Containing the Effects of Long Holding Times due to Internet Dial-up Connections

V. Ramaswami, David Poole, Soohan Ahn, Simon Byers *, & Alan Kaplan †
* AT&T Labs - Research, 180 Park Avenue, Florham Park, NJ 07932.
E-mail: vram@research.att.com
† Retired from AT&T; now an independent consultant.

***Abstract*— Internet dial-up calls have much longer holding times than voice calls. We show that in their presence, classical engineering procedures become inadequate and, without control, economies of scale get impaired. Uncontrolled systems will suffer periods of unavailability due to persistence of congestion. For digital loop carrier systems with concentration, no-dial-tone situations impacting even critical services can obtain. Call admission and overload controls that maintain good performance and high utilization are also presented.**

I. Introduction

This work was done in the context of no-dial-tone complaints from a large number of customers in a residential telephony system in which access was through a subscriber loop interface (SLIC) system and a Head End Terminal (HET). Dial tone and digit reception were provided by the switch whence inaccessibility of the switch due to circuit congestion could manifest as dial tone unavailability. Data showed that 5-8% of the calls were dial-up calls to ISPs, but they accounted for 30-45% of the busy hour usage.

We show that the uncontrolled presence of dial up ISP calls renders standard engineering and dimensioning procedures ineffective and cause persistence of congestion for noticeable durations. In digital loop carrier systems with concentration (such as those based on GR-303 [1], e.g.), that situation can lead to no-dial-tone conditions impacting even critical services like 911 calling. Inability to receive digits results in inability to determine even the type of the attempt making control quite difficult. Surmounting the problem by provisioning additional circuits or combining circuit groups will be shown to be risky. The provision of Internet offload switches may not alleviate this class of problems since these problems occur at the access portion of the network, and offload switches mainly help to relieve congestion of circuits within the core network. These counter-intuitive results involve a subtle argument similar to that involved in the famous waiting time paradox [2].

II. Circuit holding time

Fig. 1 shows a histogram of the holding times (transformed to logarithm to the base 10) of a set of about 4.5 million residential calls gathered over a week in a certain serving area. The figure shows that the holding time distribution is rather long-tailed. Although the observed data had a median of only 48 seconds, the mean was 297 seconds. These characteristics were also found to be fairly robust across other time periods and geographic locations.

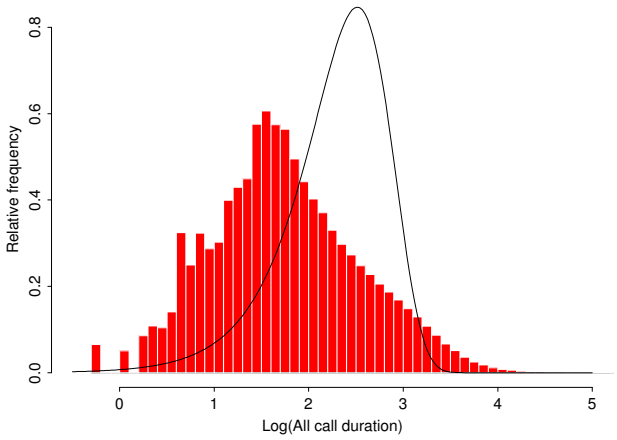

Fig. 1. Histogram of holding time (seconds) distribution for all calls on log scale (base 10). The density for the exponential distribution with the observed mean, transformed to the log scale, is shown as the curve.

In Fig. 1, we have also provided a comparison of the empirical histogram to the exponential distribution with the same mean. For convenience of visualization, both are shown after transforming to the logarithm (base 10) of the holding time. Also, some selected percentiles of the observed data and the exponential model are shown in the first three columns of Table 1. Clearly, the exponential distribution is *not* a good fit to the data. Thus, a question worth examining is whether this has an impact on trunk engineering usually done using the $M/M/c/c$ model based on the insensitivity of the blocking formula [4], [3] to the holding time distribution. We do that using phase type models.

III. A phase type fit

We fitted a phase type distribution [5],[6] to the data using the EM-algorithm [7]. Fig. 2 shows the data and our phase type fit with 4 phases, while Table 1 lists some selected percentiles. It may be noted that the phase type model provides an excellent fit to the data up to the 99-th percentile.

