Calculation of Data Flow in Local Computer Networks. Combined (Experimental + Computer Simulation) Approach

N. A. FILIMONOVA
SysAn
A. Nevskogo, 12a, 34, 630075 Novosibirsk
RUSSIA

S.I. RAKIN
Siberian Transport University
D. Koval'chuk st., 191, 630049 Novosibirsk
RUSSIA

Abstract: – In the present paper, we consider a model that can be described as a "users – network applications" in local networks. The characteristic scales of the model are: the number of users up to 1024, time of events from several seconds to hours, the transmitted data volume –10-10000Kb. This scale rise to a new model that appears between the packet-level and the global Internet level models. We introduce the notion of the elementary data flow from an Internet service and from a peer. By using these notions, we develop the "metrology" approach to the modeling of data flow in local networks. Examples are presented.

Key-Words: - Networks, Computer Networks, Data Flow, Simulation

Received: May 19, 2022. Revised: January 12, 2023. Accepted: February 17, 2023. Published: March 17, 2023.

1 Introduction

A usual presumption is that a mathematical model of any sort depends on the scale of the modeled process or system. Most of the contemporary models for data networks use as basic variables the data volume and transmission time, which usually belongs to the following scales [1-3]:

- Packet scale: here, the usual data volume is in the range 46 – 1500 bytes; the usual time unit is millisecond. This type of scale gives rise to queuing theory models and is used at the hardware level.
- Global network scale: here the usual data volume is vast. It can be estimated from the average network speed of 10 – 20 MBt/sec. The usual timescale varies from hours to months. This type of scale gives rise to stochastic process models and is used at the level of large networks.

In the present paper, we consider a model that can be described as a "users - network applications", for which the basic variables are:

- the number of users 1 – 1024 for local networks,
- time on the scale of seconds (a minimal time to start and use an application) to hours (up to 4 hours as half a business day),
- the transmitted data volume per internet session 10-10000Kb.

This naturally arising scale gives rise to a new model that appears between the packet-level models and the global Internet level models described above. This model is visibly different from either packet-level or the global Internet level models.

Our model is based on the fundamental, albeit simple, observation that data in local networks is generated in two distinct stages. First, a network peer (a human or a computer) starts a network application. Then the application itself generates a data flow by its own rule (which may or may not be affected by peer's actions).

This observation allows us to conclude that the overall process of network data flow generation has to be compounded from peer's own activity and its applications' individual data flows.

1.1 Peer's Own Activity

The total traffic depends on the elementary data stream from every e-mail service as well every peer activity. The study of the peer's activity is the subject of study of physiology, social and similar sciences. As we see, the computation of the traffic is an interdisciplinary problem that should be based on the methods of both technical and socio-economic sciences. Keeping in mind the methodological nature of this paper, we collected data on the peer's activity in student groups.

Network peer's activity is largely described by its start and end times of peer's network applications, and is largely specific to peer's business and daily routine. We have timed some standard working schedules for a network peer (a human user) working with the above mentioned applications. E.g., while using a search engine a user opens a new web-page each 30 seconds after the closing of the previous one, and then spends 6o-180 seconds browsing it. While using e-mail in intensive exchange mode, time interval between typing each e-mail is 10 min and Time of typing an e-mail is 6-7 min.

1.2 Individual Data Flows from Network Applications

The principal question here is whether a network application generates a data flow that is random and unpredictable enough, or data flow is specific for a specific application Our experiments corroborate the latter case. This allows us to introduce the notion of an elementary data flow, as a data flow that is specific to a given network application. After defining such a notion, one may describe elementary data flows generated by network applications. This may be done by experimentally collecting sufficient amounts of raw data and its subsequent statistical analysis. In this note, we present results employing the output data flow only. All measurements are collected by using TMeter software [4].

1.3 Data Flow Superposition

The problem of data flows superposition necessarily arises when the simultaneous activity of several peers is considered. The main question is: are data streams from different Internet services/peers additive? The existence of data compression implies that, generally, data has a variable volume (i.e. data is like a compressible gas, not an incompressible liquid). We will discuss this issue in detail below.

2 Elementary Data Flows Generated by Popular Network Applications

Below we present the above mentioned models stemming from our experimental data and statistical analysis.

