Features of Processing Various Self-Similar Traffic of Telecommunication Networks

Volodymyr Nakonechnyi [1], Valery Kozlovskyi [2], Andrii Toroshanko [2] and Ivan Shvets [2]

[1] *Taras Shevchenko National University of Kyiv, 60 Volodymyrska str., Kyiv, 01601, Ukraine*
[2] *National Aviation University, 1 Lubomyra Huzara Avenue, Kyiv, 03680, Ukraine*

Abstract

Considered models of mass service systems of self-similar traffic of telecommunication networks. The differences of the main ratios of mass service theory for self-similar traffic in comparison with random processes with a classical Poisson distribution, as well as with distributions with so-called "heavy tails": Pareto, Weibull, log-normal distribution, gamma distribution, beta distribution, are analyzed. Analytical expressions for evaluating the key parameters of mass service systems of self-similar traffic under conditions of stationarity and ergodicity of the request arrival process are presented. Considered models of a single-channel and multi-channel service system with shared and shared buffer memory for the incoming request queue. For self-similar traffic, the analytical dependence of the average queue length on the average network utilization rate is determined. The Hurst parameter was used to estimate the correlation function of self-similar processes. The necessity of managing the packet arrival period and other parameters of the self-similar incoming flow is shown, reducing the risk of overloading individual routes and autonomous network segments. Graphs are shown that illustrate the dependence of the required buffer memory on the utilization ratio, as well as the growth of the queue for deterministic and quasi-deterministic traffic.

Keywords [1]

Telecommunication network, self-similar traffic, mass service theory, multi-channel system, Hurst parameter, network utilization factor

1. Introduction

The processes of the functioning of networks and communication systems can be represented as a set of mass service systems (MSS), for which the characteristics of QoS service quality [1] and other performance indicators are determined. The assessment of traffic service quality indicators requires taking into account many factors in order to build adequate, scientifically based methods of calculation.

For components used to build telecommunication networks (computers, operating systems, network technologies, etc.), analytical models based on mass service theory (MST) provide an acceptable convergence of theory and practice.

The accuracy of simulation results is in all cases limited by the accuracy of the input data. In addition, even in the presence of many assumptions introduced when using MST, the obtained results are close to those that would be obtained with more detailed simulation modelling [2-6]. In addition, analysis based on MST can be performed in a shorter time than simulation.

2. Formulation of the research task

In the mathematical models of MSS, the type of input flow, the scheme of the system and the discipline of service are taken into account [1, 2].

Information Technology and Implementation (IT&I-2023), November 20-21, 2023, Kyiv, Ukraine
EMAIL: vv_k@nau.edu.ua (V. Kozlovskyi); atoroshanko@ukr.net (A. Toroshanko); ivan.shvets@gmail.com (I. Shvets); volodym.nakonechnyi@knu.ua (V. Nakonechnyi)
ORCID: 0000-0002-8301-5501 (V. Kozlovskyi); 0000-0002-0816-657X (A. Toroshanko); 0000-0001-7546-764X (I. Shvets); 0000-0002-0247-5400 (V. Nakonechnyi)

Requests arrive at MSS with some average intensity λ (number of requests per second). At any given time, there will be a certain number of requests (zero or more) in the queue; denote the average number of requests in the queue by w, the average number of requests served – by ρ, and the average waiting time is T_w. This time is averaged over all requests received at the entrance. Average service time of one request T_s −this is the time interval between sending the request to the server and leaving the serviced request from the server. Service intensity μ – this is the number of requests served per unit of time. The total average number of requests in the system is defined as r. Average time of finding a request in the system (waiting in the queue and service) – T_r. If the capacity of the queue is infinite, then requests in the system are never lost; they are only delayed during waiting and service times. Under these circumstances, the average number of sent requests is equal to the average number of incoming requests per unit of time. When the intensity of the arrival of requests at the entrance of the system increases, the time of finding requests in the system also increases, which leads to traffic jams (overload). The queue is getting longer, the waiting time is increasing. When $\rho = 1$, i.e $\lambda = \mu$, the server is saturated, working 100% of the time [1]. Therefore, the theoretical maximum intensity of the incoming flow is related to the average service time T_s as $\lambda_{\max} = {}^{1}/_{T_s}$.

