Delay Analysis of TASI with Random Fluctuations in the Number of Voice Calls

MARTIN J. FISCHER

Abstract—In this paper we consider a system where the talkspurts from several voice conversations are buffered and multiplexed over the same transmission channels. A simple mathematical model is given for the case where the number of voice calls may statistically come and go. This model is then used to numerically analyze the system and make comparisons with previously presented results.

The concept of TASI (time assignment speech interpolation) was first experimented with in the late 1940's and early 1950's by AT&T. In 1958 and 1961 two papers [1], [2] were written concerning the engineering aspects of TASI, as well as documenting results of the first actual implementation of TASI on the transatlantic submarine cable channels. The basic idea behind TASI is that in a normal telephone conversation, there are gaps and pauses between words and syllables, which allows the active parts of two or more conversations to be multiplexed over the same transmission channel and hence reduce the total number of required transmission channels. The active part of a conversation was called a talkspurt, and since there was no way to buffer the contents of the talkspurt, many talkers and channels were needed in order to take full advantage of the TASI concept.

With the introduction of packet-switched communications systems, interest has developed in the TASI implication of packetized speech. Lincoln Lab [3] considered the network implication of packetized speech, i.e., the effects packetized speech has on an overall voice quality, acceptability, and communicability when the voice conversations are transmitted in a packet-switched network. Packetizing the active parts of a voice conversation (the talkspurts) in effect is a switching concept where the philosophy of TASI is implemented in a digital network, and the packets generated by a talkspurt may be buffered when a channel is not available. Weinstein and Hofstetter conducted a study [4] that considered the trade-off between packet delay and the TASI advantage in that environment. They showed that the packetized system offers substantial improvements in the TASI advantage (number of voice calls present and waiting) over voice conversations that are allowed in the system is N with N/C(N > C) being known as the TASI advantage [4]. In steady state, let Q be the random variable representing the number of voice calls present and QT the random variable representing the number of talkspurts that are in the system, i.e., on the channels or in the buffer. If one defines \( P_{t,j} = \Pr(Q = i, QT = j) \) for \( i = 0, 1, \ldots, N \) and \( j = 0, 1, \ldots, i \), then we give a system of equations that \( P_{t,j} \) must obey and a solution for an important special case. These results are then used to determine the average number of talkspurts in the system, the average delay a talkspurt encounters in the buffer, and the complete delay distribution of a talkspurt in the buffer.

The following assumptions are made.

1) Voice calls arrive via a Poisson process with parameter \( \lambda \). The holding time of the call is exponentially distributed with mean \( \mu^{-1} \)

2) An arriving call finding \( N \) calls in the system leaves without receiving service. If accepted, the arriving call immediately generates a talkspurt. Calls may end at any time, i.e., they do not have to be in a talkspurt, but if a call ends and a talkspurt is present, the talkspurt also ends.

3) When there are \( i \) (\( \geq 1 \)) voice calls in the system, each call alternatively generates talkspurts and silent periods. The lengths of each type of period are each exponentially distributed, with mean \( \alpha^{-1} \) for silent periods and \( \beta^{-1} \) for talkspurts.

The talkspurts are served by the channels, and an arriving talkspurt finding all the channels occupied is buffered. We
note that a voice call which is accepted by the system can be in one of two states: a silent period or a talkspurt. Therefore, the number of talkspurts in the system cannot exceed the number of calls present.

Perhaps a discussion as to the implications of assumptions 2) and 3) with respect to the actual behavior of a talkspurt is in order. The last sentence of 2) assumes that if a call ends and there are talkspurts present in the system, one of these talkspurts was generated by this call and that talkspurt is removed from the system. In reality, this may not be the case; a call may end and be in a silent period. The effect of this assumption is that it simplifies the development of the original mathematical equations for the system because one does not have to concern himself with modeling whether or not a call ending his conservation was in a talkspurt. The contribution of this event to the state probabilities $P_{i,j}$ is minimal because the rate at which calls are ending is much slower than the rate at which the talkspurts are ending. For the numbers considered in this paper, the average holding time of a call was 3 min and the average duration of a talkspurt 1.23 s.