2.1 E-mail Client via Remote Server

The corresponding elementary data flow $Z_1(t_1, t_2)$ is depicted in Fig. 1 which shows a very characteristic picture of a series of isolated impulses. Such a series is always finalized by a larger impulse I associated with the transmission of the message. Here, t_1 and t_2 are, respectively, the start and end time of peer's activity. The parameters v and l are the data transmission speed and the time interval between impulses (the values of v, l, and I are solely determined by the e-mail client).

2.2 Web-surfing

A sequence of three elementary data flows corresponding to three counts of consecutive web-page access is depicted in Fig.2. Each flow $Z_2(t_1)$ is characterized by the access time t_1 (which is determined by peer's activity). A usual flow's shape is that of a step-function, in which each step has a random value V.

2.3 Skype

An elementary data flow from a Skype session is depicted in Fig. 3. This one is a continuous function $Z_3(t_1, t_2)$, where t_1 and t_2 are the start and end times of the session (peer's own activity). The data transmission speed W is a random process.

Figs 1-3 lead to the hypothesis that each network application has its own form of data flow. Our statistical analysis of the obtained experimental measurements confirms this hypothesis and gives us empirical distribution densities for random variables v, l, V, W. Thus, we obtain models for elementary data flows of the entire above mentioned network applications.

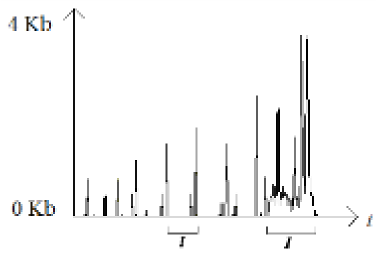


Fig. 1: Output data flow of an e-mail client

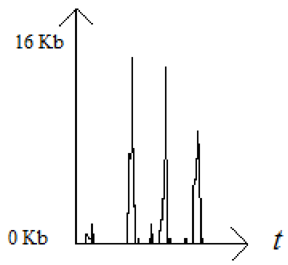


Fig. 2: Output data flow from web-surfing (accessing 3 web-pages)

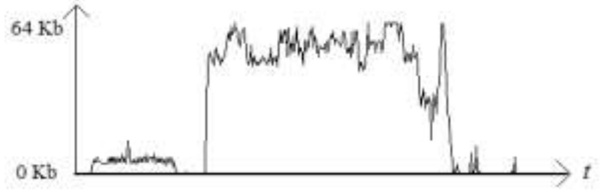


Fig. 3: Output data flow of a Skype session: audio (left) and video (right) modes

2.4 Model

Our model of network data flow generation by a single peer is represented by the following process that is decomposed into distinct stages: first, a network peer starts an application (or several applications) from a given set of software at random time t_1, and then the corresponding application (or applications) generates their elementary data flows. An essentially non-homogeneous data flow is generated as the result, which we call **a single peer's elementary data flow.**

More formally, the model $Y_i(t)$ of a single peer's data flow for the i-th peer (which is a discrete time model with time-step δ) is the following:

- first, some random values $t_1, t_2, ...$ (for Z_1 and Z_3) are generated to be used as start and end times for the aforementioned network applications, as well as t (for Z_2) to be used as access times;
- then we make time steps $t = 0$, $t \to t + \delta$;
- if $t = t_i$ then we start or stop generating the elementary data flows Z_i ($Z_1(t, t_2)$, $Z_2(t)$ or $Z_3(t, t_2)$) described above.

Such kind model/ based on the experimentally measured elementary data flows was referred in [5]as based at the "metrology" approach to the data streams. It seems, the "metrology" approach is more closed to the classical traffic [6-11] and statistical [12-15] approaches in teletraffic theory rather that the packet simulation models [16-20].

3 Flow Superposition. Conservation of Total Data Volume

The topic considered in this section can be briefly described as a question about a "conservation law" for the total data volume in a network. The existence of data compression [20] implies that, in general, data has variable volume (i.e. data is akin to compressible gas rather than incompressible liquid). However, a total volume conservation phenomenon may be observed in some networks. Such a property is equivalent to the statement that the data transmission speed is additive under data flow superposition (i.e. given two data flows Z_1 and Z_2 generated simultaneously the total data flow equals $Z_1 + Z_2$.

3.1 Data Flow Superposition for a Single Peer

Flow superposition already takes place for elementary data flows generated by a single network peer. To this end, Fig. 4 illustrates an example of a peer browsing web-pages while having a Skype session, as shown by experimental measurements.

