Modeling Call Holding Time Distributions for CCS Network Design and Performance Analysis

Vladimir A. Bolotin, Senior Member, IEEE

Abstract—The message traffic offered to the CCS signaling network depends on and is modulated by the traffic characteristics of the circuit switched calls supported by the CCS network. Most previous analyses of CCS network engineering, performance evaluation and congestion control protocols generally assume an exponential holding time of circuit switched calls. Analysis of actual holding time distributions in conversations, facsimile and voice mail connections revealed that these distributions radically differ from the exponential distribution. Especially significant is the large proportion of very short calls in real traffic in comparison with the exponential distribution model. The diversity of calls (partial dialing, subscriber busy, no answer) and services results in a multi-component call mix, with even larger proportion of short time intervals between message-generating events. Very short call holding times can have a significant impact on the traffic stream presented to the CCS network: for calls with short holding times, the different CCS messages arrive relatively close to each other, and this manifests as burstiness in the CCS traffic stream.

I. INTRODUCTION

The message traffic offered to the CCS signaling network depends on and is modulated by the traffic characteristics of the circuit-switched calls supported by the CCS network. Most previous analyses, including those associated with CCS network engineering, performance evaluation, and congestion control protocols, generally assume Poisson arrivals of circuit-switched calls having an exponential holding time with the mean of approximately 3 minutes. The effects of variations from these assumptions on CCS network capacity and performance have been largely unexamined.

In particular, very short call holding times can have a significant impact on the traffic stream presented to the CCS network: for calls with short holding times, the different CCS messages in a call (IAM through RLC) arrive relatively close to each other, and this manifests as burstiness in the CCS traffic stream. Other short intervals between events in the progress of a call also result in the correlated CCS messages. Among these events are call abandonment at an early stage, subscriber busy, or no-answer conditions.

An example described in Section II shows that, at 17% overload, the SS7 channel throughput drops by 28% under exponential holding time distribution and by 46% under actual holding time distribution, which manifests a much larger proportion of short calls in comparison with the exponential distribution. This example confirms that, for CCS performance studies, the dependence on the holding time distribution is important and needs to be taken into account by the use of appropriate probability distribution models; in particular, for precise quantification of the proportion of short calls. Other examples—call back feature and PCS application—mentioned in Section II also indicate the importance of the holding time distribution for CCS traffic characterization to be used in performance analysis.

Analysis of actual holding time distributions revealed that these distributions radically differ from the exponential distribution. Some of these differences are discussed in Section III. Especially emphasized is the large proportion of very short calls in real traffic in comparison with the exponential distribution model.

The diversity of calls (partial dialing, subscriber busy, no answer) and services results in a multicomponent call mix, with even larger proportion of short time intervals between message-generating events. Section IV discusses the components of various holding times in realistic models of circuit-switched call mix.

Among the various holding time components described in Section IV, one specific component, the connection holding time (the time from connect to release), is presented in a large majority of calls—in all completed calls. Section V completes the description of the holding time models for use in CCS performance analysis by providing models of the connection holding time with surprisingly accurate agreement to the actual data. The roots of these models and the logical steps that led to them are being published elsewhere [1].

Altogether, this paper
- demonstrates the importance of the holding time distribution and inadequacy of the exponential distribution for CCS network design and performance analysis, and
- makes models for the call holding time probability distributions available for use in CSS network performance studies.

II. IMPORTANCE OF HOLDING TIME DISTRIBUTIONS FOR CCS PERFORMANCE ANALYSIS

One important aspect of call holding time distribution in CCS performance studies is illustrated by the crucial role of this distribution in a study of focused overloads of SS7 channels.

The simulation model used by D. Smith [2] to examine focused overloads on SS7 channels included a group of "source" switches originating traffic to a "target" switch. To investigate interaction between the "sources" and the "target," traffic originated by the target switch was included in the...
model.) Several important effects were observed in that study (see [2]); for the purpose of this paper, we report the following effect on the holding time distribution on the SS7 channel throughput.

