Modelling of connection arrivals in Ethernet-based data networks

Thomas Kessler; Andreas Ahrens; Christoph Lange; Hans-Dietrich Melzer
Institute of Communications Engineering
University of Rostock, Germany

Abstract

This contribution discusses approaches for modelling Internet traffic on the basis of so-called regenerative models. Temporal interval distributions of consecutive data packets were measured in the network of the University of Rostock. The gained dataset allows to present suitable functions for the characterization and modelling of Internet traffic. As a result of the investigations it is shown, that regenerative approaches provide sufficiently exact results for modelling Internet data traffic.

1. Introduction

The increasing electronic transfer of information in almost all parts of the modern society leads to expanding demands on the efficiency of communication networks. These ever increasing needs for reliable and fast data transmission requires an anticipatory design of communication systems, in order to use network components and transmission channels efficiently. For the design of future networks it is necessary, to predict the characteristics of traffic in data networks. The data exchange in information networks – as the Internet – is based on packet-oriented transport. This data traffic can be characterized by several parameters (e.g. the temporal intervals between packets or the packet sizes). A suitable parameter often used in this context, is the distribution of the Ethernet packet interarrival times, which is subjected to temporal fluctuations. A lot of user activities can lead to a transport of long packet groups; the organization of the packet stream (e.g. securing the correct packet sequence) is provided by higher layer protocols. According to the Internet service used, the number of packets directly belonging together can vary significantly: When transferring files by FTP (File Transfer Protocol) a large number of packets belongs to the connection, whereas a Telnet transmission requires relatively few packets.

Preceding investigations, for example in [5], have shown, that the mean time interval between the arrival of two consecutive packets only characterizes the arrival process very roughly. Therefore this mean time interval alone is not sufficient to describe the data traffic. Typical approaches frequently use a characterization of the packet arrival as a stochastic process with several parameters [5, 8].

This contribution is organized as follows: Section 2 introduces the basics of classical models with their parameters. In section 3 a special regenerative model approach is discussed, which leads to useful results in modelling radio channels [12, 2, 1]. As an application, in section 4 the parameters of several distribution functions are estimated based on measurements, which were taken in the network of the University of Rostock. Finally, section 5 provides some concluding remarks.

2. Classical model basics

Important properties of packet-oriented transmission in data networks can be characterized by temporal intervals between data packets. The analysis of the measurements of data connection arrivals in the network of the University of Rostock (described in section 4 and [7]) shows, that the temporal intervals between arriving Ethernet packets can be specified by a gap process. Characteristic parameters of such a gap process are the gap density \( p(k) \) and the gap distribution \( F(k) \). The gap distribution indicates the probability, that a gap \( X \) is greater than or at least equal to a given number \( k \):

\[
F(k) = P(X \geq k) , \quad k = 1, 2, 3, \ldots , \quad (1)
\]

whereas the gap density

\[
p(k) = P(X = k) , \quad k = 1, 2, 3, \ldots , \quad (2)
\]

denotes the probability, that a gap \( X \) of length \( k \) appears.

Frequently used and well-suited practical approximations are provided, if the model is based on the independence of the gap intervals [12, 8]. These models are completely described by the gap density or the gap distribution, respectively.

Table 1 provides some characteristic probability distribution functions and probability density functions, which will be used in this contribution [3, 10, 5]. Thereby infinite gap lengths \( k \) are allowed. In a simulation they have to be

\[1\] The discrete variable \( k \) is given through the time resolution of the measurement system.
limited appropriately. Figure 1 illustrates these functions with exemplary parameters ($\beta_e = 1/2$, $\alpha_w = 1$, $\beta_w = 1$, $\beta_r = 2/3$).