In the picture, segment 1 shows the data flow from a Skype session alone, while segment 2 shows the data flow from browsing (3 web-pages were browsed without having Skype active). Then, segment 3 shows the data flow generated by having browsed the same 3 web pages during an active Skype session). Here we observe that the data flows add up (with a small margin of measurement error).

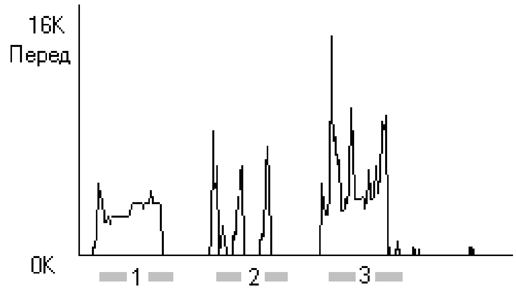


Fig. 4: Additivity of elementary data flows under superposition: 1 – Skype session, 2 – web-browsing, 3 – Skype session and web-browsing (experimental measurement with TMeter).

3.2 Data Flow Superposition for Multiple Peers

The total data flow from multiple peers in a local network, generating each a flow $Y_i(t)$ can be very closely approximated by the sum $Y_1(t) + Y_2(t) + ... + Y_n(t)$. This fact has been verified by experimental measurements for superposition of a large number (more than 100) of various data flows. Such a conclusion is legit if the total flow does not exceed the overall network capacity.

4 An Example. E-mail Clients

Now, having created the necessary models for elementary data flows generated by network applications and having established data flow additivity, we can model the total data flow from an arbitrary number of peers by using a computer program. As an example, we present our numeric results for the peers using their e-mail clients. The calculation was carried out for the parameters described below.

4.1 Peer's Network Activity

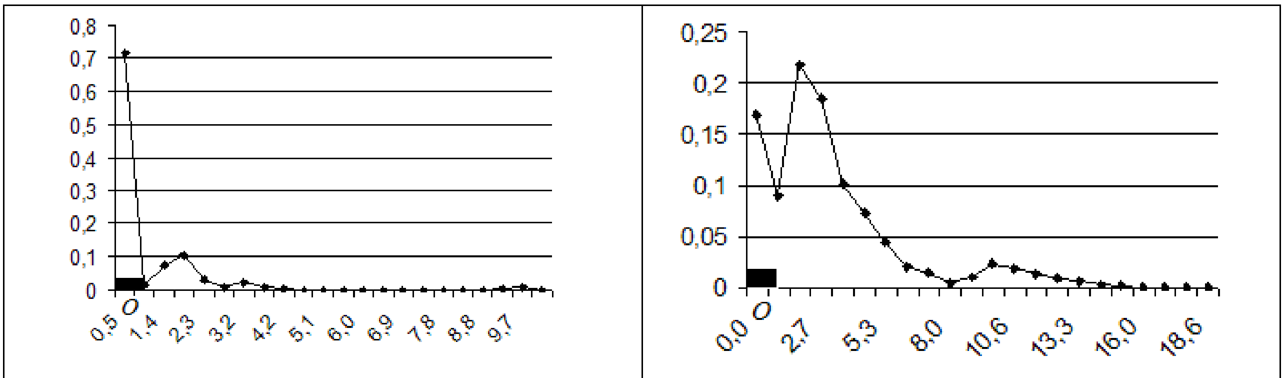
While using a search engine, a user opens a new web-page each 30 seconds after the closing of the previous one, and then spends 6o – 180 seconds browsing it. The total time of uninterrupted network activity is taken to be 4 hours (a standard half business day). The total number of peers varied from 2 to 1000.

4.2 Elementary Data Flow

An elementary flow generated by an e-mail client is depicted in Fig. 1. The statistical model of the elementary stream generated by the mail client, as well as the parameters of the model, were determined from experimental measurements. Detailed information about the elementary stream model generated by the mail client can be found in [5, 21].

4.3 Results of Computer Simulation

The total output data flows are depicted in Fig.5 and Fig.6. The abscissa shows the data transmission speed (Kb/sec), and the ordinate shows the corresponding probability. The intervals labeled in Fig.5 *O* correspond to zero data traffic.



Number of peers: 2	*Number of peers: 10*

Fig. 5: Empirical distribution density for the total data flow

When the number of peers exceeds 20, the resulting empirical distributions have a great proximity to the Gamma-distribution, as shown in Fig.6. This observation has been confirmed by using Kolmogorov-Smirnov test [22] (details may be found in [5, 21]).