We extracted dial-up ISP calls from the data. and observed that they were much longer than other calls, having a mean of 1956 seconds (compared to an overall mean of 297s) and a median of 673 seconds (compared to 48s for the combined

0-7803-7978-0/03/$17.00

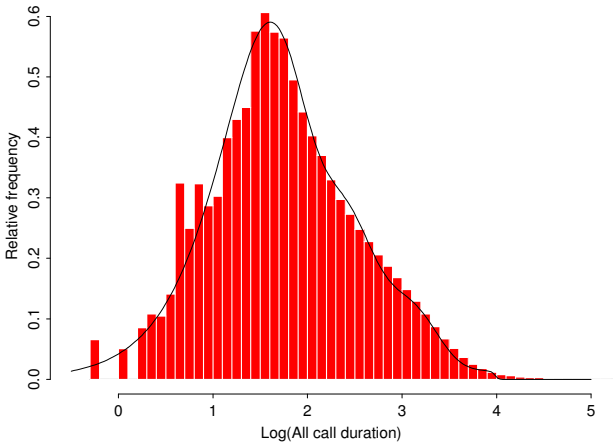

Fig. 2. Histogram of holding time distribution for all completed calls on log scale (base 10). The density for the 4-component phase type distribution fit to these data is shown as the curve.

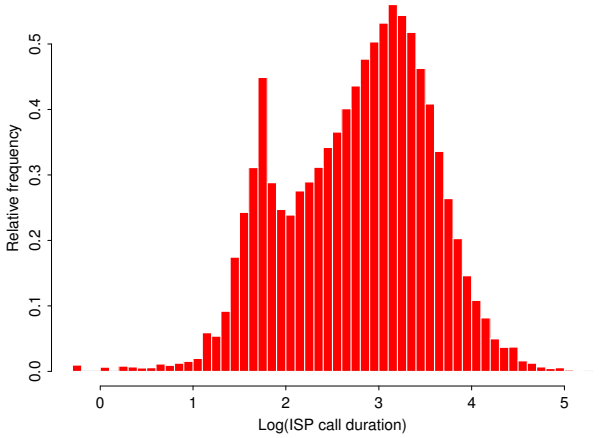

Fig. 3. Histogram of the holding time distribution for known dial-up modem calls, transformed to log scale (base 10). Note the bimodality caused by two primary groups: users who dial up to download e-mail only, and users who dial up to browse on the Internet.

data.) Fig. 3 shows a histogram of a set of several million ISP calls; compare this to Fig. 1 and observe the striking differences between them. An important issue then is to examine the impact of these facts on blocking performance.

IV. Analysis of Residuals

The steady state remaining service times of customers in service in an $M/G/c/c$ queue are independent and identically distributed with density

$$h(x) = [1 - F(x)]/\mu, \quad x > 0,$$

where $F(.)$ is the holding time (cumulative) distribution function and μ is its mean. For a phase type distribution, the above, called the excess life or residual distribution, is also a phase type distribution and is easy to compute (see [5], Chapter 3).

In our case, the computed median and mean for the residual holding time distribution were respectively 703 and 2367 seconds. Fig. 4 shows a comparison of the data on holding times, the phase type distribution fitted to it and the residual distribution obtained from the fitted phase type distribution; see also Table 1. It is clear that residual holding times are stochastically much larger than total holding times. Fig. 5 provides a comparison of the computed data from an empirical histogram obtained from a (different) random sample of 11264 calls; see how well the phase type residual distribution matches the observed.

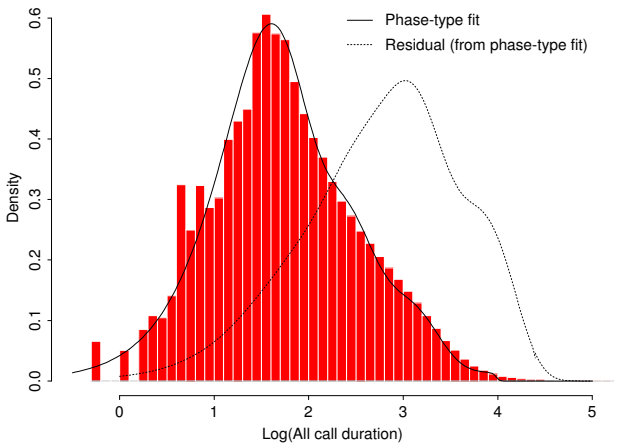

Fig. 4. Histogram of holding time distribution for all calls on log scale (base 10). The densities for the 4-component phase type fit and the residual distribution corresponding to this fit are shown as curves.

