Ensuring Access to Emergency Services in the Presence of Long Internet Dial-Up Calls

V. Ramaswami [a], David Poole [a], Soohan Ahn [b], Simon Byers [a], & Alan Kaplan [a]

[a] AT&T Labs, 180 Park Avenue, Florham Park, NJ 07932

vramaswami, dpoole, sbyers, aekaplan@att.com

[b] Department of Statistics, The University of Seoul

Seoul 130-743, S. Korea

sahn@uos.ac.kr

Abstract

Telephone availability is critical, particularly in emergency situations when people need immediate help. We used statistical data analysis and queueing models to identify the root cause of dial tone unavailability in parts of the AT&T network and to develop remedies. Our solutions restored quality service, protecting the AT&T brand name and ensuring the safety of our customers. This work also gave AT&T opportunities to reduce transit charges paid to other carriers by $15 million per year. In addition, we have filed five patent requests, of which two have been granted and the rest are pending (Chaudhury et al 2004, Kaplan & Ramaswami 2004). Furthermore, our findings have important implications for several current areas of research related to Internet and broadband technologies, call-center engineering, and network security.

In 2001, AT&T received complaints from customers in one of its market segments that they were not getting dial tone during certain times of the day. The problem was very serious. People need unfailing telephone services, particularly in emergencies. For instance, if you need to call 911 to report a heart attack or a major catastrophe, you do not want to find the telephone lines dead. Based on national statistics on 911 calling, the number of calls related to life threatening emergencies alone would be about 90 per day for the cohort of population affected.

Preliminary suspicion centered around maintenance activities related to the addition of new customer lines. But our data analysis showed no temporal or geographic correlations with maintenance activities. While most maintenance was conducted during day-time hours, the dial tone problems were occurring mainly during the evening hours when residential traffic peaks. That made us focus our attention on congestion as a potential cause.

In the affected parts of the network, AT&T was providing access using digital loop carrier systems based on a commonly adopted standard, called GR-303 (GR-303-CORE, Issue 4, 2000). Access circuit congestion could render accessing the telephone switch impossible or delayed. Since the switch is the one that provides dial tone, dial tone unavailability or dial tone delays would be the natural consequences of such congestion. But blocking measurements (at the access portion of the network) were not indicative of that. Furthermore, network engineers had followed time-tested methods based on queueing and teletraffic theories and had no a priori reason to expect congestion.

The technical ideas underlying our approach were inspired by V. Ramaswami's Ph.D. dissertation (Ramaswami & Neuts, 1980). That work obtained some counter-intuitive results for queues with disparate

traffic types, and these were relevant because in the scenario examined, customers used circuits not only for voice calls, but also for internet dial-up calls; the latter last much longer than voice calls.

We collected data on completed calls, fitted distributions to their durations (holding times), and made detailed calculations and simulations based on state-of-the art algorithmic methods of queueing theory. With these, we confirmed that there was significant chance of congestion due to the long internet calls. We could also explain the observed anomaly of low blocking rates at the access portions by identifying that blocking actually happens up-stream in the network relative to measurement points.

Once we identified the root cause, AT&T could remedy the problems quickly by rearranging circuits and balancing loads to effect a more favorable mix of business and residential lines on access groups. Unfortunately, these short-term solutions are not efficient as they entail significant labor and sometimes even customer downtime. That necessitated the development of efficient solutions for the long term in the form of automatic controls. We developed a suite of such solutions resulting in five patent applications, of which two have been granted already.

Our work resulted in a side benefit to AT&T by way of opportunities to reduce, to a tune of $15 million per year, the transit charges it pays to other carriers. We also found that our findings have major implications to several other areas of our business.

This paper is a presentation of our work in a manner accessible to a general audience. The more technically oriented reader may refer to a companion paper (Ramaswami, V. et al 2003) for additional details.

Call Holding Times

The data for 4.5 million residential calls over a week in one serving area is shown in Figure 1. The distribution has a long tail, and a mean of 297 seconds that significantly exceeds three minutes, the highly quoted value for average voice call duration. While we could attribute these to the use of circuits by some customers for internet dial-up, we also found that only a small fraction (6%) of the calls were of the internet type, and it remained to examine if that small fraction could indeed cause congestion of the type suspected.

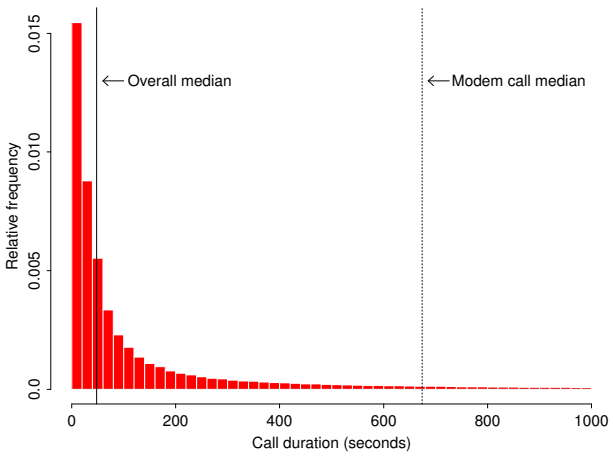

Figure 1: The histogram of holding time (call duration) distribution for calls has a long tail. The median for modem calls is seen to be much larger than the median for all calls.