The latter part of assumption 3) indicates that a call cannot go into a new talkspurt until his last talkspurt has been transmitted. On the surface, this seems unrealistic, but when one considers that if the system is properly engineered, the queuing delay of a talkspurt has to be short, and so the effect of this assumption is minimal and seems to be valid for cases of interest.

From the first two assumptions, one can see that the behavior of the random variable $Q$, the number of voice calls present in the system, can be modeled via the results of an $M/M/N/N$ loss system (see Cooper [6]). That is, if $p = \lambda/\mu$, then

$$\Pr\{Q = i\} = \frac{\rho^i}{N!} \sum_{r=0}^{N} \frac{\rho^r}{r!}, \quad i = 0, 1, 2, \ldots, N \quad (1)$$

with $\Pr\{Q = N\} = E_{\beta}(\rho, N)$ being Erlang's Loss Formula. Equation (1), with $i = N$, gives the probability that an arriving call is blocked.

In general, the balance equations for $P_{i,j}$ can be written as, for $i \leq C - 1$,

$$\lambda P_{0,0} = \mu P_{1,0} + \mu P_{1,1}$$
$$\lambda + i \mu + i \alpha \beta P_{i,0} = (i + 1) \mu P_{i+1,0} + (i + 1) \mu P_{i+1,1} + \beta P_{i,1}$$
$$\lambda + i \mu + (i - j) \alpha + j \beta P_{i,j} = \lambda P_{i-1,j-1} + (i + 1) \mu P_{i+1,j+1} + (i + 1) \mu P_{i+1,j+1} + \alpha P_{i,j-1}, \quad 1 \leq j \leq i - 1$$
$$\lambda + i \mu + i \beta P_{i,i} = \lambda P_{i-1,i-1} + (i + 1) \mu P_{i+1,i+1} + \alpha P_{i,i-1}. \quad (2)$$

For $C < i \leq N - 1$, we have

$$(\lambda + i \mu + i \alpha \beta) P_{i,0} = \beta P_{i,1} + (i + 1) \mu P_{i+1,0} + (i + 1) \mu P_{i+1,1}$$
$$(\lambda + i \mu + (i - j) \alpha + j \beta) P_{i,j} = \lambda P_{i-1,j-1} + (i + 1) \mu P_{i+1,j+1} + (i + 1) \mu P_{i+1,j+1} + \alpha P_{i,j-1}, \quad 1 \leq j \leq C - 1$$
$$(\lambda + i \mu + (i - j) \alpha + j \beta) P_{i,j} = \lambda P_{i-1,j-1} + (i + 1) \mu P_{i+1,j+1} + \alpha P_{i,j-1}; \quad (3)$$

and finally, for $i = N$, we have

$$(N \mu + N \alpha) P_{N,0} = \beta P_{N,1}$$
$$(N \mu + (N - j) \alpha + j \beta) P_{N,j} = \lambda P_{N-1,j-1} + (N + 1 - j) \alpha P_{N,j-1} + (j + 1) \beta P_{N,j+1}, \quad 1 \leq j \leq C - 1$$
$$(N \mu + (N - j) \alpha + j \beta) P_{N,j} = \lambda P_{N-1,j-1} + (N + 1 - j) \alpha P_{N,j-1} + \beta P_{N,j+1}, \quad C \leq j \leq N - 1$$
$$(N \mu + \beta) P_{N,N} = \lambda P_{N-1,N-1} + \alpha P_{N,N-1}. \quad (4)$$

Before proceeding, let us consider two special cases; the first occurs when $N = C = 1$. From (2) we have

$$p P_{0,0} = P_{1,0} + P_{1,1}$$

$$= \Pr\{Q = 1\}$$

$$= \frac{\rho}{1 + \rho}$$

or

$$P_{0,0} = \frac{1}{1 + \rho} \quad (5)$$

which is to be expected since $P_{0,0} = \Pr\{Q = 0\}$. Now from (3) and from the fact that $\Pr\{Q = 1\} = \rho/(1 + \rho)$, one gets

$$P_{1,0} = \frac{\rho \beta}{(1 + \rho) (\mu + \alpha + \beta)} \quad (6)$$
and

\[ P_{1,1} = \frac{\rho (\mu + \alpha)}{(1 + \rho)(\mu + \alpha + \beta)}. \]  

