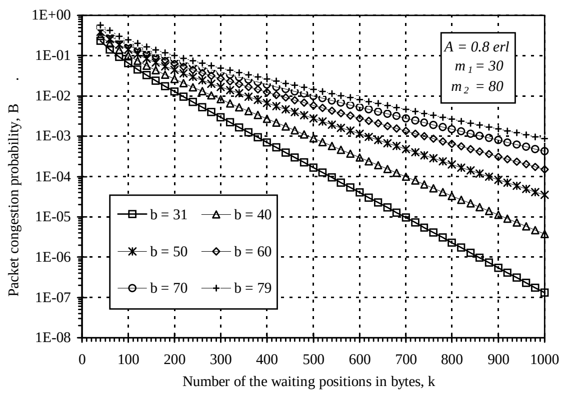

Fig.7. Packet congestion probability in discrete time single server queue with a binomial packet length distribution and different mean packet lengths between the minimum and the maximum value

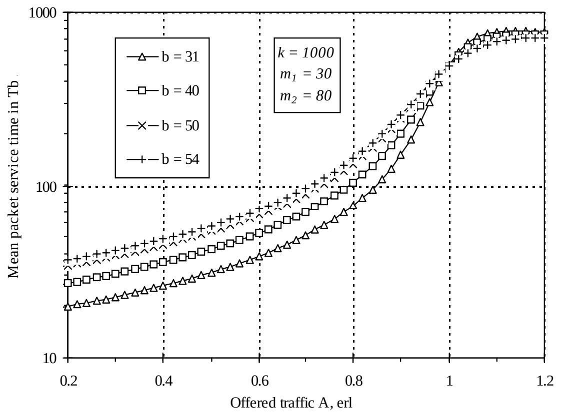

Fig.8. Normalized mean system time of the bytes in discrete time single server queue
with a truncated geometric packet length distribution and different mean packet lengths

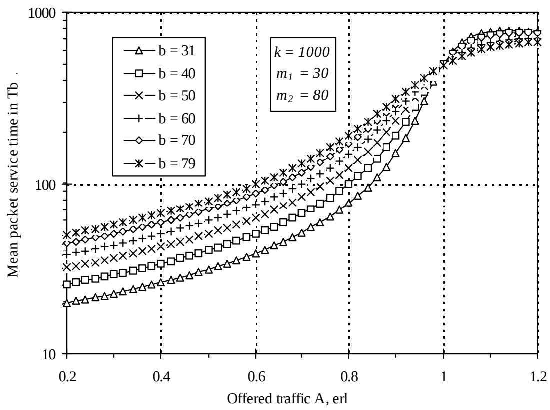

Fig.9. Normalized mean system time of the bytes in discrete time single server queue
with a binomial packet length distribution and different mean packet lengths

Figures 6 and 7 compare the packet congestion probability when the offered traffic is 0.8 erl, the range of packet length is from 30 to 80 bytes, truncated geometric and binomial distribution accordingly and different mean packet size. We can see that the influence of the mean packet length on the packet congestion probability is big.

Figures 8 and 9 present the normalized mean system time of the bytes (W/T_b) as function of the traffic intensity when the queue length is 1000 bytes, the range of packet length is from 30 to 80 bytes, truncated geometric and binomial distribution accordingly and different mean packet size. The influence of the mean packet size on the mean system time is significant when the offered traffic is smaller then 1 erl.

Conclusion

In this paper, different distributions of the packet length: truncated geometric, binomial, discrete uniform and discrete triangular are used and explained. A basic discrete-time single server teletraffic system Geo/Gx/1/k is examined in detail.

The proposed approach provides a unified framework to model discrete-time single server queue. Numerical results and subsequent experience have shown that this approach is accurate and useful in both analyses and simulations of traffic systems.

The importance of a single server queue in a case of a geometric input stream and different distributions of the packet length comes from its ability to describe behaviour that is to be found in more complex real queueing systems. It is the case in a general traffic system, which is an important feature in designing telecommunication networks and systems.

The results presented here add a new aspect to the evaluation of the discrete-time queueing system, and serve as a basis for future research on guaranteeing the quality of service

In conclusion, we believe that the presented formulas will be useful in practice.

Acknowledgements

This paper is sponsored by the National Science Funds of MES - Bulgaria in the framework of project **ВУ-ТН-105/2005** “Multimedia Telecommunications Networks Planning with Quality of Service and Traffic Management”.