Fig. 1 illustrates the throughput of the SS7 channel as a function of the source load for different call holding time distributions, the target load being constant. The source load is measured by the call rate of the first attempts (repeated attempts were included in these simulations). The offered call rate at which the maximum SS7 channel throughput for source origins occurs is about 300 \times 10^3 calls/hour.

The next two data points of the offered call rate are at about 320 \times 10^3 and 350 \times 10^3 calls/hour; they indicate the first attempt overload about 7% and 17%. At these overloads, the throughput drops significantly—from maximum of 269 \cdot 10^3 to 194 \cdot 10^3, 160 \cdot 10^3, and 146 \cdot 10^3 for the three different holding time distributions.

The holding time distributions for the calls counted in the throughput in Fig. 1 are:

- Realistic distribution constructed as described, in Sections IV and V, with the mean holding time 150 s. (Specifically, 11% busy (3 s), 11% no answer (10 s), and 78% completed calls; the latter are a mixture of 3% short calls (1.5 s) and 97% calls with the holding time distributed according to formula (5.2) with the parameters \( \alpha = 0.4, \mu_1 = 1.31, s_1 = 0.33, \mu_2 = 2.11, s_2 = 0.5. \)

- Exponential distribution with the same mean value.

- A distribution with short holding time reflecting overloads caused by mass media events; the exact distribution is uniform between 5 and 15 s.

The larger the proportion of short calls in the call mix, the more correlation and burstiness is introduced into the message stream, consequently with more impact on the system performance. For example, the probabilities that the holding time does not exceed 5 s are about 14% in a realistic distribution and about 4% in the exponential distribution. Consequently, the realistic distribution manifests a deeper drop of the throughput in overloads: 146 \cdot 10^3 versus 194 \cdot 10^3 in Fig. 1, i.e., a throughput loss of 46% versus 28%. The message traffic on the SS7 channel, being modulated by the circuit-switched calls, is especially affected by calls with short holding times.

Other potential effects of the holding time distribution may be found in new services, where signaling network traffic is generated for information exchange between switches and data bases. This traffic may depend on the call holding time distribution.

For instance, in automatic call back, the status of the called party line is monitored, i.e., periodically checked, in order to notify the caller when the line becomes free. In the call back systems with originating scanning, messages querying the status of the called party are sent at regular intervals to the terminating switch and responses are sent in the opposite direction. As was shown in a study at Bellcore, the total number of queries, i.e., the signaling traffic, greatly depends on the connection holding time.

In PCS, updates of location and status information of moving subscribers occur. Depending on the network structure, these updates may result in traffic through the signaling network. For updates that occur during established connections, the amount of traffic would depend on the connection holding time, e.g., the longer the connection the more likely that updates would occur.

These examples and, especially, the crucial effect of short calls on the SS7 throughput performance call for accurate representation of the real circuit-switched call holding time characteristics as a function of service type. The remaining part of this paper provides foundations for such characterization in various call mixes.

III. THE EXPONENTIAL MODEL

First, we show the inadequacy of the exponential distribution.

A typical example of the cumulative connection holding time distribution is illustrated in Fig. 2. This example is taken from a sample of half-million completed calls recorded in
Circuits in a telephone network are used for various kinds of call attempts and calls, with a wide range of holding times, from seconds to tens and even hundreds of minutes, depending on the call (attempt) type. Correspondingly, the holding time distribution of a call mix is a mixture of various holding time distributions \( F_i(x) \):

\[
F(x) = \sum p_i F_i(x), \quad \sum p_i = 1. \quad (4.1)
\]

The randomization probability \( p_i \) in this mixture corresponds to the proportion of the \( i \)th class calls in the mixture, \( F_i(x) \) is the holding time distribution in the \( i \)th class. To be specific, we indicate the five most typical and quantitatively significant classes of the total call flow in a typical circuit group. (Others could be added depending on the actual circumstances—e.g., circuits that are used predominantly for short calls such as credit and verification, circuits in overload when the call mix is significantly different from the one at normal load, etc.) The five classes are:

1. Various abandonments before the connection over the trunk group is established.
2. Outgoing calls encountering busy condition and being abandoned by the caller.
3. Outgoing calls encountering no-answer condition and being abandoned by the caller.
4. Outgoing calls encountering a busy and no-answer condition and leaving a voice mail message.
5. Answered, i.e., completed, calls which divide into many classes such as business, residential, facsimile, voice mail, and circuit-switched data.