TABLE I

Table of selected percentiles for (from left to right) the observed holding time data, the exponential distribution with observed mean, the phase type fit to the observed data, and the residual distribution corresponding to the phase type fit.

%'ile	Data	Exp.	Phase	Resid.
10	5.4	31.3	5.8	40.8
20	12.0	66.3	12.7	116.1
30	21.0	105.9	21.2	238.0
40	32.4	151.7	32.2	423.7
50	48.0	205.9	47.7	703.0
60	72.6	272.1	72.2	1124.6
70	120.6	357.6	120.2	1809.7
80	232.8	478.0	235.9	3285.3
90	601.8	683.9	597.6	7028.2
95	1237.8	889.7	1268.0	11073.8
99	3952.2	1367.7	4035.2	20489.2
99.5	6074.4	1573.6	7168.4	24470.7

The dramatic difference between residual and total holding times is highly significant. For instance, consider the fact that the median residual service time is 703 seconds. Suppose, based on the Erlang-B engineering formulas, one has provisioned a set of 200 circuits, and that, say, 180 of these become busy. About half the busy circuits can then be expected to

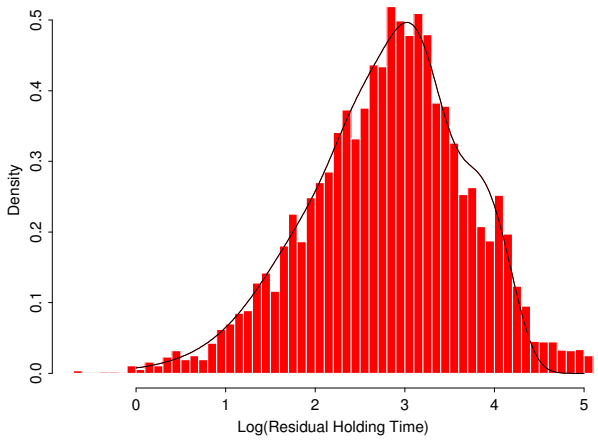

Fig. 5. Histogram of the residual holding time distribution on the log scale (base 10) for a random sample of new calls. The density for the 4-component phase type fitted residual distribution is shown as the curve.

remain continuously busy for the next 703 seconds, a period during which the system is essentially operating with at most 110 circuits. Thus, for noticeable periods, congestion and blocking for circuits could be significantly higher than the engineered level.

The explanation of the above lies in a length biasing argument similar to that in the waiting time paradox: one is likely to find a larger fraction of long holding time calls among those currently in the system than what is predicated by the overall fraction of completed calls of that type since longer calls tend to get stuck in the system and are more likely to be seen. Our simulation results (not shown here) also confirmed this phenomenon.

V. Need for Controls

One may propose the provision of larger trunk groups between the HET and the switch as a possible solution to the congestion problems noted here. Though reducing the chance of congestion at engineered loads, this, however, would not relieve the performance degradation and persistence of congestion when congestion does indeed occur. Furthermore, when one considers larger circuit groups with loads to match the given blocking rate, there are also some subtleties to consider. To this end, consider two circuit groups with 24 and 120 circuits engineered to a long run blocking rate of 0.01. Elementary calculations of the Erlang-B type yield for these systems the values 22.57 and 120.19 for the $\mu + 2\sigma$ values of the steady state number of busy circuits. Note that at these values while the smaller group still has a spare, the larger group is exhausted. Thus, if one wants to gain the maximum from economies of scale by provisioning larger trunk groups, one needs to allow for higher occupancy levels, which in our situation is often a harbinger of trouble. These perspectives were suggested to us by the work in [8]. The prudent approach therefore is to manage congestion by augmenting dimensioning procedures with sound admission and overload controls.