(7)

The second special case we consider occurs when \( \alpha \) and \( \beta \) get large while \( \eta = \alpha/\beta \) remains constant. Equations (2), (3), and (4) become, for \( i \leq C - 1 \),

\[ \rho P_{0,0} = P_{1,0} + P_{1,1} \]

\[ \ln P_{i,0} = P_{1,1} \]

\[ ((i - j)(\eta + j)) P_{i,j} \]

\[ = (i + 1 - j)\eta P_{i,j-1} + (j + 1)P_{i,j+1}, \]

\[ 1 \leq j \leq i - 1 \]

\[ \eta P_{i,i} = \eta P_{i,i-1}. \]

(8)

Again, for \( C \leq i \leq N - 1 \), we have

\[ \ln P_{i,0} = P_{i,1} \]

\[ ((i - j)(\eta + j)) P_{i,j} \]

\[ = (i + 1 - j)\eta P_{i,j-1} + (j + 1)P_{i,j+1}, \]

\[ 1 \leq j \leq C - 1 \]

\[ ((i - j)(\eta + C)) P_{i,j} \]

\[ = (i + 1 - j)\eta P_{i,j-1} + (j + 1)P_{i,j+1}, \]

\[ C \leq j \leq i - 1 \]

\[ CP_{i,i} = \eta P_{i,i-1}. \]

(9)

and finally,

\[ N\eta P_{N,0} = P_{N,1} \]

\[ ((N - j)(\eta + j)) P_{N,j} \]

\[ = (N + 1 - j)\eta P_{N,j-1} + (j + 1)P_{N,j-1}, \]

\[ 1 \leq j \leq C - 1 \]

\[ ((N - j)(\eta + C)) P_{N,j} \]

\[ = (N + 1 - j)\eta P_{N,j-1} + (j + 1)P_{N,j-1}, \]

\[ CP_{N,N} = \eta P_{N,N-1}. \]

(10)

We note that these equations are decoupled in \( i \), and the solution is easily seen to be, for \( i \leq C \),

\[ P_{i,i} = \binom{i}{j} \eta^i P_{i,0} \]

(11)

and for \( i \geq C \),

\[ P_{i,j} = \begin{cases} \binom{i}{j} \eta^i P_{i,0}, & j \leq C \\ \frac{i!}{(i-j)!C!C^{j-C}} \eta^i P_{i,0}, & j \geq C \end{cases} \]

(12)

where \( P_{i,0} \) is found using

\[ \Pr\{Q = i\} = \sum_{j=0}^{i} P_{i,j} \]

(13)

where \( \Pr\{Q = i\} \) is given by (1).

From (11)-(13), one sees that for this special case, \( P_{i,j} \) can be expressed as the product of the probability the number of voice calls in the system equals \( i (\Pr\{Q = i\}) \) and the state probability of a finite source, \( C \)-server queueing system [6, pp. 90-93].

Let us now consider the extent of the difference between the results one gets when \( \alpha \) and \( \beta \) are large and the actual results one would get without taking this limit. For this comparison, we use the first special case and consider the expected number of talkspurts in the system \( E\{Q_T\} \); from (7) we have

\[ E\{Q_T\} = \frac{\rho}{1 + \rho} \frac{\mu + \alpha}{\mu + \alpha + \beta}. \]

(14)

But from specializing the results for the second special case, we get

\[ E\{Q_T\} = \frac{\rho}{1 + \rho} \frac{\eta}{1 + \eta}. \]

(15)

From [4], the nominal values of \( \alpha^{-1} \) and \( \beta^{-1} \) are 1.34 and 1.23 s; if one assumes \( \mu^{-1} = 180 \) s, then \( E\{Q_T\} = 0.4805 \Pr\{Q = 1\} \) and 0.4786 \( \Pr\{Q = 1\} \) from (14) and (15), respectively. The difference between these two quantities is approximately 0.0019 \( \Pr\{Q = 1\} \) and so is very minimal. Thus, for the values of \( \alpha \), \( \beta \), and \( \mu \) given above, one would expect the actual solution to (2)-(4) to be very close to the solution given by (11)-(13), and we use those results for the remainder of the paper.