Number of peers: 20	*Number of peers: 100*	*Number of peers: 1000*

Fig. 6: Empirical distribution density for the total data flow MO = average value, *L* is left "zero margin", *R* is right "zero margin"

When the number of peers exceeds 2000, the empirical distribution function rather approaches the Gaussian distribution (which is the usual law of large numbers). However, the total number of peers in a local network is limited to 1024.

This is why it is absolutely necessary to use experimental measurements as an integral part of our model, which we call **a combined model.** This means that in our model, the initial experimental measurements for elementary data flows in a concrete instance of a local network are used for further computer simulation.

Table 1. Parameters of Gamma-distribution α and β, and average transmission speed (Kb/sec) in a local network as function of the number of peers n

n	20	100	1000
α	3	9.9	71
β	1.8	3	4.1
MO	5.4	29	290
$R-MO$	13.6	34	108

4.4 Application: Estimating Required Network Capacity

Table 1 features the values of parameters α, β for the Gamma-distributions with the empirical density shown in Fig.6. The average transmission speed MO equals $\alpha\beta$. Deviation to the right margin (see Fig.6) is $R-MO$. The theoretical density of Gamma-distributions is

$$\frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}.$$

Since the aforementioned data flows are described by the Gamma-distribution with parameters α, β, the required network capacity $R = R(\alpha, \beta)$ has to exceed the right zero-margin. This condition can be written as

$$\frac{1}{\beta^{\alpha}\Gamma(\alpha)} R^{\alpha-1} e^{-\frac{R}{\beta}} = \delta,$$

where δ is a small number (in our computations, $\delta = 0.002$).

By solving the above equation with respect to R, we find the required minimal network capacity as a function of the number of peers. Since the values of α, β in the above equation are given in Table 1, solving it is a routine computation.

This solution is legit for the case of using solely the e-mail client. If there are several applications employed by peers, we have to take into account how many peers are using each. This can be done with an analogue to the total probability formula. It is required, of course, to develop models for all elementary data flows from all network applications used.

5 An Example. Web-surfing

Now, we present our numeric results for the case of multiple peers using web browser for the Web-surfing. The calculation was carried out for the parameters described below.

5.1 Peer's Network Activity

While using a search engine, a user opens a new web-page each 30 seconds after the closing of the previous one, and then spends 6o – 180 seconds browsing it. The total time of uninterrupted network activity is taken to be 4 hours (a standard half business day). The total number of peers varied from 2 to 1000.

5.2 Elementary Data Flow

An elementary flow generated by a web-page visit is depicted in Fig. 1. The statistical models of the elementary flow, as well as the parameters of the model, were determined from experimental measurements. Detailed information about the elementary stream model generated by the mail client can be found in [21].

5.3 Results of Computer Simulation

Figs 7-17 show the data transfer rates (experimental) in Kb when clicking on hyperlinks (web pages), relative empirical frequencies and their distribution over intervals.

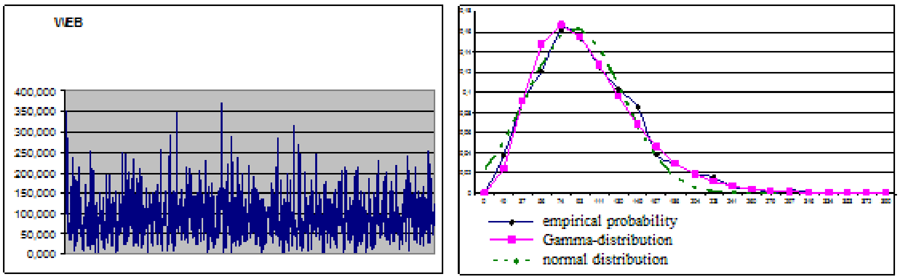


Fig. 7: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 50.

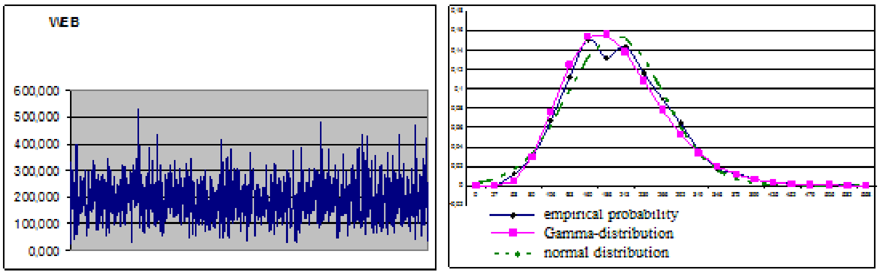


Fig. 8: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 100.