VI. Controls

We have observed that when congestion occurs one is highly likely to find many circuits occupied by long holding time ISP dial-up connections. This points to a solution based on the following principles: (a) under congested conditions, do not admit new ISP dial-ups; (b) in extreme congestion situations, terminate an ongoing ISP connection to make room for a voice call. With regard to the former, note that one may select a small threshold T and reject an ISP attempt when the number of free circuits F is less than T immediately following digit analysis when a decision has to be made to route the call. This allows the system to maintain, with a high probability, a circuit just for the purpose of giving dial tone and receiving digits. The threshold T can be quite small since dial tone and digit reception take a very short amount of time. However, if a call attempt finds fewer than T free circuits at the end of digit analysis and happens to be a voice call, then we may accept it provided the number I of ongoing ISP calls in the system is greater than a pre-assigned threshold K; in that case, one of the ongoing ISP connections is terminated to prevent a possible no-dial-tone condition in the near future. Note that this effectively attempts to maintain, with a high probability, T circuits for providing dial tone and digit reception. However, this is accomplished without reserving some specific set of T circuits and thereby avoids the hassle of having to switch a call to a different circuit after digit analysis has been made and also obviates concerns about failures within the set of reserved circuits. The scheme presented in Fig. 6 embodies these ideas and is a simplified version of the controls developed by Ramaswami & Kaplan; the actual control has many other hooks, such as for instance to control the effects of rapid re-attempts by modems, etc. In practice, the choice of K has to be made judiciously such that (a) under engineered loads and under moderate departures therefrom, the chance of terminating an ISP call during its lifetime is small, and (b) the chance of repeated hits on the same caller is made negligible. The latter could be achieved by controlling the overall premature termination probability for ISP calls and by selecting the call to be terminated with care, such as, for example, by selecting the call that has been on for the longest amount of time or has exceeded a given threshold of time.

VII. Performance Results

The performance results reported here are obtained using appropriate Markov chain models. A sample set of results is presented here for the case of 96 circuits to illustrate how the proposed controls work. The mean duration of the call setup phase is assumed to be 3 seconds and the means for voice and ISP calls are chosen to match the observed values of 190 and 1956 seconds respectively. The small set of results presented here are for illustration only; extensive computations over a range of values of T and K are needed in practice to fine tune the system.

Table 2 provides a comparison of the situation without any control to the case where we use the controls $T = 1$, $K = 26$.

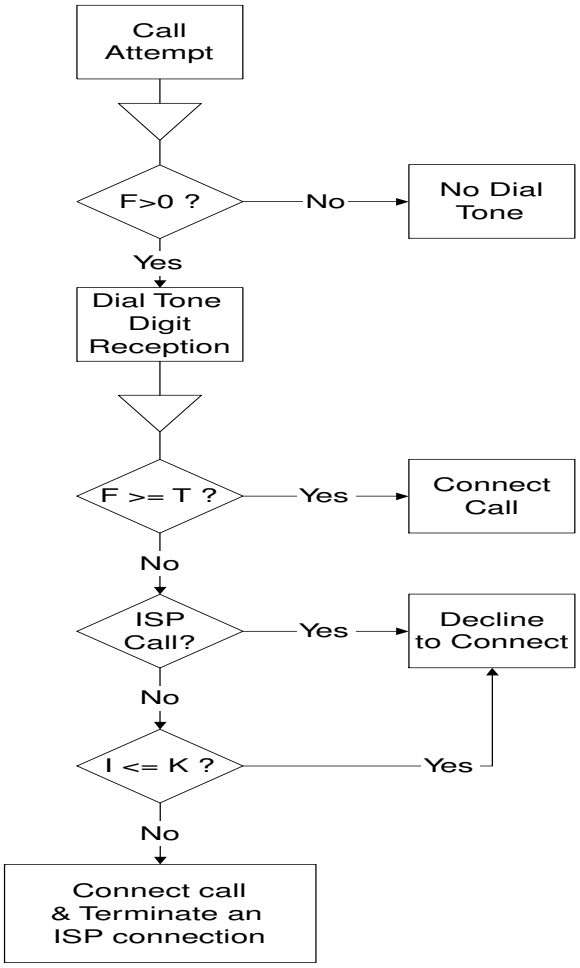

Fig. 6. Flow chart of the control algorithm.