In the first approximation, the length of the request stream is taken as an infinite stream. This means that the average frequency of applications does not change when they are lost. If the length of the stream is limited, then the amount of requests that can be expected at the system entrance is reduced by the number of requests currently in the system; this usually results in a proportional decrease in the average frequency of applications.

If an infinite queue size is assumed, the waiting time can grow to infinity. Under the conditions of a limited queue, some applications in the system may be lost. In practice, of course, any queue is limited. In many cases, this does not lead to a significant difference in the analysis [1, 2].

3. Models of mass service systems

The simplest and most frequently used in MSS are service disciplines FIFO (First came In – First came Out) and LIFO (Last In – First Out) [3, 6, 8]. In computer and telecommunication networks, other service disciplines can also be chosen [9], for example:

− FIRO (First came In − in Random order came Out. Another name − SIRO (Service In Random Order);

– SPT (Shortest requests are Processing First);

– PRS (Priority Requests Service), service according to priority.

In practice, the service discipline is chosen for reasons of acceptable service time. For example, in a node with packet switching, it is possible to provide for the sending of the shortest packets first or, conversely, the longest packets. This choice is determined by the nature of traffic and quality of service requirements.

3.1. Single-channel MSS model

This is the simplest model of SMO [10, 11] (Figure 1). The central element of the system is a server that serves the incoming flow of applications. These applications enter the service system. If the server is free, the request is served immediately. Otherwise, the application becomes a queue for service.

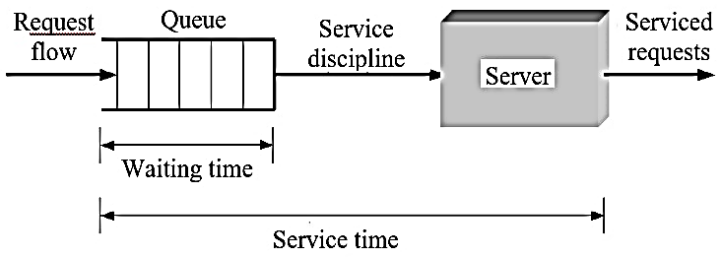


Figure 1: Single-channel MSS model

When the server has finished servicing the current request, it is removed from the queue. If there are requests in the queue, one of them immediately enters the server according to the service discipline used. The server in this model can perform some auxiliary services in processing requests. Examples: a processor provides a service to processes; the data transmission line provides the service of transmission of packets or frames; an I/O device provides read or write requests. When the system is saturated, when $\rho \to 1$, the queue grows to infinity. In practice, in a single-channel system, the intensity of the input flow is limited to 70% to 90% relative to the theoretical maximum.

3.2 A model of a multi-channel system with a common queue

In figure 2 shows a model of a multi-channel service system with a shared buffer memory of the input queue of requests. A common queue with a given service discipline is used for all requests.

If a request arrives at a time when at least one server is free, it is immediately sent to that server. All servers are assumed to be identical; therefore, if more than one server is available, it does not matter which server is selected for service. If all servers are busy, a queue begins to form. As soon as one server becomes free, the request is dequeued according to the current service discipline.

