Statistical Data Handling Program of Wireshark Analyzer and Incoming Traffic Research

Veniamin Tarasov <tarasov-vn@psuti.ru>,
Sergey Malakhov <malakhov-sv@psuti.ru>
Volga Region State University of Telecommunications and Informatics, 77 Moskovskoe sh., Samara, 443090, Russian Federation

Abstract. The identification of the distribution laws of intervals is particularly sophisticated problem, at the same time the traffic as a random process tends to be constantly changing. Therefore it is important to know the numerical characteristics of these intervals or their moments. In this paper we propose to use the Wireshark analyzer to determine such characteristics. The paper presents a plugin to the Wireshark traffic analyzer to calculate the moments of the random variable – the interval between packets of incoming traffic. The article also presents the analytical solution for the average waiting time for a QS type H2/M/1. Here H2 is the 2nd order hyperexponential distribution law of the input flow time intervals. The final result is obtained as a solution of Lindley's integral equation using the method of spectral decomposition. It is shown that in this case the distribution laws of intervals between input flow requirements can be approximated at the level of their three first moments. The joint use of these results allows to fully analyze the incoming traffic by queuing methods. The obtained results demonstrate the fact that the classical M/M/1 system shows optimistic results in comparison with the considered system. Therefore, the approach can be successfully applied in the modern teletraffic theory where packet delays in the incoming traffic are significant.

Keywords: traffic analyzer, wireshark program, numerical characteristics of random variables, Lindleys equation, method of spectral decomposition.

DOI: 10.15514/ISPRAS-2015-27(3)-21

For citation: Tarasov Veniamin, Malakhov Sergey. Statistical Data Handling Program of Wireshark Analyzer and Incoming Traffic Research. *Trudy ISP RAN/Proc. ISP RAS, vol. 27, issue 3, 2015*, pp. 303-314. DOI: 10.15514/ISPRAS-2015-27(3)-21.

1. Introduction

The identification of the distribution laws of intervals is particularly sophisticated problem, at the same time the traffic as a random process tends to be constantly

changing. It is known, the queuing theory is based on the laws of distribution of intervals between income and service requirements. Therefore it is important to know the numerical characteristics of these intervals or their moments. In this paper we propose to use the Wireshark analyzer to determine such characteristics [[1]].

2. Description of the program Wireshark

Wireshark (previously, Ethereal) is a traffic analyzer for Ethernet computer networking technology and some others. In June 2006 the project was renamed Wireshark due to trademark issues [[1]].

The functionality provided by Wireshark is very similar to the capabilities of the tcpdump program, but Wireshark has a graphical user interface and additional features for sorting and filtering information. The program allows the user to view all the traffic through the network in real time, shifting the network card to promiscuous mode. (Eng. Promiscuous mode) (Fig. 1).

Wireshark is an application that can display the structure of a wide variety of network protocols, and therefore allows parsing network packets, showing the value of each field protocol at any level. The use of Pcap packet capture library allows capturing data only from those networks that are supported by this library. However, Wireshark can work with multiple formats of input data an open data files captured by other programs that enhances the capture.

The features include:

- deep analysis of hundreds of protocols, with the regular addition of new ones;
- capturing network traffic in real time, followed by analysis at any time;
- standard three-pane packet browser (standard package has three regions);
- cross-platform: there are versions for most types of UNIX, including Linux, Solaris, FreeBSD, NetBSD, OpenBSD, Mac OS X, as well as for Windows;
- The captured from network information can be viewed by using the graphical user interface or by using the TTY-mode utility TShark;
- the most powerful sorting and filtering in the industry;
- a great opportunity to VoIP analysis;
- read / Write a large number of file formats capture: tcpdump (libpcap), Pcap NG, Catapult DCT2000, Cisco Secure IDS iplog, Microsoft Network Monitor, Network General Sniffer® (compressed and uncompressed), Sniffer® Pro, and NetXray®, Network Instruments Observer, NetScreen snoop, Novell LANalyzer, RAD-COM WAN / LAN Analyzer, Shomiti / Finisar Surveyor, Tektronix K12xx, Visual Net-works Visual UpTime, WildPackets EtherPeek / TokenPeek / AiroPeek, and many other;
- capture files that compressed with gzip can be unpacked immediately;

- capturing real-time data can be effected via Ethernet, IEEE 802.11, PPP / HDLC, ATM, Bluetooth, USB, Token Ring, Frame Relay, FDDI, and the other (depending on the platform);
- decoding support for many protocols, including IPsec, ISAKMP, Kerberos, SNMPv3, SSL / TLS, WEP, and WPA / WPA2;
- Highlighting rules can be applied to the package list for quick, in-intuitively analysis;
- output data can be exported to XML, PostScript®, CSV, or plain text.