The control helps to drive the "no-dial-tone" probability to near zero and provides much better blocking performance for voice at a small expense of ISP calls. Note that although under our controls there is blocking even after getting dial tone, since ISP calls form only a small fraction of the totality of calls, the actual number of attempts blocked would be significantly reduced.

An important performance measure is the chance that an ISP call is pre-empted during its lifetime. For the various values of ρ considered, these probabilities are 0.058, 0.165, 0.255 and 0.303 respectively. We see that this probability increases noticeably (only) under significant overloads. Given that pre-emption occurs only when the number of ISP calls is more than 26 and that we may select the pre-empted call judiciously (e.g., as the oldest) we can make sure that the same caller is not hit repeatedly.

Finally, for values of ρ in the range 0.8 to 1.1, Table 3 provides the conditional performance of the system given that the number of circuits busy is at least $\min(96\rho,\ 96)$. We note that without control, the system performance degrades drastically as the load increases. The control helps to make that degradation graceful and particularly so for voice calls, while maintaining negligible no-dial-tone probabilities as one would desire to have. The fraction of ISP calls rejected is not significantly higher than voice calls, and yet that buys much by way of performance yielding a graceful degradation of service.

A detailed version of this paper will appear elsewhere.

ACKNOWLEDGMENTS

We thank Aswath Rao, Gagan Chaudhury and Ward Whitt for valuable comments, and Glen Zdroik for assistance with data collection. Soohan Ahn was supported partially by the Post-Doctoral Fellowship Program of the Korea Science & Engineering Foundation.

REFERENCES

[1] GR-303-CORE, Issue 4, "IDLC Generic Requirements, Objectives & Interfaces," Dec. 2000, Telcordia.
[2] W.Feller, *An Introduction to Probability Theory and its Applications,* 2nd ed., vol. 2, New York: John Wiley & Sons, 1971.
[3] B.A. Sevast'yanov, "An ergodic theorem for Markov processes and its application to telephone systems with refusals," *Teor. Verojatnost. I Primenen,* vol. 2, pp. 106-116 (Russian with English Summary), 1957.
[4] R.W. Wolff, *Stochastic Modeling and the Theory of Queues,* London: Prentice-Hall, 1988.
[5] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling,* Philadelphia: SIAM & ASA, 1999.
[6] M.F. Neuts, *Matrix-Geometric Solutions in Stochastic Models,* Baltimore: Johns Hopkins Univ.Press, 1981.
[7] S. Asmussen, O. Nerman and M. Olsson, "Fitting phase-type distributions via the EM algorithm," *Scand. J. Stat.,* vol. 23, pp. 419-441, 1996.
[8] V. Ramaswami and M.F. Neuts, "Some explicit formulas and computational methods for infinite server queues with phase type arrivals," *J. Appl. Probab.,* vol. 17, pp. 498-514, 1980.

TABLE II
TABLE OF NO DIAL TONE (NO-DT) AND CALL REJECTION PROBABILITIES. THE LOAD FACTORS CORRESPOND TO ρ VALUES 0.8, 0.9, 1.0 AND 1.1.

	No Ctrl	T=1 and K=26		
Erlang	No DT	No DT	Reject Voice	Reject ISP
76.8	.0045	.00004	.0005	.0043
86.4	.0282	.00019	.0049	.0155
96.0	.0772	.00057	.0210	.0374
105.6	.1374	.00136	.0531	.0727

TABLE III
CONDITIONAL PERFORMANCE MEASURES FOR THE SYSTEM GIVEN THAT THE NUMBER OF CIRCUITS BUSY IS AT LEAST $\min(96\rho,\ 96)$.

	No Ctrl	T=1 and K=26		
Erlang	No DT	No DT	Reject Voice	Reject ISP
76.8	.0082	.0001	.0011	.0090
86.4	.0559	.0008	.0171	.0539
96.0	.0996	.0013	.0432	.0770
105.6	.1513	.0022	.0789	.1080