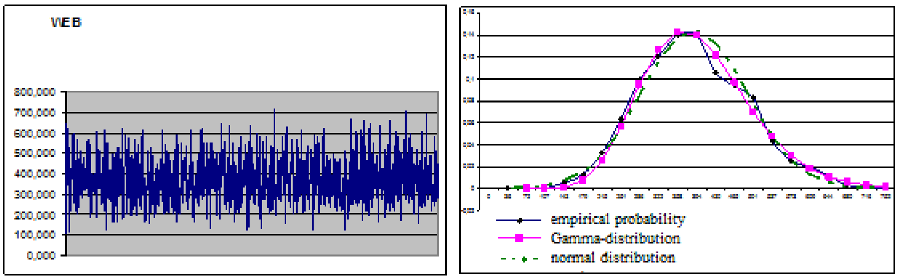


Fig. 9: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 200.

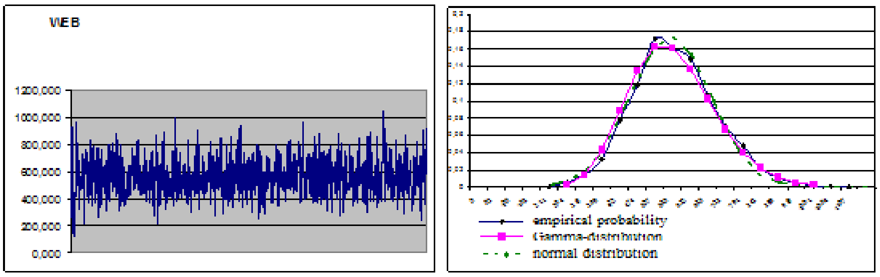


Fig. 10: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 300.

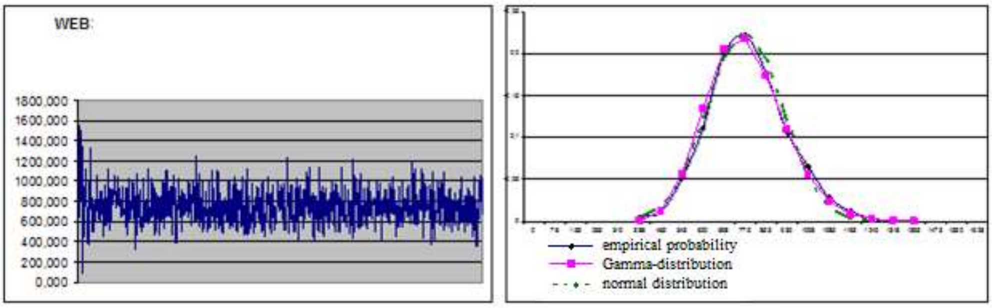


Fig. 11: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 400.

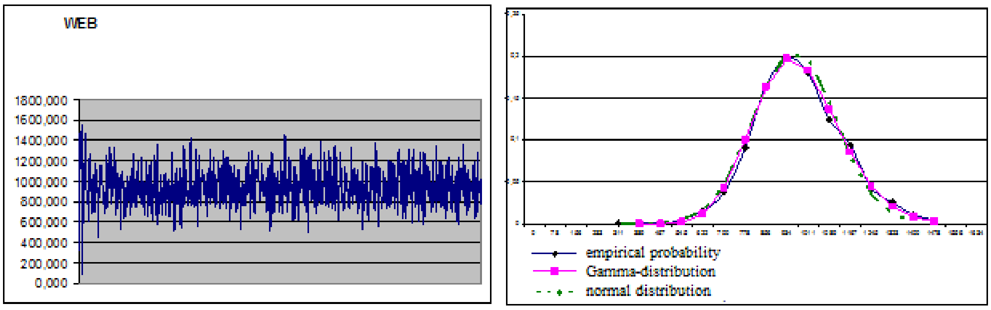


Fig. 12: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 500.