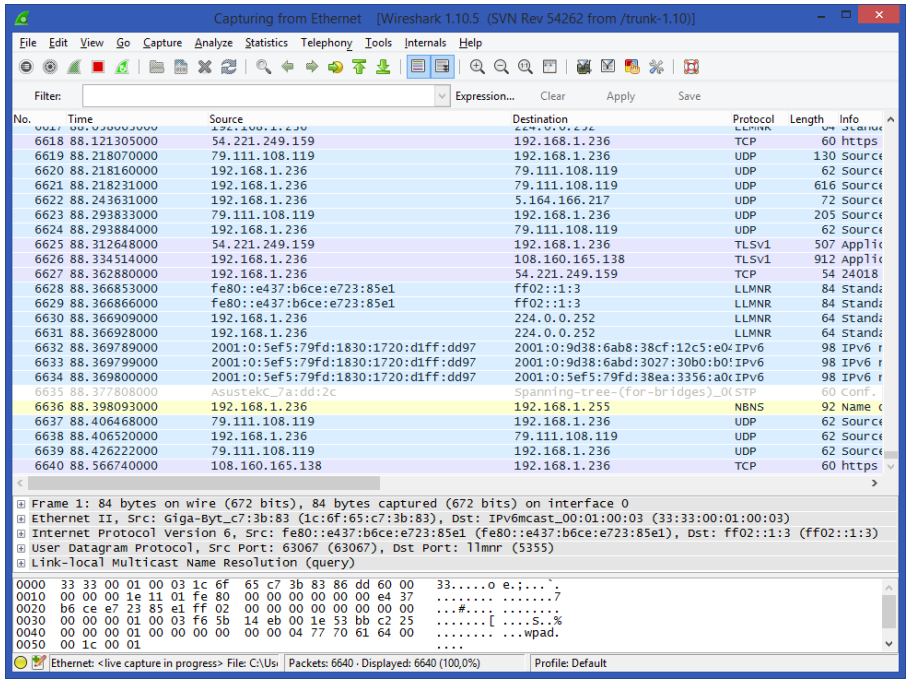

Fig. 1. The example of a network traffic capture by Wireshark.

CSV is one of the formats of data export, convenient for viewing (Fig. 2). This file can be opened in any text editor or spreadsheet editor for analysis and calculation of performance.

However, it is difficult to process the data in case of intense traffic even in the spreadsheet editor. Furthermore the traffic data can be stored in more than one file. This article describes a software solution for the calculation of the numerical characteristics of packet arrival intervals. The main advantage of this analyzer is his work on a small scale of time (microseconds), in contrast to the same program NetFlow Analyzer, which captures packets-per-minute rate.

3. Determination of the moments of the interarrival time of incoming traffic

The program developed by the authors of the present paper allows, in addition to the analyzer, to retrieve the packet arrival times, isolated the incoming traffic from the entire data set received by Wireshark. Next, using the well-known formulas of mathematical statistics, it can be defined the moment characteristics of the timing. We use the statistics to the third order statistical properties, which provides representations of the distribution of the intervals.

For example, the coefficient of variation shows the difference from a Poisson traffic flow and with asymmetry gives an indication of the degree of weight in the distribution tails.

The average value of the interval between adjacent packets

$$\bar{\tau} = \frac{1}{N}\sum_{k=0}^{N}\left(t_{k+1} - t_k\right)$$

where t_k – packet arrival times, N – the number of intervals analyzed.

Custom dispersion $D = \overline{t^2} - \bar{\tau}^2$,

where $\overline{t^2} = \frac{1}{N}\sum_{k=0}^{N}\left(t_{k+1} - t_k\right)^2$ – the second initial moment.

The coefficient of variation $c = \sigma / \bar{\tau}$, where $\sigma = \sqrt{D}$.

Asymmetry $A_s = (\overline{t^3} - 3 \cdot \overline{t^2} \cdot \bar{\tau} + 2\bar{\tau}^3) / \sigma^3$,

where $\overline{t^3} = \frac{1}{N}\sum_{k=0}^{N}\left(t_{k+1} - t_k\right)^3$.