Some expected value results are now easily written. The expected number of talkspurts in the system \( E\{Q_T\} \) is

\[ E\{Q_T\} = \sum_{i=0}^{N} \sum_{j=0}^{i} jP_{i,j}; \]

(16)

the expected number of talkspurts in queue \( E\{Q_T^Q\} \) is

\[ E\{Q_T^Q\} = \sum_{i=C+1}^{N} \sum_{j=C}^{i} (j-C)P_{i,j}; \]

(17)
and the average waiting time of a talkspurt in the buffer

\[ \mathbb{E}[W_T^q] = \mathbb{E}[Q_T^q]/\bar{\alpha} \]  

(18)

where

\[ \bar{\alpha} = \alpha \sum_{i=0}^{N} \sum_{j=0}^{i} (i-j)P_{i,j} \]  

(19)

and \( Q_T^q \) and \( W_T^q \) are the random variables representing the number of talkspurts in the buffer and the waiting time of a talkspurt in the buffer.

We now develop an equation for the probability that the waiting time in the buffer is less than or equal to \( t \), \( \Pr[W_T^q \leq t] \). In order to do so, we need an expression for the probability \( Q = i \), and an arriving talkspurt that is accepted into the system finds \( j \) talkspurts already there; let us denote this probability by \( R_{i,j} \). From [7], one can show that

\[ R_{i,j} = \frac{(i-j)P_{i,j}}{\sum_{r=0}^{i} \sum_{i=0}^{N} (i-r)P_{i,r}} \]  

(20)

and so

\[ \Pr[W_T^q] = 1 - \sum_{i=C+1}^{N} \sum_{j=C}^{i} \sum_{r=0}^{i-C} e^{-\beta \alpha t} \frac{(\beta \alpha t)^r}{r!} R_{i,j} \]  

(21)

The probability of zero wait in the buffer is

\[ \Pr[W_T^q = 0] = 1 - \sum_{i=C+1}^{N} \sum_{j=C}^{i} R_{i,j} \]  

(22)

The probability distribution of the total waiting time in the system for a talkspurt is found by convoluting \( \Pr[W_T^q \leq t] \) with the exponential service time of a talkspurt.

One can see from (20) that \( R_{i,j} \neq P_{i,j} \), but it is of interest to discuss what the physical difference is between these two probabilities. For both, the number of calls present is the same, but for \( R_{i,j} \), an arriving talkspurt finds \( j \) already present, whereas \( P_{i,j} \) represents the case where someone outside the system randomly observes the system and finds \( i \) voice calls and \( j \) talkspurts present. Consider the case where \( i = 1 \), the call alternatives between a talkspurt and a silent period. Thus, the probability an arriving talkspurt finds one talkspurt in the system equals 0 for this case, whereas \( P_{1,1} = \eta \Pr(Q = 1)/(1 + \eta) \neq 0 \) for \( \eta \neq 0 \). For a more in-depth discussion of the difference between \( R_{i,j} \) and \( P_{i,j} \), see Cooper [6, ch. 3].

### III. SYSTEM ANALYSIS

Using the mathematical results of Section II, we now numerically investigate the performance of the system when one considers the statistical fluctuations in the number of voice calls. This investigation is carried out in two areas: first, we ascertain the differences in results that are obtained when one considers the cumulative distribution function of the waiting time as opposed to only the mean. Second, an investigation is carried out to determine the effect on performance when one fixes the number of voice calls as compared with the case where the calls may statistically come and go.

As we pointed out in Section II, we have used the solution, given by (11)-(13), in our analysis. In order to determine how much error was introduced by this solution, we have made some comparison with results of a simulation model of that actual system as modeled via (2)-(4). These results are given in Table I. For various values of \( N \) \((C = 10)\), the results one would get for the expected number of talkspurts in a system via (11)-(13) and a simulation are given for two levels of blocking: \( PB = 0.2 \) and \( PB = 0.6 \). As one can see, there is close agreement between the two results, and one can only conclude that the solution given by (11)-(13) is very close to the actual solution of (2)-(4).