Many studies on voice call holding times have shown that the exponential distribution is often an adequate model for holding times to assess trunk group performance. We compared the empirical histogram to the exponential distribution with the same average (Figure 2 and Table 1). We found that the exponential distribution is not a good fit to the data; it overestimates the fraction of short calls and terribly underestimates the fraction of long ones.

Classical trunk engineering, the method used in the field, is based on the Erlang-B formula predicting long run blocking rates. That formula is derived under the assumption of exponentially distributed holding times and the familiar $M/M/c/c$ queueing model (Wolff 1988). It is well-known (Sevast'yanov 1957 and Wolff 1988) that (with Poisson arrivals) the long run blocking probability, one of the trunk engineering criteria, depends on the service-time distribution only through its mean; in other words, the shape of the distribution does not matter. In light of this result, which simplified much of traffic engineering for telephony, we wondered whether the lack of a good fit by the exponential distribution should matter.

A phase type fit

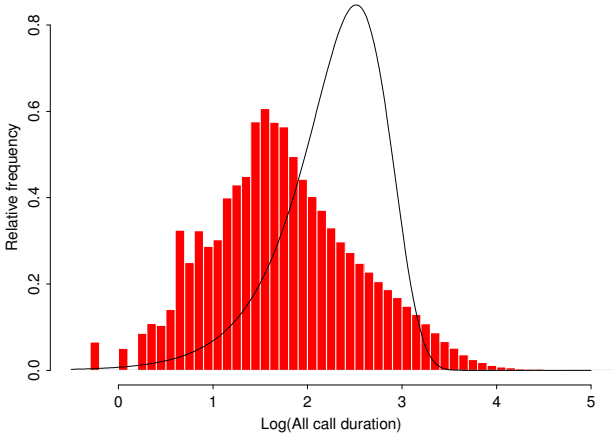

Figure 2: Note that the density for the exponential distribution with the observed mean plotted as a curve is not a good fit to the empirical histogram of holding time. For convenience of visualization, we have used a logarithmic scale (base 10) for holding times in seconds.

For a detailed queueing analysis, we fitted a phase type distribution to the data. Phase type distributions can be obtained as distributions of the time until absorption in a finite state Markov chain with one absorbing state, and they include as special cases mixtures and convolutions of exponential distributions. They are dense in the class of all distributions on the nonnegative real line; that is, they can approximate any histogram shape (Neuts 1981, Latouche & Ramaswami 1999). To fit a phase type model to the data, we employed a maximum likelihood procedure using an Expectation Maximimization (EM) algorithm (Asmussen et al 1996).

We obtained a good fit to our data using a phase type distribution based on a Markov chain with only five states (Figure 3 and Table 1). We also tried fitting with six and seven states, with no noticeable improvement. Note from Table 1 that the phase type model provides an excellent fit to the data up to the 99-th percentile.

With a well-fitting model and powerful algorithms based on matrix-geometric methods for queues (Neuts 1981, Latouche & Ramaswami 1999), we could perform extensive computations of both steady-state (long-term) and transient (time-dependent) performance measures for a queueing model representing the system. These agreed remarkably well with many field observations; for example, although internet calls account for only 5-8% of all calls in the data sets, the fraction of circuit usage due to such calls is very high (35-45%).

Internet Dial-up Calls

To examine Internet dial-up calls in detail, we extracted data on them using destination numbers. We saw that calls to Internet Service Providers (ISPs) were much longer than other calls, having a mean of 1,956 seconds (compared to an overall mean of 297 seconds) and a median of 673 seconds (compared to 48 seconds for the combined data.); see Fig. 4.

The list of ISP access numbers we used, though extensive and covering large ISPs, was not exhaustive. Therefore, we had to be inventive in obtaining the

Table 1: The table presents the percentiles for the observed holding times (Data), the exponential distribution with observed mean (Exp.), the phase type fit to the observed data (Phase), and the residual distribution corresponding to the phase type fit (Resid.) Note that the exponential distribution is a bad fit, the phase type model provides an excellent fit, and finally that the computed residual percentiles are much larger than what is observed.

%'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

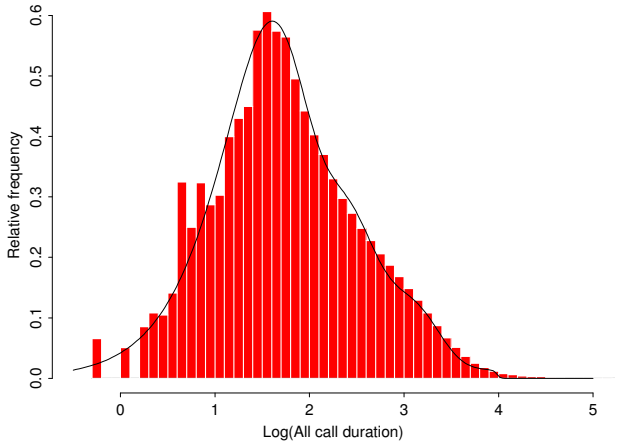

Figure 3: In this, we show the observed holding time distribution of calls on log scale (base 10) as a histogram and the density of the 5-component phase type distribution fitted to it as a curve.