Bibliography

[Atencia, 2005] Atencia I. and P. Moreno. A single-server G-queue in discrete-time with geometrical arrival and service process. Perform. Eval. 59: pp. 85-97 (2005)

[Atencia, 2006] Atencia I, P. Bocharov and P. Moreno. A discrete-time Geo/PH/1 queueing system with repeated attempts. Информационные процессы, Том 6, N: 3, стр. 272-280 (2006).

[Cao, 2004] Cao J., W. Cleveland and D. Sun. Bandwidth estimation for best-effort Internet traffic. Statist. Sci., Volume 19, Number 3 (2004), pp. 518-543.

[Dan, 2005] Dan G., V. Fodor, and G. Karlsson, "Packet size distribution: an aside?" in Proc. of QoS-IP'05, pp. 75–87, February 2005.

[Farber, 2002] Farber J., S. Bodamer and J. Charzinski. Measurement and Modelling of Internet Traffic at Access Networks, Proceedings of the EUNICE'98,1998,196-203.

[Janevski, 2003] Janevski T., D. Temkov, A. Tudjarov: Statistical Analysis and Modelling of the Internet Traffic. ICEST Sofia, 2003, pp. 170-173.

[Mirtchev, 2006] Mirtchev S., G. Balabanov and S. Statev, New Teletraffic Models in the IP Networks. National Conference with Foreign Participation, Telecom'2006, Varna, Bulgaria, 2006 (in Bulgarian).

[Pitts, 2000] Pitts J.and J. Schormans. Introduction to IP and ATM Design and Performance - 2nd Ed., John Wiley & Sons, 2000.

[Salvador, 2004] Salvador P., A. Pacheco and R. Valadas. Modelling IP traffic: joint characterization of packet arrivals and packet sizes using BMAPs. Computer Networks, Volume 44, Issue 3 , 2004, pp. 335-352.

[Vicari, 1996] Vicari N. and P. Tran-Gia. A numerical analysis of the Geo/D/N queueing system. Technical Report 04, COST-257, 1996.

[Zang, 2001] Zhang Z. and N.Tian. Discrete Time Geo/G/1 Queue with Multiple Adaptive Vacations, Queueing Systems, Volume 38, Number 4, August 2001, pp. 419-429.

Authors' Information

Seferin Mirtchev – Technical University of Sofia, Kliment Ohridski St., N:8, Bl.1, Sofia-1000, Bulgaria; e-mail: stm@tu-sofia.bg

TOWARDS USEFUL OVERALL NETWORK TELETRAFFIC DEFINITIONS

Stoyan Poryazov

Abstract. A detailed conceptual and a corresponding analytical traffic models of an overall (virtual) circuit switching telecommunication system are used. The models are relatively close to real-life communication systems with homogeneous terminals. In addition to Normalized and Pie-Models Ensue Model and Denial Traffic concept are proposed, as a parts of a technique for presentation and analysis of overall network traffic models functional structure; The ITU-T definitions for: fully routed, successful and effective attempts, and effective traffic are re-formulated. Definitions for fully routed traffic and successful traffic are proposed, because they are absent in the ITU-T recommendations; A definition of demand traffic (absent in ITU-T Recommendations) is proposed. For each definition are appointed: 1) the correspondent part of the conceptual model graphical presentation; 2) analytical equations, valid for mean values, in a stationary state. This allows real network traffic considered to be classified more precisely and shortly. The proposed definitions are applicable for every telecommunication system.

Keywords: Overall Network TrafficTheory, ITU-T Definitions, Virtual Circuits Switching.

ACM Classification Keywords:

1. Introduction

The first what we need for usable Overall Network and Terminal Traffic Theory, is a complete set of clear, precise and useful definitions, particularly for overall network characteristics.

State of the art: Expressions "offered traffic" and "demand traffic" are not found in "ETSI Publications Download Area" [http://pda.etsi.org/pda/queryform.asp] and in [ANSI 2001]. The ITU-T definition of offered traffic is not valid for real telecommunication systems [Poryazov 2005] and that one for demand traffic is simply absent, despite usage in ITU-T Recommendations of expression "demand traffic" three times, and of "traffic demand" – 50 times.

Objective of the research: To trigger discussions towards establishing stable fundaments of Overall Network Traffic Theory.