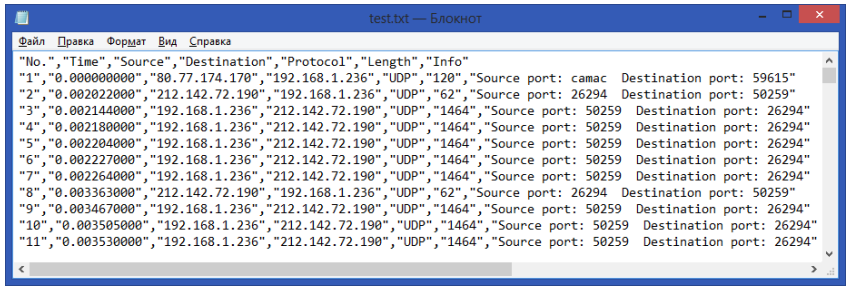

Fig. 2. The example of the data exported to the CSV format.

If a large amount of data is divided into several blocks, then these formulas are determined by the average group, and then their mean values.

4. Time data analysis software and Results

To calculate the moments of the interval between adjacent packets, we developed a program, which selects only the data related to the inbound packet from the input file, containing the capture of a network traffic data, and calculates intervals and moments.

The features include:

- sample timing of the data packets arrived at said host;
- calculation of the time intervals between the incoming packets;
- calculation of the torque characteristics for intervals of received packets;
- saving time of the data packets arrived in binary and text format;
- saving data packet arrival intervals in binary and text formats;
- output and saving torque characteristics in a text format.

The program handles text files containing the data as shown in Fig. 2 or similar.

For the program the two classes (in terms of object-oriented programming) are developed:

- TrafficLogParams – stores the packet arrival time, their intervals and calculates the torque characteristics. Also provides the methods to store and download the data from files;
- LogParser – static class that produces an analysis of the input file and adds data to the TrafficLogParams class.

The input of LogParser main method is the file name and IP-address of the host. Each line of the source file is processed and from the selected data on the time and two IP-address - the address of the sender and the recipient's address. If the recipient field matches the host IP-address, then the packet arrival time is added to the array such times in TrafficLogParams class.

```
public static TrafficLogParams TextFileParser(string fileName, string ip, bool isIncoming)
{
TrafficLogParams log = new TrafficLogParams();
StreamReader file = new StreamReader (fileName);
string[] currentLine;
int lineNumber = 0;
int ipIndex;
if (isIncoming)
        ipIndex = 2;
else
        ipIndex = 1;
while (!file.EndOfStream)
```



```
{
        currentLine = GetDataArray (file.ReadLine().Trim());
        lineNumber++;
        try
        {
        if (MinimizeIp (currentLine[ipIndex]) == MinimizeIp (ip))
        {
        log.AddTime(ParseDouble(currentLine [0]));
        }
        }
        catch (FormatException ex)
        {
MessageBox.Show(string.Format("{0}\nСтрока = {1}", ex.Message, lineNumber));
        }
}
file.Close();
return log;
}
```

The second most important method of LogParser splits the input string into elements, checking every element belonging to the format of time or IP-address, and returns them as an array.

```
private static string[] GetDataArray(string input)
{
string[] data = new string[3];
string currentValue = "";
int symbolIndex = 0;
int valueIndex = 0;
while (symbolIndex < input.Length && valueIndex < 3)
{
        while (symbolIndex < input.Length && (char.IsDigit(input[symbolIndex])
|| IsSeparator(input[symbolIndex])))
        {
        currentValue += input[symbolIndex];
        symbolIndex++;
        }
        if (currentValue != "")
        {
                if ((IsDouble(currentValue) || IsIp(currentValue)))
        {
```

```
        data[valueIndex] = currentValue;
        valueIndex++;
        }
        currentValue = "";
        if (valueIndex >= 3)
        {
        symbolIndex = input.Length;
        }
        }
        while (symbolIndex < input.Length && !char.IsDigit(input[symbolIndex])
&&
        !IsSeparator(input[symbolIndex]))
        {
        symbolIndex++;
        }
}
return data;
        }
```