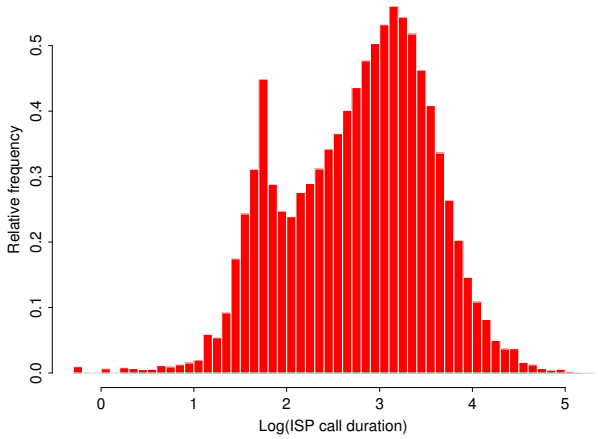

Figure 4: In this histogram of the holding time distribution for dial-up Internet calls, transformed to log scale (base 10), the bimodality is caused by two primary groups: users who dial up to download e-mail only and users who dial up to browse on the Internet.

distribution of voice call durations (Figure 5). We

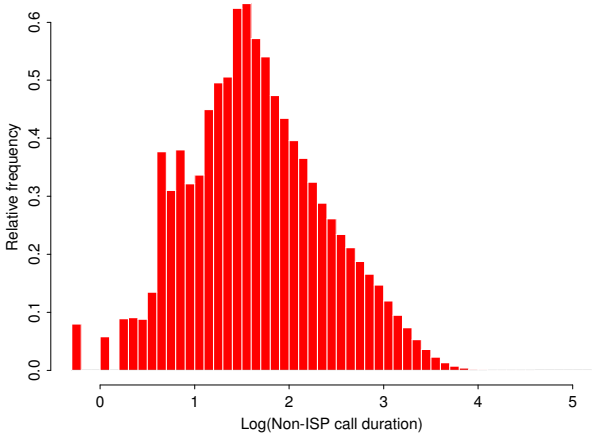

Figure 5: This histogram of the duration of voice calls, transformed to log scale (base 10), bears striking differences from the histogram for Internet dial-up calls shown in Figure 4 both in relative magnitude and distributional properties.

achieved this by considering only incoming calls to residential numbers. The resulting median and the mean for voice call durations were 40 and 190 seconds respectively. This median is close to the median value of the combined data, a natural consequence of the preponderance of voice calls in the system; the mean is more sensitive to extreme values.

Analysis of residuals

We noted already that although our data sets showed only a small percentage (5–8 per cent) of calls to be of the internet type, yet a large portion (35–45 per cent) of circuit usage was due to calls of this type. A quick way to understand this is to consider an infinite server model with two types of calls with respective average durations of 190 and 1,956 seconds and arrival rates 0.95 and 0.05 per second. The expected numbers of calls of the two types in the system can then be seen to be respectively 0.95*190=180.5 and 0.05*1956=97.8. In other words, on average, a fraction of approximately 97.8/(97.8+180.5)=35% of the busy servers will be serving the longer call type. Thus, the low value of the percentage of calls of the internet type does not imply that their effect is in-

significant; we need to weight this fraction against appropriate differences in the durations. This recognition that a large fraction of circuits can be held by the longer type of calls immediately led us to the realization that the chance is high that once a circuit is grabbed by a call, it could be held for a very long duration of time. To make this precise, we considered the "residual holding times" (see below for explanation) in the queueing model.

Given an absolutely continuous service time distribution $F(x)$, we can show that in steady state, the 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 $F(\cdot)$ is the holding time cumulative distribution function and μ is its mean. For a phase type distribution, this distribution (called the excess life or residual distribution), is also a phase type distribution and is easy to compute (Latouche & Ramaswami 1999, Chapter 3).

In our case, the computed median and mean for the residual holding time distribution were respectively 703 and 2,367 seconds. Figure 6 and Table 1 show that the predicted residual holding times are stochastically much larger than total holding times.

To determine whether the residual distribution computed from our fitted phase type distribution is consistent with real data, we collected a sample of residual holding times from a set of calls not used in the fitting procedure. We obtained a sample by fixing a time of day (6:00 PM) and recorded the remaining holding times of all calls in progress at that instant. The match between the computed residual distribution and the empirical data on residuals was remarkably good (Figure 7) except at the very highest values. The discrepancy at the very high values is caused by our omitting outliers – about 0.1% – in the data while fitting the phase type model to holding times. This empirical verification validated our modeling approach and gave greater credence to our conclusions.

One interprets the residual service time distribution as representing the remaining holding times of

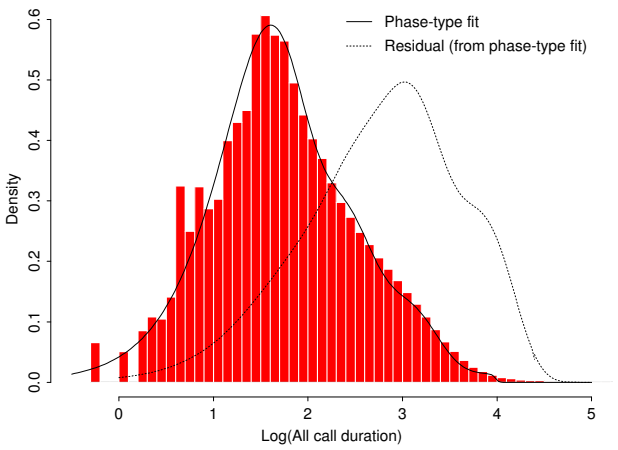

Figure 6: In this, we show the histogram of holding time distribution for all calls on log scale (base 10) and the densities for the 5-component phase type fit and the residual distribution corresponding to this fit. Note that the residual holding times are stochastically much larger than (total) holding times.

connections in progress at the sampled epoch. Given this, 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 formula, we had provided 200 circuits and that, say, 180 of these become busy. We can expect about half the busy circuits (90) to remain continuously busy for the next 703 seconds, leaving the system to operate with at most 110 circuits. That we have a large community of users and only a small percentage of calls are Internet calls implies that the reduction in capacity will not be accompanied by a corresponding reduction in demand for circuits. In short, for noticeable periods, congestion and the resultant blocking for circuits could be much higher than the engineered level.