<table>
<thead>
<tr>
<th>Type</th>
<th>Probability density $p(k)$</th>
<th>Distribution $F(k)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>$\beta_e e^{-\beta_e k}$</td>
<td>$e^{-\beta_e k}$</td>
</tr>
<tr>
<td>Weibull</td>
<td>$\beta_w (-\beta_w k)^{\alpha_w - 1} e^{-(\beta_w k)^{\alpha_w}}$</td>
<td>$e^{-(\beta_w k)^{\alpha_w}}$</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>$\beta_r^2 k e^{-\beta_r^2 k^2}$</td>
<td>$e^{-\beta_r^2 k^2}$</td>
</tr>
</tbody>
</table>

The fundament of the further investigations is the definition of the probability $p_p(n)$. This probability describes the arrival of at least one Ethernet packet within an interval $n$. With the assumption that the down-times of the system and the packet-length are neglected, the probability $p_p(n)$ can be determined over the gap distribution $F(k)$ of successive packets in the following form:

$$p_p(n) = p_d(n) \cdot \sum_{k=0}^{n-1} F(k) .$$

(3)

Thereby, the parameter $p_d(n)$ describes the average probability for the arrival of Ethernet packets. This is illustrated in Fig. 2.

3. Regenerative models according to Wilhelm and Ahrens

In this contribution, the statistical independence of the individual data packets is presumed. This is an assumption, which is frequently regarded as a good practical approximation [12] and in our case it seems justified, since many independent data connections were watched and evaluated at the test point in the net of the university. In addition, the affiliation of the packets to definite connections isn’t taken into account. Simultaneously, all data packets are considered equal rights and to have be independent from each other. Therefore the complete information lies in the time of the arrival or in the temporal gaps between the arrival of packets. We neglect at this point that Ethernet packets can have various lengths (64 Byte to 1518 Byte [6]). Under these prerequisites the temporal distances between the incoming data packets can be considered as significant parameter for the description of the arriving traffic in an Ethernet based net.

As shown in [12], for the determination of the distribution function of the gaps between successive errors in radio channels, like the short wave channel, the approach

$$\sum_{k=0}^{n-1} F(k) = n^\alpha \quad 1 \leq n \leq n_{\text{max}}$$

(4)

was chosen. This approach was further investigated by Ahrens in [2]. The parameter $\alpha$ in (4) defines the traffic intensity and the value $n_{\text{max}}$ the maximal possible gap length. Using (4), one gradually yields the gap distribution

$$n = 1 : F(0) = 1$$
$$n = 2 : F(0) + F(1) = 2^\alpha$$
$$n = 3 : F(0) + F(1) + F(2) = 3^\alpha$$
$$\vdots$$
$$n \leq n_{\text{max}} : F(0) + F(1) + \cdots + F(n - 1) = n^\alpha .$$

Therewith, the distribution function $F(k)$ can be described in compact notation as follows:

$$F(k) = \begin{cases} (k + 1)^\alpha - k^\alpha & 0 \leq k < n_{\text{max}} \\ 0 & k \geq n_{\text{max}} \end{cases}.$$
A traffic-intensity parameter $\alpha = 1$ with (5) yields

$$F(k) = \begin{cases} 1 & 0 \leq k < n_{\text{max}} \\ 0 & k \geq n_{\text{max}} \end{cases}.$$  

(6)

The gap distribution function $p(k) = P(X=k)$ leads with

$$F(k) = p(k) + p(k+1) + p(k+2) + \cdots$$  

(7)

and

$$F(k+1) = p(k+1) + p(k+2) + \cdots$$  

(8)

to the result

$$p(k) = F(k) - F(k+1).$$  

(9)

Furthermore we yield

$$p(k) = \begin{cases} 1 & k = (n_{\text{max}} - 1) \\ 0 & k \neq (n_{\text{max}} - 1) \end{cases}.$$  

(10)

The weakness of this model attempt is clearly recognizable from (10). It leads to a deterministic gap process. The probability density degenerates into a discrete spectral line at $(n_{\text{max}} - 1)$. Minimizing the weakness of this model at $p(n_{\text{max}} - 1)$ different solutions were presented in [12] and [2]. In [12] equation (5) is multiplied with the value $e^{-\beta k}$.