The method checks if the input symbol is a separator "." or ",". Such testing is important only for the time data, as in some countries, the fractional part is separated by a comma (for example, in Russia), rather than a point. It is for the reason, when a string representation of a number is converted to its equivalent real number denoting the time, the standard method is not used programming language, and its modification depends on the regional settings.

```
private static double ParseDouble(string value)
{
if (CultureInfo.CurrentCulture .NumberFormat.NumberDecimalSeparator == ".")
{
        value = value.Replace(',', '.');
        }
        else
        {
        value = value.Replace('.', ',');
        }
return double.Parse(value);
}
```

When comparing the IP-address of the host with the IP-address on the current line of the log file to minimize the usual pro-IP-address to the general form. In other words, IP-address will be equal 010,014,000,011 10.14.0.11.

The program was used to analyze the data file of the traffic coming to the proxy server of the university with almost an hour-long data set. The input file contains

more than 2150000 rows, which could not be processed manually. Were obtained the following results (Fig. 3):

File Help	
Initial moment of the 1st order:	5,097781e-003
Initial moment of the 2nd order:	3,325837e-004
Initial moment of the 3rd order:	5,505049e-005
Dispersion:	3,065963e-004
Variation coefficient:	3,434807e+000
Asymmetry:	1,025441e+001
Packets count:	628183
Ready!	

Fig. 3. The result of the analysis program log files.

5. Research of queuing system h2/m/1

The data indicate that the analyzed traffic differs from a Poisson (coefficient of variation c = 3,43 instead of 1), the asymmetry value As = 10,25 indicates that the distribution of intervals between the packets of traffic relates to a heavy-tailed distributions. For example, for Poisson flow of As = 2. The calculation of the characteristics of such traffic requires appropriate mathematical apparatus. For the analysis of such traffic the authors of [[2]] proposed the new results for the system H2/M/1. We will describe the basic results from the article.

It is known, as example from [[3]], to study queuing systems (QS) G/G/1 the integral equation of Lindley is used:

$$W(y)=\begin{cases}\int_{-\infty}^{y} W(y-u)dC(u), & y\geq 0\\ 0, & y<0\end{cases} \quad (1)$$

where $W(y)$ is the probability distribution function (PDF), the waiting time in line requirements $C(u)$ is the PDF limiting random variable, $U=\lim_{n\to\infty} U_n = x_n - t_{n+1}$, and x_n is the time of the n-th service requirement C_n, and is the time interval between the t_{n+1} arrival of the requirements C_n and C_{n+1}.

To solve (1), a spectral method is used that reduces to using the expression $A^*(-s)\cdot B^*(s)-1$ and finding a representation as a product of two factors, which would give a rational function of s [3]. Thus, to find the latency distribution, the following spectral decomposition is used:

$$A^*(-s)\cdot B^*(s)-1=\frac{\psi_+(s)}{\psi_-(s)} \tag{2}$$

where $\psi_+(s)$ and $\psi_-(s)$ are rational functions of s, which can be factored. The functions $\psi_+(s)$ and $\psi_-(s)$ must satisfy certain conditions [3]:

1. For $\mathrm{Re}(s)>0$, the function $\psi_+(s)$ is analytic without zeros in the half-plane.

2. For $\mathrm{Re}(s)<D$, the function $\psi_-(s)$ is analytic without zeros in the half-plane, (3)

where D is a positive constant determined from the following condition:

$\lim\limits_{t\to\infty}\frac{a(t)}{e^{-Dt}}<\infty$.

Moreover, the functions $\psi_+(s)$ and $\psi_-(s)$ must have the following properties:

$$\begin{aligned}&\text{for}\quad \mathrm{Re}(s)>0 \quad \lim_{|S|\to\infty}\frac{\psi_+(s)}{s}=1;\\ &\text{for}\quad \mathrm{Re}(s)<D \quad \lim_{|S|\to\infty}\frac{\psi_-(s)}{s}=-1.\end{aligned} \tag{4}$$

We know that all the main characteristics of QSs are derived from the average waiting time, and therefore all subsequent calculations will be performed with respect to the average waiting time in the queue requirements.