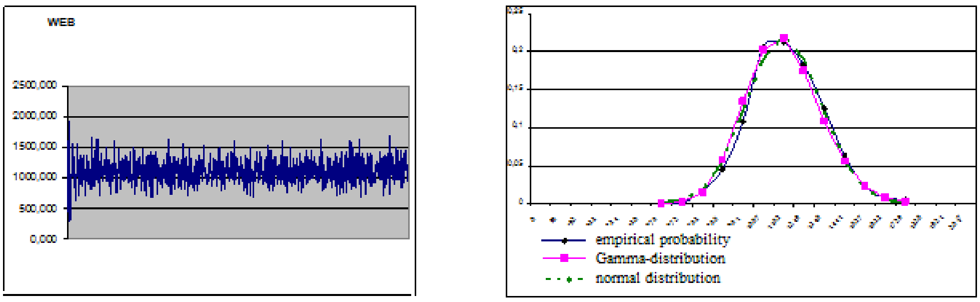


Fig. 13: The gamma function, normal density function, and the density, determined by using computer simulation. The number of users is 600.

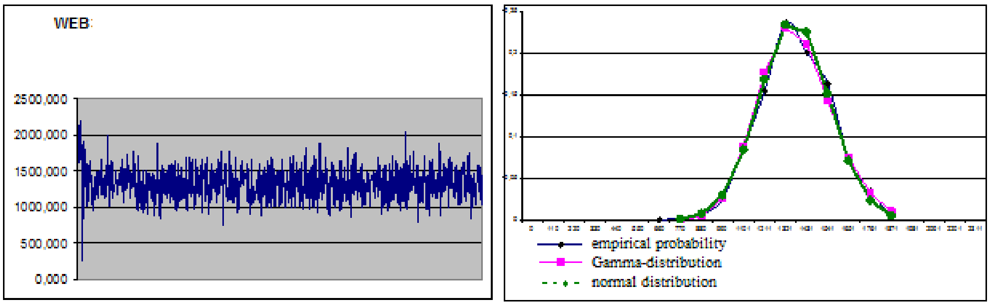


Fig. 14: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 700.

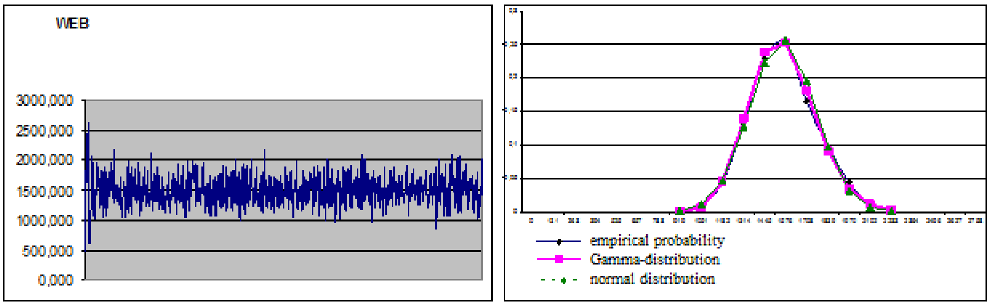


Fig. 15: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 800.

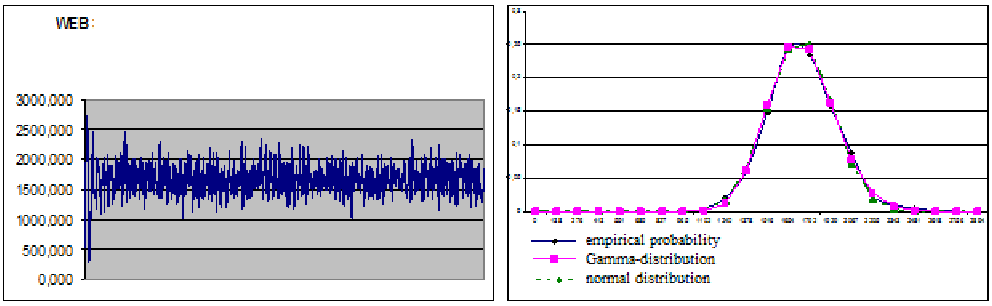


Fig. 16: Computer simulated traffic (left) and gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 900.