For $F(k)$, this leads to the setup

$$F(k) = [(k + 1)^\beta - k^\beta] \cdot e^{-\beta k} \quad \text{for} \quad k = 0, 1, 2, 3, \ldots,$$  

(11)

with the auxiliary condition $\lim_{k \to \infty} e^{-\beta k} = 0$ for $\beta > 0$. The analysis of the model qualities leads to the result, that traffic-intensity parameters in the range of $0.5 \leq \alpha \leq 1$ are possible.

A smaller traffic factor describes a higher exploitation of the net, since in this case, the probability $p(0)$ increases, that after a packet in the distance of $k = 0$ the next packet arrives. These coherences illustrates figure 3. The proposed channel model is described by two parameters and also mathematically manageable well. The assumption that successive gaps are statistically independent is regarded as a good practical approximation. Models with this requirements are described as regenerative models in the literature [12, 2].

![Figure 3: Gap distribution function $F(k)$ for short gaps $k$](image)

**4. Practical Application**

The investigations rely on measurements, which were taken by a measuring system within the data network of the University of Rostock. The measurements and the measuring system are presented in [7]. The measuring point is located on the net-sided interface of the VPN concentrator (virtual private network) of the University of Rostock. The aggregate data traffic from and into the Internet of 10 halls of residence and of the WLAN (Wireless Local Area Network) of the University of Rostock are merged at this point. Here, on average 300 out of approximately 15000 potential users are active at the same time. Before further data processing, the measured data are made anonymous. A detailed explanation of the IP traffic characteristics is given in [7].

The data are gripped by a 1000BaseSX monitor interface. They are recorded for both communication directions separately. Caused by the user behaviour, the bandwidth is exploited asymmetrically, as it is typical in IP access networks: The data traffic in the downstream (from the Internet to the user) is much higher than in the upstream (from the user to the Internet). For the investigations in this contribution, packet arrival times of $N = 2047$ or $N = 65535$ consecutive Ethernet packets with IP payload are recorded, respectively.

For the purposes of mathematical modelling of the distribution of the intervals between arriving Ethernet packets, the arrival of the data packets is understood as a stochastic process. In principle, any time interval which is greater than a minimum time interval can appear between two data packets. According to [6] the least possible length of an Ethernet packet is 64 byte, which leads to a minimum time interval between two consecutive packets of $\Delta t_{\text{min}} = 512$ ns at a transmisson speed of 1 Gbit/s. An interframe spacing is added by the particular Ethernet implementation, which is in general not specified by the manufacturer. Additional idle times between two consecutive packets caused by the CSMA/CD (Carrier Sense Multiple Access/Collision Detection) do not appear, because the connection between the monitor interface and the measuring system works in full-duplex mode.

A time stamp with resolution $t_a = 2^{-32}$ s [4] was used for determining the packet arrival times. By that, a time pattern of $k t_a$ is introduced and the random variable becomes discrete. The dead time $\Delta t_{\text{min}}$ is not considered in the following investigations and a normalization was done.
in that way, that in each measured data sample a gap of the length zero appears.

The aim of the investigation is to approximate the measured gap interval distribution \( F_m(k) \) by a suitable distribution function \( F(k) \). Based on this, the statistical properties of Internet traffic are characterized with relatively few parameters and networks can be simulated (e.g. for design purposes).

As a quality parameter for a "good" approximation between the measured data \( F_m(k) \) and a given distribution function \( F(k) \) the mean square error

\[
E_{\text{min}} = \sum_{k=0}^{K_{\text{max}}} |F(k) - F_m(k)|^2
\]  

is used and minimized [11, 9]. The value \( K_{\text{max}} \) describes the maximum gap length considered in the model.

The figures 4 and 5 show the measured gap distribution functions together with approximating distribution functions. As a result it can be stated, that the Weibull distribution approximates best the measured gap distributions in the sense of minimizing the mean square error. Within the tables 2 and 3 the optimal parameters of each distribution function are given together with the minimum energy of error according to (12).

According to these results, the Weibull distribution is well suited for modelling the statistical properties of Internet traffic by two parameters (\( \alpha_w \) and \( \beta_w \)), as can also be seen in [5]. Based on this, a simulation of data traffic can be established.