Notice that the five classes are enumerated in the order of the increasing average holding time. The proportion of calls in the first class is probably several percent; the holding time is on the order of seconds. Typically, the next three classes together account for more than 20% of all calls, with most of this in classes 2 and 3 and the approximate average holding time under 5 and 15 s, correspondingly. The voice mail messages are, on average, shorter than 30 seconds.

Two sequential holding times are generated by a call in classes 4 and 5—the ringing time and the connection holding time (the time from connect to release). Currently, no data are available for the distribution function of the ringing time. An approximate model is a uniform distribution between 0 and 18 s.

The classification of calls and call attempts on a circuit group suggests principles of realistic holding time characterization, along with ways of modeling holding times in simulation studies and, in addition, emphasizes the fact that the share of short holding times in a realistic call mix is radically larger than the share quantified by a single one-distribution model (such as the exponential model).

### IV. Modeling Call Time Distribution in a Call Mix

#### A. Holding Time Distribution for All Calls

For realistic modeling of intervals between messages on a SS7 channel, a variety of holding times should be taken into account. As a general example, a reasonable and manageable approach would be the following characterization of calls holding times in a circuit-switched route supported by a CCS channel.

#### B. The Connection Time Distribution

The connection holding time (the time from connect to release) is present in the majority of calls and, in contrast to all other time intervals, is spread over a wide range of values from seconds to hours. A good knowledge of the connection time distribution is needed to make it possible to sort out
all completed calls according to their holding time. This is why studies of various time distribution components in mixture (4.1) begin with analysis of the connection time distribution. Section V contains recommendations that add this most important component to formula (4.1). In Section V, we outline the results of a comprehensive analysis and modeling of the connection holding time distribution in order to make this knowledge available for CCS/SS7 studies. The complete description of the analysis is being published elsewhere [1].

V. MODELS FOR CONNECTION HOLDING TIME DISTRIBUTION IN COMPLETED CALLS

A. Connection Holding Time Distribution on a Single Line

The probability model for connection holding time is based on two notions. One is a human perception of time on a logarithmic scale, and the other is a fundamental fact that the holding time distribution in a call mix is a mixture of distributions. In a group of subscribers, individual subscriber holding times combine and the total time, as a random variable, becomes a mixture of individual distributions with different parameters.

In Fig. 3, a histogram of the one-thousand-call sample is shown (cut at 600 s), along with the exponential density “fit” superimposed on the histogram. To enhance the presentation of the exponential distribution model given in Section III, we use the density rather than the cumulative distribution function. The obvious discrepancies in this fit are: the exponential distribution misses the significant rising part of the histogram—from 0 to 10–20 s, and the density drops below 1 at 517 s ignoring the 17% tail over 600 s. In the range between 60 and 300 s, the exponential density does not indicate a good fit, either. The coefficient of variation of the total sample is 1.7, significantly larger than the value 1 of the exponential distribution. Thus, the exponential distribution function deviates from the empirical distribution function way beyond any statistically reasonable limits.

The change from the linear time scale to a logarithmic time scale turns the histogram in Fig. 3 into the histogram in Fig. 4. The x-axis is expressed here as the (common) logarithm of the holding time. Excluded from Fig. 4 are 20 calls with holding time under 3 s. These very short calls (“erroneous connections”) form a distinctly separate distribution which is characterized in section 5.2. Analysis of the remaining sample of 980 calls shows that the single-subscriber holding time can be modeled by the normal distribution, with surprising accuracy, as illustrated in Figs. 4 and 5. The parameters shown in these figures refer both to the natural time scale (T) and logarithmic time scale (τ), so that the normal density and distribution curves N(μ, σ) are drawn with μ = E[log10(T)] = 1.80 and σ = E[log10(T)] = 0.46.

The closeness between the empirical distribution and the fitting normal distribution is clearly seen in Fig. 5, where the middle stepwise curve is the empirical distribution function, the upper and the lower stepwise curves indicate the two-sided simultaneous contours for the 5% error probability (Kolmogoroff–Sminoff test), and the dotted line is the normal distribution function.