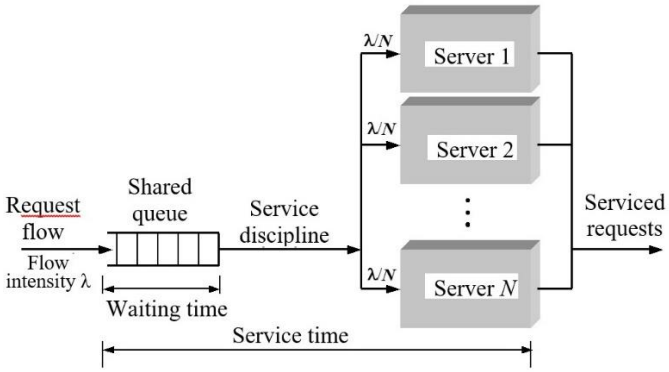


Figure 2: Multi-channel service system with a common queue

Except for service intensity ρ, all parameters used in the analysis of a single-channel system have the same meaning. If used N identical servers, then the average intensity of maintenance of the system as a whole is equal to $N\rho$. This term is often associated with traffic intensity u, which is numerically equal to the intensity of the incoming flow of requests λ. The theoretical maximum of the relative service intensity is equal to $N \times 100\%$, and the theoretical maximum intensity of the incoming flow is $\lambda_{\max} = {}^{N}/_{T_s}$.

3.3 A model of a multichannel system with a split queue

In figure 3 shows a multi-channel system with a separate buffer memory. Such a system can be interpreted as a parallel structure of single-channel service systems. Although the structural changes are not fundamental, the operating characteristics of the depicted system may differ from those previously discussed. The key characteristics of a queue with multiple serving devices are similar to those of a single-channel system. An infinite volume of buffer memory and an infinite size of the queue are assumed, with the distribution of the queue among all serving devices (servers). It is commonly believed that the discipline of service in the order of arrival (FIFO) is implemented. In the case of a multi-channel service system, if all servers are assumed to be identical, the choice of a specific server for the next request does not affect the service time.

4. Determination of key parameters of self-similar traffic

To estimate the average queue size r under conditions of stationarity and ergodicity of the

application arrival process, Little's formula is used [1, 9, 11]:

- for a single-channel service system:

$$r = \lambda T,\ r = w + \rho; \tag{1}$$

- for N -channel service system:

$$\rho = {\lambda T_r}/{N},\quad u = \lambda T_s = \rho N,\ r = w + N\rho,\ \text{де}\quad T_r = T_w + T_s\,. \tag{2}$$

Accordingly, Little's formulas can be used to connect the number ρ with the intensity of incoming requests λ and the time the request was in the system T_s: $\rho = \lambda T_s$.

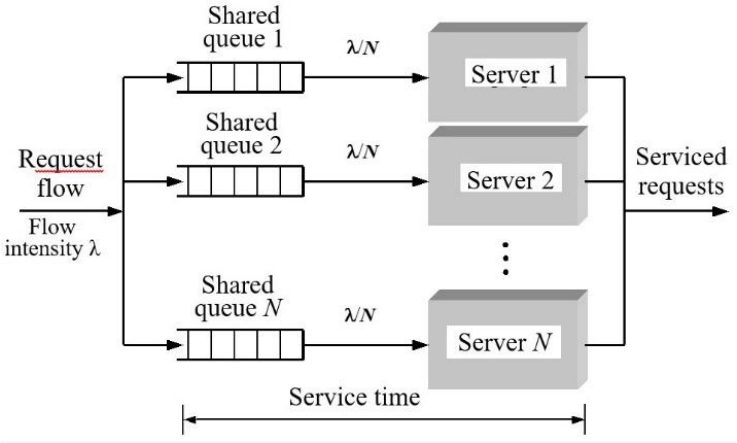


Figure 3: Multi-channel service system with split buffer queue

Thus, the following a priori information is necessary for the analysis of MSS: the intensity of the incoming flow of requests, the average service time, and the number of service channels. Based on this information, you can get asymptotic estimates of the average number of requests in the queue, the average waiting time, and the total time the requests are in the system.

It should be taken into account that request flows may not be distributed according to Poisson's law, but according to other probabilistic laws with so-called "heavy tails" [11]. These are the Pareto, Weibull, log-normal distribution, gamma distribution, beta distribution, and some other less popular ones.