The above observations help to reconcile why engineering based on the Erlang-B formula does not provide adequate protection from such congestion. The Erlang-B formula yields only the steady state blocking probability, which is the long-run blocking rate

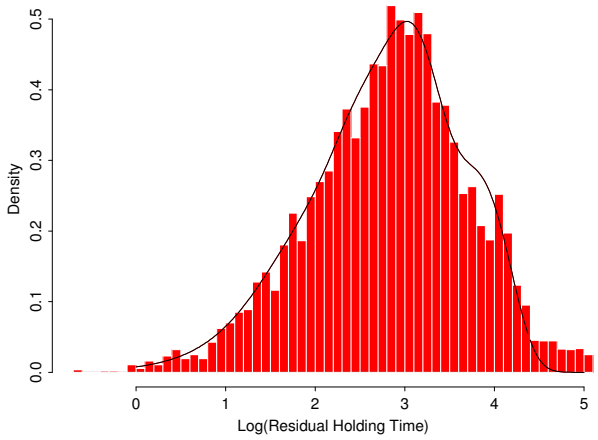

Figure 7: We show the histogram of the residual holding time distribution on the log scale (base 10) for a random sample of calls along with the density curve for the 5-component phase type fitted residual distribution. The remarkable match validated our modeling approach.

over an infinite horizon of time. This long-run performance measure is not a good descriptor of the short-run (transient) blocking rates that govern customers' experience. The mathematical underpinnings of our findings lie in a length-biasing argument similar to that in the waiting time paradox, namely, the interval in a Poisson process covering a randomly chosen point has twice the mean of a typical interval between events (Feller 1971). When sampling at a random epoch, we are likely to find a larger fraction of long-holding-time calls among those currently in the system than we would predict based on the overall fraction of calls of that type, because long calls tend to get stuck in the system and are more likely than short calls to be seen by the observer.

To examine these effects, we simulated the transient blocking probabilities for two access groups of 96 circuits, both engineered to obtain the same level (0.01) of blocking in steady-state and to have the same mean holding times matching the real data. In one system, the holding times follow our phase type distribution matching the data, while in the other system the holding times follow an exponential distribution with the same mean. With long holding times, convergence to steady-state occurs more slowly (Figure 8, Figure 9). In practical terms, long-holding-time calls lead to long "relaxation times" for the system; that is, when congestion occurs, the congestion will also persist.

Need for Controls

Large circuit groups are more efficient than small ones in the sense that for a given blocking rate, they can support a greater use of resources. For instance, circuit utilization that attains a 1% blocking rate is only about 64% for a 24-circuit group, while it is about 86% for a 120-circuit group. Thus, combining circuit groups produces an economy of scale. This may suggest that we could increase the size of the access groups to solve the congestion problem.

When we enlarge circuit groups without increasing the number of users, we will decrease the chance of congestion, because we provide more resources for the same load. Doing so, however, may be uneconomical because of the resulting low utilization of circuits. Furthermore, in our context of long holding times,

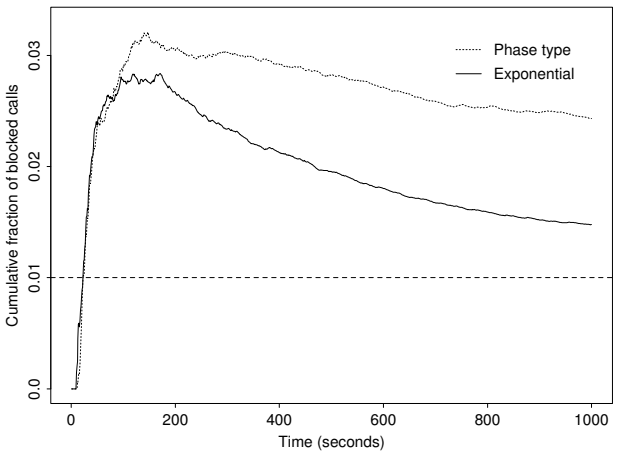

Figure 8: Comparison of cumulative fractions of attempts blocked for two 96 circuit access groups, both starting with 90 busy circuits and engineered to have long term blocking rate is 0.01, shows that with long holding times, it takes much longer for the system to recover from congestion.