Consider QS H2/M/1, where H2 designates the hyperexponential distribution 2nd order arrival time requirements in a density function

$$a(t)=p\lambda_1 e^{-\lambda_1 t}+(1-p)\lambda_2 e^{-\lambda_2 t} \tag{5}$$

and M – notation exponential law services with a density function

$$b(t)=\mu e^{-\mu t} \tag{6}$$

The Laplace transform of (5) has the form

$$A^*(s)=p\frac{\lambda_1}{s+\lambda_1}+(1-p)\frac{\lambda_2}{s+\lambda_2} \tag{7}$$

and function (6):

$$B^*(s)=\frac{\mu}{s+\mu} \tag{8}$$

Now we define (2) for the distributions (5) and (6) from (7) and (8):

$$\begin{aligned}&\frac{\psi_+(s)}{\psi_-(s)}=\left[p\frac{\lambda_1}{\lambda_1-s}+(1-p)\frac{\lambda_2}{\lambda_2-s}\right]\frac{\mu}{\mu+s}-1=\\ &=\frac{[p\lambda_1(\lambda_2-s)+(1-p)\lambda_2(\lambda_1-s)]\cdot\mu-(\lambda_1-s)(\lambda_2-s)(\mu+s)}{(\lambda_1-s)(\lambda_2-s)(\mu+s)}=\\ &=\frac{\mu(a_0-a_1 s)-(\lambda_1-s)(\lambda_2-s)(\mu+s)}{(\lambda_1-s)(\lambda_2-s)(\mu+s)},\end{aligned} \tag{9}$$

where the coefficients $a_0=\lambda_1\lambda_2$, $a_1=p\lambda_1+(1-p)\lambda_2$.

The numerator of the right side of (9) is a third degree polynomial $s(s^2-c_2 s-c_1)$, and it remains to determine the coefficients for the decomposition of the factors. The coefficients of the polynomial are:

$c_1=\mu[\lambda_1(1-p)+\lambda_2 p]-\lambda_1\lambda_2$, $c_2=\lambda_1+\lambda_2-\mu$. Then the expression (9) can be factored:

$$\frac{\psi_+(s)}{\psi_-(s)}=\frac{s(s^2-c_2 s-c_1)}{(s-\lambda_1)(\lambda_2-s)(\mu+s)}=\frac{s(s+s_1)(s-s_2)}{(s-\lambda_1)(\lambda_2-s)(\mu+s)},$$

where $-s_1=-(\sqrt{{c_2}^2/4+c_1}-c_2/2)$ is the negative root of the quadratic equation in the numerator, and is the $s_2=\sqrt{{c_2}^2+c_1}+c_2/2$ positive root.

Further, omitting some calculations, we obtain the Laplace transform of the density function of the waiting time: $W^*(s)=\frac{s_1(s+\mu)}{\mu(s+s_1)}$. Hence $\frac{dW^*(s)}{ds}=\frac{s_1\mu(s_1+s)-s_1(s+\mu)\mu}{\mu^2(s+s_1)^2}$. Using the properties of the Laplace transform, we find that the average waiting time is

$\overline{W}=-\left.\frac{dW^*(s)}{ds}\right|_{S=0}=\frac{-s_1^2\mu+\mu^2 s_1}{\mu^2 s_1^2}=\frac{1}{s_1}-\frac{1}{\mu}$. Finally, the average waiting time is

$$\overline{W}=\frac{1}{s_1}-\frac{1}{\mu} \tag{10}$$

where $s_1=\sqrt{c_2^2/4+c_1}-c_2/2$, $c_1=\mu[\lambda_1(1-p)+\lambda_2 p]-\lambda_1\lambda_2$, $c_2=\lambda_1+\lambda_2-\mu$.

6. Practical use of the results

Consider the result (10) for example, the input distribution, with a heavy tail (fig. 3). Using the Laplace transform (7) we can determine the initial moments of the distribution (5):

$$\begin{cases} \overline{\tau}_{\lambda} = \dfrac{p}{\lambda_1} + \dfrac{(1-p)}{\lambda_2} \\ \overline{\tau_{\lambda}^2} = \dfrac{2p}{\lambda_1^2} + \dfrac{2(1-p)}{\lambda_2^2} \\ \overline{\tau_{\lambda}^3} = \dfrac{6p}{\lambda_1^3} + \dfrac{6(1-p)}{\lambda_2^3} \end{cases}.$$

Next, substituting the results obtained in step 1 from the initial moments of the distribution of intervals between bursts to determine the unknown parameters of the input distribution (5): λ_1, λ_2 and p, we obtain the following system of equations:

$$\begin{cases} \dfrac{p}{\lambda_1} + \dfrac{(1-p)}{\lambda_2} = 5.0978e-003 \\ \dfrac{2p}{\lambda_1^2} + \dfrac{2(1-p)}{\lambda_2^2} = 3.3258e-004 \\ \dfrac{6p}{\lambda_1^3} + \dfrac{6(1-p)}{\lambda_2^3} = 5.5050e-005 \end{cases} \quad (11)$$

The solution of (11) in the package Mathcad yields the following results: p≈ 0.950, $\lambda_1 \approx 417.985$, $\lambda_2 \approx 17.556$.