Fig. 1 presents two results in terms of variability of the presence of voice calls. Three curves are presented for different values of voice blocking probability; for each curve the average delay in the buffer is given as a function of the number of voice calls which is present. The expected number of voice calls may vary, but the expected number of calls which is present is the same as in the fixed case.

Table I. Comparisons of Mathematical Solution and Simulation for the Expected Number of Talkspurts in System \((C = 10, \alpha^{-1} = 1.34 \text{ s}, \beta^{-1} = 1.23 \text{ s})\)

<table>
<thead>
<tr>
<th>N</th>
<th>Equations (11)-(13)</th>
<th>Simulation</th>
<th>Equations (11)-(13)</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>4.71</td>
<td>5.11</td>
<td>5.44</td>
<td>5.32</td>
</tr>
<tr>
<td>13</td>
<td>5.18</td>
<td>5.22</td>
<td>5.93</td>
<td>5.99</td>
</tr>
<tr>
<td>14</td>
<td>5.63</td>
<td>5.72</td>
<td>6.42</td>
<td>6.47</td>
</tr>
<tr>
<td>15</td>
<td>6.11</td>
<td>6.16</td>
<td>6.90</td>
<td>6.95</td>
</tr>
<tr>
<td>16</td>
<td>6.61</td>
<td>6.62</td>
<td>7.42</td>
<td>7.44</td>
</tr>
<tr>
<td>17</td>
<td>7.06</td>
<td>7.13</td>
<td>7.94</td>
<td>7.95</td>
</tr>
</tbody>
</table>
knowledge has not been presented before in the literature for this type of system. Most engineering studies consider only the average delay or the variance of the delay; we give results for the probability distribution function of delay for various values of voice blocking. From the curves of Fig. 2, one can determine the time \( t_{0.95} \) such that \( \Pr\{W_T^Q < t_{0.95}\} = 0.95 \). The lines at the top of the curves compare \( t_{0.95} \) with the expected waiting time in the buffer for each of the three voice blocking probabilities. From this figure, \( t_{0.95} \) is about seven or eight times the expected delay in the buffer.

These results are further examined in Figs. 3 and 4. There we consider the sensitivity of \( \Pr\{W_T^Q < t\} \) to small, \( C = 5 \), and large, \( C = 30 \), numbers of voice channels. As one can see, the sensitivity is much greater for the small number of voice channels than for the large number of channels. Several points can also be made from these figures; first, as the TASI advantage increases, the probability distribution \( \Pr\{W_T^Q < t\} \) decreases for a fixed value of \( t \). Second, for the cases where it applies, the 95th percentile point is again about seven or eight times the expected delay in the buffer. Finally, these figures give us further insight into the findings of [4] which stated that the TASI advantage could be improved on even for a small number of voice channels, this percentage is rather low and might not be acceptable in practice, whereas for a large number of voice channels, this percentage is much higher and probably acceptable in practice. The point to be made is that even though improvements in the TASI advantage can be obtained for small numbers of voice channels, when one considers expected delay in the buffer, the percentage of packets that experience a delay larger than the average might be unacceptable.

IV. CONCLUSIONS

We have presented a mathematical model that predicts the performance of a talkspurt in a packetized voice communications system where voice calls are randomly allowed to come and go. Numerical applications of this model involved comparisons with previously generated results as well as new results for the complete probability distribution of delay. These results can be summarized as follows.

1) Significant differences in the average delay in the buffer can be obtained, depending on the block probability of voice calls (Fig. 1).

2) Fixing the number of voice calls in the system results in shorter average buffer delays than in a system where the
Fig. 3. Sensitivity of waiting time distribution for small number of voice channels ($C = 5$) ($a^{-1} = 1.34 \text{ s}, \beta^{-1} = 1.23 \text{ s}, PB = 0.2$).