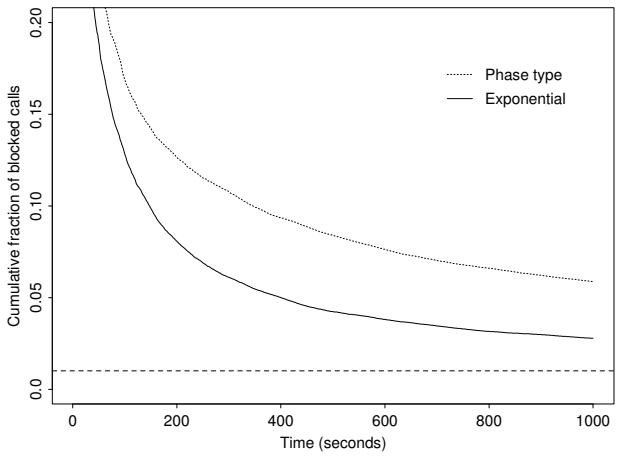

Figure 9: Comparison of cumulative fractions of attempts blocked for two 96 circuit access groups, both engineered to have long-term blocking rate of 0.01 and starting with all circuits busy, shows that congestion and high blocking rates persist much longer for the system with long holding times.

the behavior of the system under congestion — which from a mathematical perspective depends on the conditional distributions given high occupancy levels — would not, however, improve significantly; this fact can be verified through simulations or through analytic computations. Thus, providing additional circuits will not protect the system from unanticipated overloads due to sudden changes in traffic patterns or due to circuit failures.

In addition, in large circuit groups with loads to match a blocking rate, one must consider some subtleties. 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. At these values, the smaller group still has a spare circuit, but the larger group is exhausted. This shows that the economy of scale in large server groups comes with a potentially increased risk. To achieve the efficiencies, we have to load the larger groups to higher levels, and that increases the risk of saturation. In the presence of long holding times, that risk translates to persistent congestion when congestion occurs.

We noted that blocking rates at access points did not indicate congestion, an anomaly we observed in the data. At the access portions, the individual trunks each had only 24 circuits, while upstream in the access portion they were combined into larger groups of 120 or more circuits with an added level of concentration. The above discussion shows that problems could occur upstream with no downstream indication. In addition, we identified another cause of the anomaly. Small circuit groups behave like finite source systems that are self-regulating in the sense that the arrival rates decrease as more circuits get busy (since busy sources cannot generate new calls). But such effects diminish as we increase server group sizes, and large service groups tend to behave like infinite source systems. Being large, the groups upstream in the access network do not exhibit the self-regulating behavior of those downstream.

The prudent approach is to manage congestion by augmenting the procedures for determining the system sizes with sound admission and overload controls, so that we can maintain service quality at moderate overloads and protect essential services even under significant overloads. However, controls should not trigger too frequently and cause customer complaints.

While uncontrolled systems behave badly in the presence of long holding times, we showed that they are good candidates for control. In the presence of long holding times, after a control relieves congestion, the system will remain in the uncongested state for long periods of time (Figure 10). This is so since if blocking rates of two systems over the long haul is the same but one suffers persistence of congestion when it occurs, then to obtain the same long run average, the system subject to persistent congestion should also have compensating periods of good performance that are much longer. Thus, our situation is one where controls can be effective in that once they are exercised, the system will return to a stable state and remain there for significant amounts of time.

Congestion Controls

We have observed that when congestion occurs many circuits are likely to be held by ISP dial-up

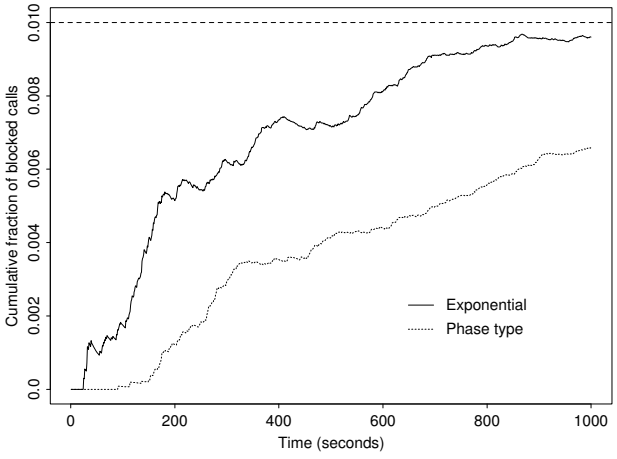

Figure 10: Comparison of simulated cumulative fractions of blocked attempts shows that uncongested periods are also relatively more persistent in the presence of long holding times. Here two systems of 96 access circuits, both starting with only 81 busy circuits, are compared. Note that for the phase type model with long holding times, the observed blocking rates are consistently below the one for the comparable exponential model.

connections that last long. Thus, we need to control ISP calls, and for this we must first be able to identify such call attempts. We may also consider giving high priority to certain types of calls, such as those dialing 911. For all this, we need to have the ability to receive the digits dialed by the customer so that from it we can discern the type of the call that is being attempted. Thus, one of our main efforts was to ensure that the customer can get dial tone and dial the digits with a very high probability. To that end, we effected techniques that are tantamount to reserving a small number of circuits for the dial-tone and digit-reception phase of the call. Once we did that, we developed a solution based on the following principles: (1) under congested conditions, do not admit new ISP dial-ups; (2) under extremely congested conditions, terminate an ongoing ISP connection to make room for a voice call. With regard to (1), we can set a small threshold T and reject an ISP attempt when the number of free circuits F is less than T immediately following digit analysis and prior to the decision to route the call. The system then can maintain, with a high probability, the availability of a circuit for giving dial tone and receiving digits. The threshold T can be quite small since dial tone and digit reception take very little time; this is comforting since a large value for T would result in rejecting too many ISP attempts. However, if a call attempt finds fewer than T free circuits at the end of digit analysis and happens to be voice, we may accept it provided the number I of ongoing ISP calls in the system is greater than a preassigned threshold K; in that case, we terminate one of the ongoing ISP connections to prevent a no-dial-tone condition in the near future. We thus maintain, with a high probability, a favorable call mix in the system resulting in frequent release of circuits for being available for providing dial tone and digit reception. We do this without reserving some specific set of T circuits; that avoids the hassle of switching calls to different circuits after digit analysis and also obviates concerns about failures in the reserved circuits.