B. Connection Holding Time Distribution in a Homogeneous Line Group

The following characterization of the holding time distribution in groups of subscriber lines is based on a comprehensive study [1] of a complete set of individual local originating call records for hundreds of residential lines in a geographical area for one month. The total number of individual call records is 530,786. The holding time distribution was analyzed for units of data; each unit contained all individual call records for a single-hour period (700–1500 calls/hours in the afternoon hours, 1500–2000 calls/hour in the evening). The sample mean holding time varies between 2.5 and 7 minutes, the coefficient of variation is about 2, with some variation within 1.8–2.2.
The distribution function of the connection holding time may be modeled, with great accuracy, by a mixture of three random variables:

\[
F_3(x) = p_3 \cdot F_1(x) + (1 - p_3) \times \left( \alpha \cdot F_1(x) + (1 - \alpha) \cot F_2(x) \right).
\]

(5.1)

\(F_3(x)\) characterizes very short calls (under 3 s). The proportion \(p_3\) of these calls varies about 1–3%, and the holding time distribution is complex (close to a mixture of two normal distributions on a linear time scale) but can be approximated by a uniform distribution (linear time scale). In most cases, accepting a fixed value of 1.5 s will be satisfactory. The connection time of all other calls is modeled by a mixture of two normal distributions on logarithmic time scale with the mixing (randomizing) probability \(\alpha\):

\[
F(x) = \alpha \cdot F_1(x) + (1 - \alpha) \cdot F_2(x).
\]

(5.2)

Denote the three first moments of the mixing distributions by \(\mu_i, \sigma_i^2, \gamma_i^3\) \((i = 1, 2)\) and the moments of the mixture \(F(x)\) by \(\mu, \sigma^2, \gamma^3\).

It can be shown that

\[
\alpha = \frac{\mu_2 - \mu}{\mu_2 - \mu_1} = \frac{\mu_2 - \mu}{\Delta}, \quad \Delta = \mu_2 - \mu_1
\]

(5.3)

and, for symmetric distributions \(F_i(x)\) \((\gamma_i = 0)\),

\[
\sigma_i^2 = \sigma^2 + \frac{\gamma_i^3}{3(\alpha - i + 1)\Delta} = \frac{4}{3}(1 - \alpha)\Delta^2 + \frac{(\alpha + i - 1)^2}{3}\Delta^2, \quad i = 1, 2.
\]

(5.4)

Equations (5.3) and (5.4) provide an algorithm for fitting a mixture \(F(x)\) of two distributions (5.2) to a given sample distribution \(F_n(x)\). The sample points in this procedure are \(\tau_i = \log_{10}(T_i), \quad i = 1, 2, \ldots, n\) where \(\{T_i\}\) is the original sample.

To define the five parameters of the mixture of the two normal distributions, we find the best fit by varying \(\mu_1, \sigma_1, \gamma_1\) and \(\mu_2, \sigma_2, \gamma_2\). The procedure is as follows.

1. Compute the first three Fisher’s k-statistics [3] of the sample \(m = \frac{1}{n} \sum_{i=1}^{n} \tau_i, s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tau_i - m)^2, \gamma^3 = \frac{1}{n-2} \sum_{i=1}^{n} (\tau_i - m)^3\).
2. Vary \(\mu_1, \sigma_1, \gamma_1\) and for each pair \((\mu_1, \mu_2)\)
   - find \(\alpha(\mu_1, \mu_2)\) using formula (5.3) and substituting for \(\mu\) in it;
   - find \(\sigma_1\) and \(\sigma_2\) using formula (5.4) and substituting the obtained \(\alpha(\mu_1, \mu_2)\) for \(\alpha, s\) for \(\sigma\), and \(\gamma\) for \(\gamma_1\);
   - calculate \(\varepsilon(\mu_1, \mu_2) = \max |F_n(x) - F(x; \mu_1, \mu_2)|\).
3. The best fit could be defined as \((\mu_1^*, \mu_2^*)\) such that

\[
\varepsilon^* = \varepsilon(\mu_1^*, \mu_2^*) = \min_{\mu_1, \mu_2} \varepsilon(\mu_1, \mu_2).
\]

(5.5)

This process is illustrated in Table II, where each line corresponds to a fixed \(\mu_1\) and shows that \(\mu_2\) for the given \(\mu_1\), minimizes \(\varepsilon(\mu_1, \mu_2)\). The bold line indicates the best fit.