For example, for the Pareto distribution, the main relations have the following form:

- probability density

$$f(x) = {\alpha}/{k}\left({k}/{x}\right)^{\alpha+1},\ (k \text{ i } \alpha < 0) - \text{distribution parameters;} \tag{3}$$

- probability function:

$$F(x) = 1 - {k}/{x}^{\alpha},(x > k,\ \alpha > 0); \tag{4}$$

- average value

$$E[X] = {\alpha}/{\alpha - 1}\,k,\ (\alpha > 1). \tag{5}$$

Real random processes, of course, preserve the property of self-similarity only up to a certain limit. This measure of the statistical stability of the process under multiple scaling is defined by the so-called Hurst parameter or related self-similarity parameter. A random process *x(t)* is statistically self-similar with the Hurst parameter *H* (0,5 ≤ *H* ≤ 1), if for any *a* > 0 process ${x(at)}/{a^H}$ has the same statistical characteristics as the process $x(t)$ itself:

- mathematical expectation

$$M[x(t)] = {M[x(at)]}/{a^H} \tag{6}$$

- dispersion

$$D[x(t)] = {D[x(at)]}/{a^{2H}} \tag{7}$$

- correlation function

$$R(t,a) = {R(at,a\tau)}/{a^{2H}} \tag{8}$$

The more H, the longer the property of self-similarity is preserved under multiple scaling. At $H = 0{,}5$ this property is practically absent.

Correlation functions of self-similar processes with a large Hurst parameter decay more slowly than those of ordinary random processes, and the decay has, as a rule, an oscillatory character. It was established that the constant component of the correlation function decreases according to the law $c_1 t^{-c_2 a}$, where c_1, c_2 – constants, a – scale parameter.

Accordingly, the spectral density of the process theoretically tends to infinity at a frequency approaching zero. Ratios (1-7) can be useful as asymptotic approximations of real processes.

Such specific characteristics are inherent not only to data traffic (TCP, FTP protocols), but also to signal traffic (SS7 protocol), VBR-video, Ethernet/ISDN and some others. Physically, they are caused by a high degree of grouping of packets at client sites, in routers and switching nodes of information communication networks. Even if the source generates a regular stream of packets, the data is delivered to the consumer in bursts interspersed with idle intervals. The reasons for this are the limited speed of network devices, insufficient volume of buffers, etc. In addition, self-similar traffic has a special structure that is preserved during multiple scaling. In real processes, there is some outliers with a relatively small average traffic level. Due to such bursts of load, network characteristics also deteriorate: losses, delays, jitter of packets when passing through network nodes increase [12].

Methods for calculating the requirements for networks of new generations (channel bandwidth, buffer capacity, etc.) based on Markov models and Erlang or Little formulas, which were successfully used in the design of telephone networks, can give unreasonably optimistic solutions and lead [7, 13].

With the self-similar nature of the traffic, the dependence of the average duration of the queue (respectively, the required size of the buffer) q from the average utilization ratio has the following form:

$$q = \rho^{1/2(1-H)} \Big/ (1-\rho)^{H/(1-H)}$$

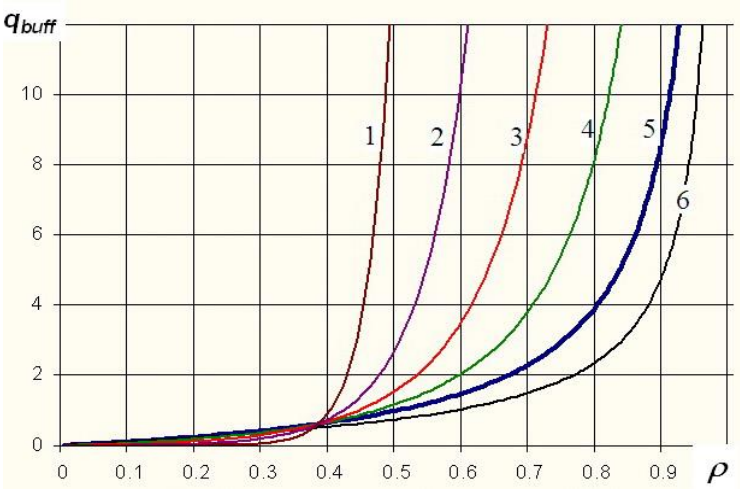
At H=0,5 this formula is simplified:

$$q = \rho / (1-\rho)$$

which is a classical result of MSS with the simplest input flow and exponentially distributed service time (*M*/*M*/1). For a system with deterministic maintenance time (*M*/*D*/1) a classic result looks like this:

$$q = \rho / 1-\rho - \rho^2 / 2(1-\rho).$$

In figure 4 shows the results of calculations of the dependence of the required buffer memory q_{buff} from the utilization factor $\rho = \lambda/\mu$ for different inbound traffic patterns. Calculations are made as for Poisson flows of requests *M*/*M*/1 і *M*/*D*/1, as well as for self-similar flows.



(1 – *H*=0,6; 2 – *H*=0,8; 3 – *H*=0,7; 4 – *H*=0,4; 5 – *M*/*M*/1; 6 – *M*/*D*/1)

Figure 4: Dependencies of the required buffer memory on the utilization ratio $\boldsymbol{\rho}$.

The graphs clearly show that for self-similar traffic already at $\rho \approx 0{,}4$ a larger memory resource of buffer devices is required than for the classic model *M*/*M*/1, which is considered the least favorable compared to others (for example, with a constant or Gaussian service time distribution). The rate of growth of the required amount of memory increases with the increase of the Hurst parameter, which is mainly due to the degree of grouping of homogeneous packets and bursts of network load.

It can also be concluded that simply increasing the buffer memory (hardware or software) is ineffective. With the expected increase in the share of data traffic in the total volume, the degree of self-similarity will increase, and the dependence $\rho(q_{buff})$ will grow more and more sharply. To eliminate or at least reduce the harmful effect of traffic similarity, methods of regulation or shaping of the incoming flow (policing - shaping) are usually used. Ideally, this results in a deterministic or close to deterministic application order. With deterministic traffic (deterministic order of incoming applications and deterministic processing time), the queue growth graph is a linear-broken line (Figure 5).

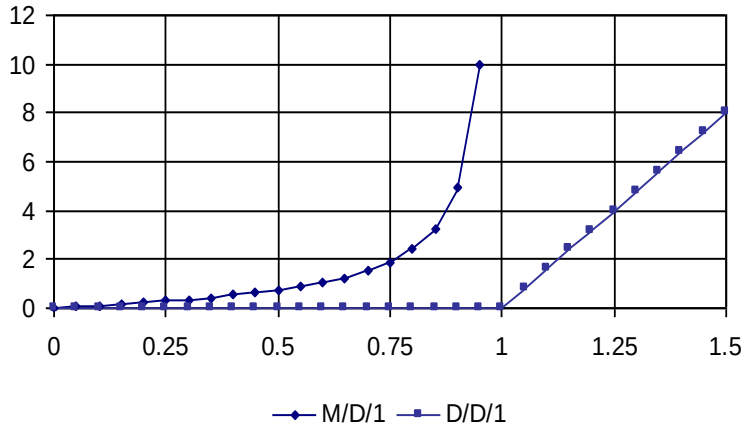


Figure 5: Graph of queue growth with deterministic traffic

In practice, both the traffic at the output of the shaper and the packet processing time are quasi-deterministic (we denote them by QD). In figure. 6 shows graphs for the relevant cases.