Table 2: Optimum parameters of the distribution functions
(Measured: 31.03.2003, \( N = 2047 \))

<table>
<thead>
<tr>
<th>distribution</th>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( E_{\text{min}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>-</td>
<td>0.00095</td>
<td>34.38</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.715</td>
<td>0.00100</td>
<td>24.15</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>-</td>
<td>0.00150</td>
<td>87.32</td>
</tr>
<tr>
<td>Wilhelm</td>
<td>0.974</td>
<td>0.00075</td>
<td>29.99</td>
</tr>
</tbody>
</table>

Besides the Weibull distribution the Wilhelm distribution shows a good compliance with the measured data. It can be stated, that the statistical properties of the packet arrival process are well approximated by distribution functions with two parameters, whereas distribution functions with only one single parameter not provide such a good result. Thus, the mean time arrival of two consecutive packets is not sufficient to describe the packet arrival process.

The aim of this contribution was, to find model approaches for creating gap structures and the description of their characteristic properties with a required accuracy. If the gap process is created according to (11), a uniformly distributed random variable \( Y \) is equated with the value \( F(k) \) and the corresponding gap \( k \) can be determined. The gap process can be understood as a sequence of intervals and therefore the gap distribution or the gap density ensures a sufficient description of the regenerative model, respectively. Besides the modelling of the Ethernet packet arrival times the prediction of characteristic parameters (e.g. the traffic factor \( \alpha \)) plays a dominant role. With the assumption, that these model approaches are only valid for limited time intervals, the prediction of dominant parameters becomes more important. First solution approaches can be found in [1]. Appropriate strategies are not investigated in this contribution and remain for further study.
Table 3: Optimum parameters of the distribution functions (Measured: 04.04.2003, \(N = 65535\))

<table>
<thead>
<tr>
<th>distribution</th>
<th>(\alpha)</th>
<th>(\beta)</th>
<th>(E_{\text{min}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential</td>
<td>-</td>
<td>0.00095</td>
<td>22.31</td>
</tr>
<tr>
<td>Weibull</td>
<td>0.838</td>
<td>0.00098</td>
<td>19.83</td>
</tr>
<tr>
<td>Rayleigh</td>
<td>-</td>
<td>0.00140</td>
<td>65.59</td>
</tr>
<tr>
<td>Wilhelm</td>
<td>0.988</td>
<td>0.00098</td>
<td>21.27</td>
</tr>
</tbody>
</table>

5. Conclusion

The Internet data traffic depends on packet-oriented transmission and it can be characterized by temporal intervals between arriving data packets and the packet length. In this contribution, only the packet interarrival times are considered. Based on measured gap interval distributions, suitable distribution functions and their parameters are determined. As a quality parameter, the minimum mean square error was used. Especially the Weibull distribution is well suited for describing the statistical properties of data traffic in Ethernet-based networks. The regenerative approach according to Wilhelm provides useful results, too. However, the applicability of this model concerning the simulation of Ethernet traffic deserves further study.

The measured data were determined over relatively short time intervals (seconds or minutes): Therefore the validity of the results is limited to these short time ranges. Whether the results can be assigned to longer time ranges has to be clarified and if necessary verified in further investigations. Therefore, data on long time network traces over several days and weeks (depending on the amount of network traffic and in consideration of the capacity of our measurement system) and in significant different networks in regards of size and business has already been collected and is currently analyzed.

Based on the results of this contribution, packet-oriented data traffic with given parameters can be simulated for e.g. planning and designing communication networks. To accomplish this in an increasingly powerful way the next steps in our research are to look deeper into the structure of Internet traffic. That means in particular to asssing each IP (Internet Protocol) packet to its data flow which is characterized by source and destination IP address, protocol number and the transport layer protocol port number. This approach allows for statistical analysis of concurrent connections within the traffic. Furthermore if it is possible to assign data flows to network applications such as WWW (World Wide Web) or Peer-to-Peer applications (e.g. eDonkey, Napster) and to certain users we should be able to create models fulfilling the needs of application based QoS (Quality of Service) and SLA (Service Level Agreement) planning.

References