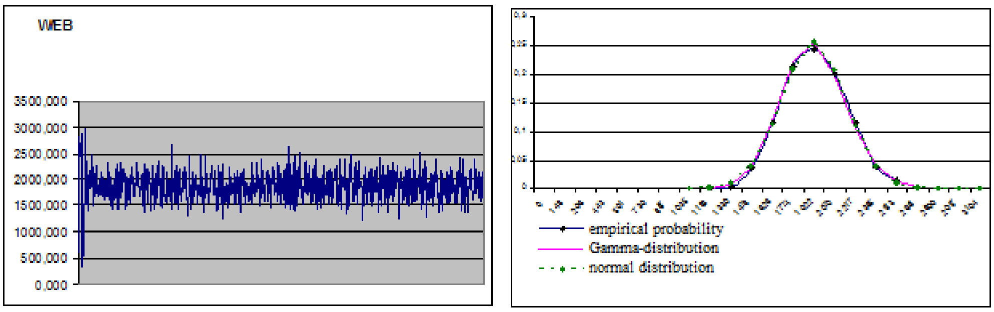


Fig. 17: Computer simulated traffic (left) and Gamma function, normal function, and the density, corresponding to the computer simulated traffic. The number of peers 1000.

5.4. Justification of the Constructed Density of the Data Flow Rate

Visually, there is a good match between the plots of the distribution functions determined from the computer simulations and the plots of the distribution density functions of the gamma distribution in Figs 7-17. We present the statistical justification for this conclusion. We propose the following hypothesis: the data rate determined from our computer simulation has a gamma distribution with the parameters indicated in Table 2.

Table 2. Peer number n and parameters of Gamma-distribution α and β

n	50	100	200	300	400	500	600	700	800	900	1000
α	4	8	15	20	30	37	43	52	59	67	70,1
β	25	25	26,4	29	25,9	26	26,8	26,4	26,4	26	27,4

To verify/reject this hypothesis, we use the Kolmogorov-Smirnov test [22], which consists of the following: as a measure of the discrepancy between the theoretical and statistical distributions, the maximum value of the modulus of the difference between the statistical distribution function $F(x)$ and the corresponding theoretical distribution function $F^*(x)$ is considered:

$$D = \max | F(x) - F^*(x) | .$$

The critical value of the Kolmogorov–Smirnov test is calculated by the formula $\lambda = D\sqrt{n}$, where n is the number of relative empirical frequencies. The probability $P(\lambda)$ is determined from the table from [22] ($P(\lambda)$ is the probability that, due to purely random reasons, the maximum discrepancy between $F(x)$ and $F^*(x)$ will be no less than the one that is actually observed [22]). If $P(\lambda)$ is close to 1, then the hypothesis of the gamma distribution of the computer-simulated data transfer rate is accepted. At a value close to zero, this hypothesis is rejected, see [22] for details.

The calculated values λ and $P(\lambda)$ are presented in Table 3. Since the probabilities $P(\lambda)$ from Table 3 are close to 1, then we accept the hypothesis of the gamma distribution of the simulated frequencies.

Table 3. Critical value of the Kolmogorov-Smirnov test λ and $P(\lambda)$.

n	50	100	200	300	400	500	600	700	800	900	1000
λ	0,305	0,29	0,42	0,23	0,324	0,35	0,26	0,35	0,37	0,44	0,33
$P(\lambda)$	1	1	0,997	1	1	1	1	1	1	0,997	1

Note that the hypothesis: the data rate determined from our computer simulation has a normal distribution with the density

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-a}{\sigma}\right)^2}$$

with the parameters indicated in Table 4 is also valid. The calculated values λ and $P(\lambda)$ for this hypothesis are presented in Table 5. One can use Gamma-distribution or normal distribution, as it is more convenient.

Table 4. Peer number n and parameters of normal distribution

n	50	100	200	300	400	500	600	700	800	900	1000
a	88	199	389	574	774	960	1152	1365	1496	1730	1920
σ	45	68	99	121	135	150	173	183	220	210	230

Table 5 Critical value of the Kolmogorov–Smirnov test λ and $P(\lambda)$.

n	50	100	200	300	400	500	600	700	800	900	1000
λ	0,2	0,37	0,44	0,29	0,38	0,39	0,32	0,37	0,08	0,42	0,34
$P(\lambda)$	1	1	0,997	1	1	1	1	1	1	0,997	1

Data from Table 2 or Table 4 may be used for estimating required network capacity, see Section 4.4 for details.

6 Conclusion and Prospective

The “metrology” approach presented in this paper is a new method most suitable for the modeling of data flow in local networks. It directly accounts for specific peers’ activity and specific characteristics of Internet services in use. This approach rise to a new model that appears between the packet-level and the global Internet level models. This approach is evidently integrated with the global Internet model.

Progress in the area under discussion implies continued collection of information about the data streams generated by Internet services, and progress in the summation of functions distributed according to the law of the Gamma distribution.