We must choose K judiciously so that (1) under engineered loads and under moderate departures therefrom, the chance of terminating an ongoing ISP call is small, and (2) the chance of repeated hits on the same caller is negligible. We could achieve (2) by controlling the overall probability of premature termination for ISP calls and by selecting calls to be terminated carefully, for example, terminating the longest call or one that has exceeded a given threshold of time. The heavy tailed nature of ISP holding times ensures a high probability that we will find an ISP call with a long elapsed time during periods of congestion, and since preempted callers return as the youngest in the system, they will be unlikely to be hit again. We validated thse intuitive considerations with our model computations and simulations.

Performance Results

A simplified version of our control is presented in Figure 11. We describe a sample set of results for 96 circuits to illustrate how it works. We assume the mean duration of the call setup phase to be 3 seconds, and we used the observed values of 190 and 1,956 seconds for means for voice and ISP calls.

We compare the situation with no control to that

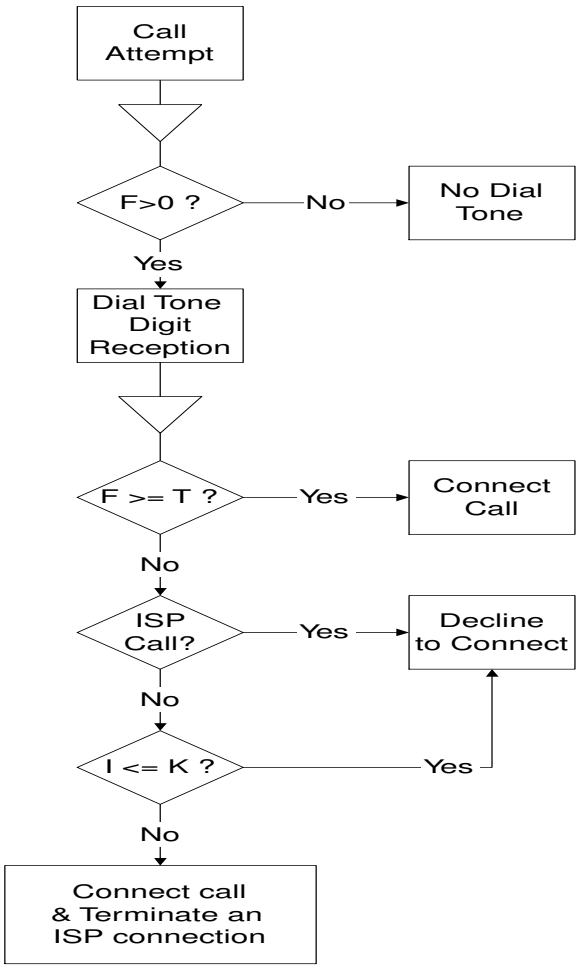

Figure 11: The control algorithm is presented as a flow chart. In this F denotes the number free circuits in the system, I the number of circuits held by ISP callers, and T and K are two thresholds to be chosen suitably.

in which we use the controls $T = 1$, $K = 26$ (Table 2). 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 in ISP calls. Although under our controls there is blocking even after getting a dial tone, since ISP calls form only a small fraction of the total calls, the actual attempts blocked are few.

An important performance measure is the chance that an ISP call is pre-empted midstream. For the various load levels considered, these probabilities are 0.058, 0.165, 0.255 and 0.303, respectively. The probability increases noticeably only under loads far above levels at which networks usually operate. Given that preemption occurs only when ISP calls number more than 26 and that we select preempted calls judiciously (for example, the oldest), we can make sure that we do not interrupt the same caller repeatedly. If the system becomes congested, the control reduces congestion quickly by changing the call mix and increasing the churn of circuits by allowing more voice calls, which are shorter.

Table 3 provides a comparison of performance in the presence of congestion. Without control, the system performance degrades drastically as the load in-

Table 2: We give the no-dial-tone (No-DT) and call rejection probabilities for the system of 96 circuits with no control and with the control. The load factors (Erlang) used 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 3: For a 96 circuit system, we provides the conditional no-dial-tone and call rejection probabilities given that the number of circuits busy is at least min(96ρ, 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

creases. The control makes that degradation graceful, particularly for voice calls, while maintaining negligible no-dial-tone probabilities for all . The fraction of ISP calls rejected is reasonably close to the case of not exercising any control, and yet those selective rejections buy much by way of performance, yielding a graceful degradation of service.

Thus, the admission and overload controls embodied in our algorithm work well in preventing no-dial-tone conditions while at the same time providing high circuit utilization and satisfactory performance to both types of callers.