Fig. 4. Sensitivity of waiting time distribution for large number of voice channels ($C = 30$) ($a^{-1} = 1.34 \text{ s}, \beta^{-1} = 1.23 \text{ s}, PB = 0.2$).

calls randomly come and go, but the expected number of calls present is the same in the fixed system (Fig. 1).

3) New results for the probability distribution of delay in the buffer are easily computed and tabulated so that comparisons of expected delay and percentile points of this distribution can be made. For the examples considered here, the 95th percentile point was about seven times the expected delay (Fig. 2).

4) Previous analyses [4] and [5] for fixed number of calls have shown that substantial increases in the TASI advantage can be gained when the contents of talkspurts are buffered, even for a small number of voice channels. Our analysis has shown that one gets shorter average delays and smaller overflow probabilities when the calls are allowed to randomly arrive and leave, thus further strengthening these results. However, our new results indicate that the 95th percentile point of the waiting distribution is seven times the mean. Furthermore, the percentage of packets that see a greater than the average delay for the smaller number of voice channels might not be acceptable in practice (Figs. 3 and 4), thus indicating that further analysis and study are required on this aspect of the problem.

ACKNOWLEDGMENT

The author would like to thank W. Cohen of the Defense Communications Engineering Center for his helpful discussions on the engineering relevance of this problem. Furthermore, I am grateful to Dr. C. Weinstein and Dr. R. Berger of M.I.T. Lincoln Laboratories for allowing me access to early versions of their works.
Balanced HDLC Procedures: A Performance Analysis

WERNER BUX, KARL KÜMMERLE, AND HONG LINH TRUONG

Abstract—The prime objective of the present paper is to analyze the performance of HDLC Balanced Class of Procedures, i.e., to quantitatively study the interaction among a multiplicity of parameters which are procedure specific, characterize the properties of the transmission medium, and identify the operational characteristics and requirements. The approach taken is to consider two kinds of operation: a saturated case characterized by maximum throughput as the most suitable measure of performance, and a nonsaturated situation for which waiting and transfer times are the appropriate measures. The analysis is performed by means of both simulation and analytic techniques. Within the simulation model, the information transfer phase of the HDLC procedure was implemented in full detail. The key idea of the analytic approach is to use a so-called virtual transmission time, a quantity comprising both the real constant transmission time of a message and the duration of recovery actions in case of transmission errors. It allows the performance measures to be represented by explicit and easily-computable expressions. The results provide a fundamental insight into how the most relevant parameters interact and determine performance.

I. INTRODUCTION

The data-link control level in data and computer communication networks contains control functions such as addressing, frame numbering, error recovery, and flow control to reliably transfer data and control information across a physical communication link. Both national and international standardization bodies and computer manufacturers have defined data-link control procedures, e.g., ISO High-Level Data-Link Control Procedure (HDLC) [1]-[3], ANSI Advanced Data Communication Control Procedures (ADCCP) [4], Digital Equipment Digital Data Communications Message Protocol (DDCMP) [5], IBM Synchronous Data-Link Control Procedure (SDLC) [6].

The objective of the present investigation is to study the performance of HDLC when it is operated in the asynchronous balanced mode (ABM) over a point-to-point configuration. We are interested in: 1) identifying the essential parameters which determine protocol performance, and 2) analyzing their impact on performance under various conditions. In three recent papers [7]-[9] we provided detailed analyses of the balanced class of procedures through simulation techniques and analytic modeling, respectively; the present contribution combines the major results obtained there and represents a comprehensive performance analysis of the HDLC balanced class of procedures.

The protocol was implemented in full detail within the framework of a simulation study [7], the results of which provided a basic insight into the interaction among procedure-specific parameters, parameters characterizing the transmission medium, and the required operational parameters. The analytic model—structured such that it reflects all relevant parameters and mechanisms—uses the new concept of a virtual transmission time, a quantity comprising both the real transmission time of an information frame and the duration of recovery actions in case of transmission errors.

Related performance investigations of full-duplex data links are given in [10] for the throughput behavior of SDLC and the Unbalanced Classes of HDLC, and in [11], [12] for

REFERENCES

Martin J. Fischer, for a photograph and biography, see p. 26 of the January 1980 issue of this TRANSACTIONS.