References:

[1] Liu, D.; Xu, X.; Liu, M.; Liu, Y. Dynamic traffic classification algorithm and simulation of energy Internet of things based on machine learning. *Neural Comput. Appl*., 33, **2021,** 3967–3976.

[2] WłodarskI, P. Impact of Long-Range Dependent Traffic in IoT Local Wireless Networks on Backhaul Link Performance. In*: International Conference on Computational Science—ICCS 2020*; Springer: Cham, Switzerland; Berlin/Heidelberg, Germany, 2020 pp. 423–435.

[3] Data Volume of Global Consumer Internet Traffic from 2017 to 2022, by Subsegment. Available online: https://www.statista.com/statistics/267194/forecast-of-internet-traffic-by-subsegment/ (accessed on 1 June 2022).

[4] TMeter Freeware Edition. *Available online: http://tmeter.ru/* (accessed on 1 June 2022).

[5] Filimomova N.A, Kolpakov A.G, Rakin S.I. “Metrology” Approach to Data Streams Initiated by Internet Services in the Local Networks. *Computers* **2022**, 11(9), 138.

[6] Erlang, A.K. Probability and telephone calls. *Nyt Tidsskr. Mat. B* 20, **1909**, pp33–39.

[7] Erlang, A.K. Losning af nogle problemer fra sandsynlighedsregningen af betydning for de automatiske telefoncentraler. *Elektroteknikeren* 13, **1917**, pp 5–13.

[8] Pollaczek, F. Über eine Aufgabe der Wahrscheinlichkeitstheorie. *Math. Z.* 32, **1930**, 64–100.

[9] Kendall, D.G. Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *Ann. Math. Stat*. **1953**, 24, 338–354.

[10] Roberts, J.; Mecca, U.; Virago, J. (Eds.) *Broadband Network Teletraffic: Performance Evaluation and Design of Broadband Multiservice Networks*, Final Report of Action COST 242; Springer: Berlin/Heidelberg, Germany, 1996.

[11] Akimaru, H.; Kawashima, K. *Teletraffic: Theory and Applications (Telecommunication Networks and Computer Systems)*, 2nd ed.; Springer: New York, NY, USA, 2012.

[12] Mekid, S. *Metrology and Instrumentation: Practical Applications for Engineering and Manufacturing*; Wiley-ASME Press: Hoboken, NJ, 2022.

[13] Crowder, S.; Delker, C.; Forrest, E.; Martin, N. *Introduction to Statistics in Metrology*; Springer: Cham, 2020.

[14] Loh, T.H. (Ed.) *Metrology for 5G and Emerging Wireless Technologies (Telecommunications)*; The Institution of Engineering and Technology: London, 2022.

[15] Kurose, J.F.; Ross, K.W. *Computer Networking: A Top-Down Approach*, 6th ed.; Pearson: Boston, MA, 2012.

[16] Hui, J.N. *Switching and Traffic Theory for Integrated Broadband Networks; Springer*: New York, 2012.

[17] Schmid, S.; Wattenhofer, R. Dynamic Internet Congestion with Bursts. In: *High Performance Computing—HiPC 2006*; Robert, Y., Parashar, M., Badrinath, R., Prasanna, V.K., Eds.; Springer: Berlin/Heidelberg, Germany, 2006.

[18] Jeruchim, M.C.; Balaban, P.; Shanmugan, K.S. *Simulation of Communication Systems: Methodology and Techniques*; Springer: New York, 2000.

[19] Rorabaugh, C.B. *Simulating Wireless Communication Systems*; Prentice Hall: Hoboken, NJ, 2004.

[20] Dumas, J.G.; Roch, J.L. *Foundations of Coding: Compression, Encryption, Error Correction*; Wiley: New York, 2015

[21] Filimonova, N.A. Model of elementary stream in Internet. In: *Distributed Informational and Computational Resources*, 26–30 November 2012; Siberian Branch of Russian Academy of Sciences: Novosibirsk, Russia, 2012.

[22] Olofsson, P.; Andersson, M. Probability, *Statistics, and Stochastic Processes,* 2nd ed.; Wiley: New York, 2012.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

N.A. Filimonova developed general methodology, simulation, algorithms and experiments. S.I. Rakin was responsible for the general organization, resources, preparation of the manuscript.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself.

There is no external fund

Conflict of Interest

The authors have no conflicts of interest to declare that are relevant to the content of this article.