Other Enhancements

The congestion controls we proposed need to be augmented by control of reattempts. In practice, computers and modems reattempt at much faster rates and more often than people making voice calls. It is customary in telephony to manage the dial tone queue with a Last-in-First-Out (LIFO) scheme so that processing capacity is not wasted on customers who get impatient and disconnect. While this scheme in classical telephony helps to maintain high levels of good throughput (Forys 1983, Zhao & Alfa 1995), in our context it would defeat the intent of our overload control policies. Circuits released by the overload control would have a high chance of being grabbed immediately again by long holding time calls since such calls reattempt at a faster rate and more frequently than voice calls. Thus, we had to develop new queueing strategies. We set up separate queues and weighted polling schemes that in effect introduced delays between successive attempts by modem calls under congestion.

The notion of disconnecting an ongoing call is anathema in the telephone world. Although we showed that our controls would be exercised rarely, that once exercised they might not be invoked again soon, and that they could be implemented in a way avoiding hitting the same caller more than once (say, in a week or a month), the service provider may want to avoid disconnecting ongoing calls altogether. However, that would rquire system development and modification. We developed approaches that convert ISP connections transparently to packet mode (in which connections are shared by calls which transmit their payload in chunks called packets and thereby do not waste capacity) and also various compression techniques to further minimize bandwidth wastage.

The procedures may have many other uses, and they are the subjects of three pending patent applications.

Research Issues We were inspired by problems in an area (circuit switching) that is now considered classical and in its decline, but it offers many useful pointers for further research in new areas, including those related to the Internet and broadband technologies:

(1) In high speed communications, application payloads can differ in size and access speeds. We need to examine the effects of these differences carefully, particularly as they affect short-term performance. During congestion, the system may show a bias towards having large numbers of faster connections or heavier payload applications. This is an important topic for research as it pertains to the efficient multiplexing of highly disparate services on a common packet network.

(2) Various dynamic routing algorithms determine when to take an alternate route for a call that cannot be routed directly. The alternate route may require more resources (trunks or bandwidth paths), leading to increased blocking on those portions. The computations for developing such algorithms are almost entirely based on steady-state (long-term, infinite-horizon) blocking calculations and assumptions of exponentiality, which could be misleading when holding times are disparate. A similar situation can obtain in packet routing schemes used in high speed networks, such as those based on label switching. We need a clearer understanding of the stochastic length biasing effects in these areas to prevent unanticipated problems of congestion.

(3) In detecting denial of service attacks caused by transmission of large payloads, simple-minded algorithms based on observed payload sizes along a path could cause many false alarms if they do not take into account the disparity in normal payload sizes and their effect on observed measurements.

(4) In modeling the performance of call centers, which are estimated to employ about 3% of the U.S. workforce, analysts use the existence of what are called Quality and Efficiency Driven (QED) regimes

of operation that ensure both service quality (few delays) and operational efficiency (high operator usage) simultaneously. (Halfin & Whitt 1981). These results are based on assumptions of exponentiality and steady-state performance measures, and we need to consider the impact of exceptionally long service times on such systems. Without some controls to handle exceptional demands, calculations based on asymptotic results alone may lead to unwanted surprises.

In short, our work opens up several important research issues.

Concluding Remarks

Our work provided an increased ability to maintain high levels of call completion rates and circuit usage. These, combined with the attendant avoidance of many truck rolls for frequent load balancing and maintenance activities, will save millions of dollars in capital and maintenance costs. Our data analysis revealed not only that a large percentage of circuit usage was attributable to internet calls, but that such calls were being routed to ISPs via third party switches. That discovery also created an opportunity for AT&T to effect significant savings of the order of $15 million per year in transit costs to other carriers through more efficient routing of such calls.

The non-trivial and counter-intuitive results of our work stem from certain length biasing effects, and the quantification of those effects would have been impossible without detailed modeling and interpretation of results based on practical experience. Arguments based on averages miss the significant temporal variability in customer perceived performance. The probabilistic analysis explains the source of such variability and points to solution approaches that are meaningful.

In the words of Dr. Hossein Eslambolchi, President, AT&T Global Networking Technology Services:

> "AT&T prides itself as a leader in telecommunications research, and this work exemplifies that leadership. Its non-trivial, subtle, and counter-intuitive findings of high practical value demonstrate the capabilities of our innovative researchers and our ability to bring their talent to serve our customers expeditiously and well."

The value of Operations Research is often asserted in monetary terms. But what monetary value can one attach to preventing potential loss of lives? We will never know how many lives this work that ensures access to emergency services has saved or will save, but we do know that our network no longer suffers from dial-tone related problems. In that, our work demonstrates that O.R. is not only "a science of better" but also a science of the safer and more secure.

References

Asmussen, S., Nerman, O., Olsson, M. 1996. Fitting phase-type distributions via the EM algorithm. *Scand. J. Stat.,* 23(4), 419-441.

Chaudhury, G.L., Heyman, D.P., Kaplan, A.E., Ramaswami, V. 2004. Method for preventing overload condition in a circuit switched arrangement. United States Patent 6,731,740, issued: May 4, 2004.

Kaplan, A.E., Ramaswami, V. 2004. Method for preventing overload condition in a circuit switched arrangement. United States Patent 6,826,873, issued: November, 2004.

Feller, W. 1971. *An Introduction to Probability Theory and its Applications.* 2nd ed., vol. 2. John Wiley and Sons, New York.

Forys, L.J. 1983. Performance analysis of a new overload control strategy. *Proc. 10th International Teletraffic Congress (ITC 10)*, Montreal, Canada, Paper No. 4.