The best fit is illustrated in Fig. 6 where a histogram of a single-hour data sample is shown along with the two normal density curves and their mixture. The parameters on the left are the same as in Fig. 4, with the third moment \(\gamma(\tau)\) added; the parameters on the right describe the two normal components shown by the dotted lines. The sample contains 1,650 individual records, 21 of which are shorter than 3 s (in this sample, only about 1%). They are not included in the histogram.

As in Section V-A, the empirical c.d.f. function \(F_n(x)\) and the confidence band are calculated. They are shown in Fig. 7. The dotted line in Fig. 7 is the fitting c.d.f. \(F(x)\). The asymptotic two-sided confidence bounds for the 5% error probability are \(\varepsilon = 1.36 \cdot n^{-0.5} = 0.0337\). The maximal difference between the empirical c.d.f. and the fitting distribution is \(\varepsilon^* = 0.0139\). The closeness between the empirical distribution and the fitting distribution is evident in Fig. 7.

Detailed analysis of 60 single-hour samples was performed in the study reported in [1] (30 days of one month with data for one morning hour and one evening hour). Invariably, all of them showed the same startling agreement between the empirical distribution and the fitting mixture of normal distributions.
TABLE III
EXAMPLES OF MIXTURES OF TWO LOGNORMAL DISTRIBUTIONS

<table>
<thead>
<tr>
<th>Connection Holding Time</th>
<th>Parameters of Mixture</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Type</td>
</tr>
<tr>
<td></td>
<td>Residential 642 sec</td>
</tr>
<tr>
<td></td>
<td>Residential 161 sec</td>
</tr>
<tr>
<td></td>
<td>Facsimile 60 sec</td>
</tr>
</tbody>
</table>

C. Connection Holding Time Distribution in a Heterogeneous Line Group

In a heterogeneous mix of calls generated by residential and business lines, a homogeneous business component is modeled, similar to the residential lines, by the mixture (5.1) of three distributions but the parameters of the distribution are significantly different. Therefore, this heterogeneous mix of calls in telephone exchange should be characterized by as many as six distributions in one mixture.

Other types of calls will introduce other components into the total, even more heterogeneous, mixture of distributions of connection holding times. Finally, the total mixture becomes a component in mixture (4.1). The distributions of other significant connection types, such as facsimile and voice mail, are currently under study.

D. Examples of Connection Time Distribution

Three examples given in Table III provide default values for characterizing the mixture depending on the average holding time. Note that the first two examples are valid for connections in which the coefficient of variation is about 2. This high variability value is typical of residential traffic. The coefficient of variation for business line traffic is under study. The third example is based on the preliminary results of facsimile connection holding time study. In this case, the coefficient of variation is 1.23.

VI. Conclusion

This paper demonstrates the importance of the call holding time distribution for the CCS network design and performance analysis. It shows that the exponential distribution drastically underestimates the proportion of short calls. Models for the call holding time probability distributions are proposed.

The implications of the holding time analysis to CCS network engineering, performance evaluation, and congestion control protocols are currently being examined by others and will be reported elsewhere. In particular, the model described by formula (4.1) has been incorporated in a number of CCS/SS7 network simulation models in Bellcore.

For more comprehensive characterization of the holding time, it is necessary:
- to quantify the time components before completion (Section IV-A) more accurately, and
- to describe the model probability distribution functions of other connection holding times such as facsimile, voice mail, and data.

There is no doubt that the classification and characterization of the holding time distributions provided in this paper will increase awareness in the CCS research and engineering community of the potential effects of the holding time distributions and will provide input for various SS7 performance analyses.

REFERENCES


Vladimir A. Bolotin, for a photograph and biography, see this issue, p. 378.