In case of load of the channel equals to 0.4, intermediate parameters: $c_1 \approx 10999{,}4$; $c_2 \approx$ -54.655, $s_1 \approx 135.707$ and the average waiting time $\overline{W} \approx 5.329 \cdot 10^{-3}$ s.

For comparison, let us look to the average waiting time for an M/M/1 system. In this case, the intensity of service equals to $\mu \approx 490.196$, and the channel loading $\rho = 0.4$.

Then the average waiting time of packets $\overline{W} = \dfrac{\rho/\mu}{1-\rho} = \dfrac{0.4/490.196}{1-0.4} = 1.36 \cdot 10^{-3} s$.

Thus the queuing model taking into account the distribution and its weight in the tail of the input, gives a delay about four times larger than the classical model.

7. Conclusion

This paper has presented how optimistic are the results given by classical M/M/1 system in comparison to the system in the case of high H2/M/1 weightiness tail of the distribution of the input stream. Therefore, the approach can be successfully applied in the modern teletraffic theory where packet delays in the incoming traffic are significant.

Note that the distribution, which contains three unknown parameters λ_1, λ_2 and p, allows to use the moment equations to approximate the unknown input distribution in the first three moments.

References

[1]. Wireshark official web-site URL: http://www.wireshark.org/
[2]. Tarasov V.N., Bakhareva N.F., Gorelov G.A. Mathematizheskaya model trafica s tyazhelohvostnym raspredeleniem na osnove sistemy massovogo obsluzhivaniya H2/M/1. [Mathematical model of traffic from heavy-tailed distributions with based queuing system H2/M/1]. Infocommunicationye tehnologii, 2014, no. 3, pp.36-41.
[3]. Kleinrock L. Queueing Theory. Tran. from English. edited by V.I. Neumann. M. Mechanical Engineer-ing, 1979.

Программа статистической обработки данных анализатора wireshark и исследование входящего трафика

Вениамин Тарасов <tarasov-vn@psuti.ru> ,
Сергей Малахов <malakhov-sv@psuti.ru>
ПГУТИ, 443090, Россия, г.Самара, Московское ш., д 77

Аннотация В работе представлена программа-дополнение к анализатору трафика Wireshark для расчета моментов случайной величины - интервала между пакетами входящего трафика. Приведено аналитическое решение для среднего времени ожидания для СМО типа H_2/M/1. Здесь H_2 - гиперэкспоненциальный закон распределения 2-го порядка интервалов времени входного потока. Конечный результат получен путем решения интегрального уравнения Линдли методом спектрального разложения. Показано, что в этом случае законы распределения интервалов между требованиями входного потока можно аппроксимировать на уровне их трех первых моментов. Совместное использование этих результатов позволяет полностью анализировать входящий трафик методами массового обслуживания.

Ключевы слова: анализатор трафика, программа Wireshark, числовые характеристики случайной величины, интегральное уравнение Линдли, метод спектрального разложения.

DOI: 10.15514/ISPRAS-2015-27(3)-21

Для цитирования: Тарасов Вениамин, Малахов Сергей. Программа статистической обработки данных анализатора wireshark и исследование входящего трафика. Труды ИСП РАН, том 27, вып. 3, 2015 г., стр. 303-314 (на английском языке). DOI: 10.15514/ISPRAS-2015-27(3)-21.

Список литературы

[1]. Wireshark official web-site URL: http://www.wireshark.org/
[2]. В.Н. Тарасов, Н.Ф. Бахарева, Г.А. Горелов «Математическая модель трафика с тяжелохвостным распределением на основе системы массового обслуживания H_2/M/1» // Инфокоммуникационные технологии, 2014 г., №3, с.36-41.
[3]. Клейнрок Л. Теория массового обслуживания. Пер. с англ. под редакцией В.И. Неймана. М. Машиностроение, 1979. – 432 с.