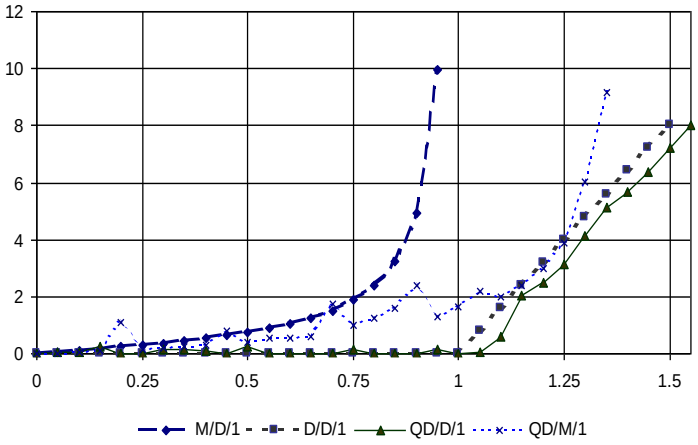


Figure 6: Graph of queue growth with quasi-deterministic traffic

5. Conclusions

Models of heterogeneous computer network traffic, which has self-similar properties, are analyzed in the work. The considered models of single-channel and multi-channel service system with shared and separate buffer memory for the input queue of requests. Their advantages and disadvantages are shown. Analytical expressions for evaluating the key parameters of mass service systems of self-similar traffic under conditions of stationarity and ergodicity of the application arrival process are presented.

For self-similar traffic, the analytical dependence of the average queue duration on the average network utilization rate is determined.

To eliminate traffic bursting caused by the similarity of the incoming stream, it is necessary to control its parameters, first of all, the period of arrival of packets. Thanks to this, the rate of growth of queues in the buffer memory of switching nodes slows down. As a result, the risk of overloading individual routes and autonomous network segments is reduced..

6. References

[1] Giambene G. Queuing Theory and Telecommunications: Networks and Applications; 2nd edition. – Springer NY, 2014.

[2] Floudas C.A., Pardalos P.M. Encyclopedia of optimization: 2-d ed. – Springer science+buisiness media, LLC, 2009. – 4646 p.

[3] Bonaventure O. Computer Networking: Principles, Protocols and Practices. Release. – cnp3book, 2018.

[4] Stallings W. Foundations of Modern Networking: SDN, NFV, QoE, IoT, and Cloud. New Jersey; Pearson Education, Inc., Old Tappan, 2016.

[5] Popovs′kyy V.V. Telekomunikatsiyni systemy ta merezhi. Struktura ta osnovni funktsiyi. Kharkiv.: SMIT, 2018. http://www.znanius.com/3534.html

[6] Speidel J. Introduction to Digital Communications / Springer Nature Switzerland AG, 2019.

[7] Kurose J.F., Ross K.W. Computer Networking: A Top-Down Approach. – 7th ed. – Pearson Education, Inc., 2017.

[8] Frenzel L.E. Principles of electronic communication systems, 4th Ed. McGraw Hill Education, 2016.

[9] Lesnaya N.N. Development of control algorithm intelligent multiservice networks. Problems efficiency infrastructure. Collected Works, issue 11. Kyiv, 2015. 150-155.

[10] Tanenbaum Andrew S., Maarten Van Steen. Distributed systems: principles and paradigms. Pearson Education. Inc. Pearson Prentice Hall, Upper Saddle River, NJ 07458, 2007.

[11] Vinogradov N.A., Savchenko A.S. Comparative analysis of the functionals of optimal control corporate computer network. Journal of Qafqaz University (Mathematics and Computer Science). Vol. 1, Nr. 2., 2013. 156-167.

[12] Di Giorgio A., Pietrabissa A., Priscoli F.D., Isidori A. Robust Output Regulation for a Class of Linear Differential-Algebraic Systems. – IEEE Control Systems Letters. – 2018. – Vol. 2, No. 3. – P. 477-482.

[13] Forbs C., Evans M., Hastings N., Peacock B. Statistical Distributions: 4-th Edition. – John Wiley & Sons, Inc., Hoboken, New Jersey, 2011. – 212 pp.