GR-303-CORE, Issue 4, 2000. IDLC Generic Requirements, Objectives & Interfaces. Dec. 2000, Telcordia.

Grassmann, W.K., Taksar, M.I., Heyman, D.P. 1985. Regenerative analysis and steady state distributions for Markov chains. *Operations Research,* 33(5), 1107-1116.

Halfin, S., Whitt, W. 1981. Heavy–traffic limits for queues with many exponential servers. *Operations Research*, 29(3), 567-587.

Latouche, G., Ramaswami, V. 1999. *Introduction to Matrix Analytic Methods in Stochastic Modeling.* SIAM and ASA, Philadelphia, PA.

Neuts, M.F. 1981: *Matrix-Geometric Solutions in Stochastic Models.* Johns Hopkins Univ. Press, Baltimore, MD.

Ramaswami, V., Neuts, M.F. 1980. Some explicit formulas and computational methods for infinite server queues with phase type arrivals. *J. Appl. Probab.,* 17(2), 498-514.

Ramaswami, V., Poole, D. Ahn, S., Byers, S. Kaplan, A.E. 2003. Containing the effects of long holding time calls due to Internet dial-up connections. *Proc. IEEE PACRIM Conf.*, Victoria, Canada.

Sevast'yanov, B.A. 1957. An ergodic theorem for Markov processes and its application to telephone systems with refusals. *Teor. Verojatnost. I Primenen,* 2(1), 106-116 (Russian with English Summary).

Wolff, R.W. 1988. *Stochastic Modeling and the Theory of Queues.* Prentice-Hall, London, England.

Zhao, Y.Q., Alfa, A.S. 1995. Performance analysis of a telephone system with both patient and impatient customers. *Telecommunication Systems,* 4(3), 201-215.

APPENDIX

Queueing Models & Methods

For phase type holding times, the variables defining the Markov chain are the number of callers in various states (idle, dial-tone and digit-reception, transmitting voice or data) as well as various phases of the phase type model describing their type and durations. In our models, to limit an explosion in dimensionality, we used exponential distributions for the idle and dial-tone and digit-reception phases, but we used the phase type distribution fitted to data for call durations.

The model may involve a large number of states depending on the number of circuits and the number of phases. Even with exponential distributions for the three call states and two types of calls, the dimensionality of the Markov chain is $\binom{N+3}{3}$ for a set of N circuits and grows quickly with N; for instance, for $N = 120$, we have a total 280,840 states in the Markov chain. With a p state Markov chain describing holding times, the state space dimensionality is $\binom{N+p+2}{p+2}$, and exact computations become quite delicate. Thus, for moderately large values of N, our recommendation is to use an approximation with two or three phases to determine the steady state probabilities but to include more detail for transient computations, such as for the probability of preempting an ISP call during its lifetime. For large values of N, approximations based on the infinite-server model work well.

To handle large models, in addition to standard techniques for dealing with sparse matrices (for example, storing only nonzero entries and their locations), it helps to reduce the size of the model up front by eliminating states that are visited with very small probabilities. For certain algorithms, such as the state reduction method (Grassmann et al 1985), this step is essential; otherwise, programs could terminate abnormally with a division by zero due to near reducibility of the model. One simple way to trim the state space is to use the steady-state results for an infinite-server model to discard states with negligible steady-state probabilities; the necessary steady-state results are available in simple form even in the case of phase type models (Ramaswami & Neuts 1980, Theorem 8.8). As an example, with three seconds as the average for the dial-tone and digit-reception steps, we found that (at meaningful ranges of load) we could comfortably truncate the number of circuits in the call setup (dial-tone/digit-reception) phase at a small value (for example, 8 while considering a circuit group of size 120.)

Even for moderately large trunk group sizes (92 to 116 are typical of the field), we could make detailed computations with the above approaches because the models are highly structured (for example, block tri-diagonal matrices with embedded tridiagonal blocks), and one can use efficient algorithms based on standard matrix analytic methods (Latouche & Ramaswami 1999).

Once we have determined the steady state distri-

bution of the Markov chain, we can obtain various conditional stationary distributions, such as the conditional distribution given the number of busy circuits. These conditional distributions help, among other things, to assess how service level degrades as congestion builds up in the system; our goal was also to effect a graceful degradation of service under overloads. We also used these conditional distributions to quantify the fraction of ISP calls in the system at various levels of utilization.

We can compute the probability of forcibly terminating an ISP call as the absorption probability in a suitably chosen Markov chain. Through computations similar to those for evaluating a phase type distribution (Latouche & Ramaswami 1999, Chapter 2), we can also evaluate the conditional distribution of the elapsed lifetime of a forcibly terminated call, which has a bearing on customer experience for ISP calls. Our analysis shows that for well-engineered systems operating with our controls, the elapsed time of an ejected ISP call is of the order of six hours or more with a high probability, making it highly unlikely that an ongoing transmission is impaired. In our analysis, we assumed that the call to be terminated is chosen randomly from among ongoing ISP calls (instead of the oldest), which provides pessimistic estimates for performance measures, such as the expected elapsed time of a call suffering premature termination.

For these computations, we assigned for the initial state of the Markov chain the steady-state distribution at the instant of arrival of an ISP call which, because we assume Poisson arrivals, is the same as the stationary distribution of the original Markov chain. The two absorbing states correspond to two events: normal termination of the marked ISP call; its preemption by an incoming voice call whose acceptance without preemption would not leave T free circuits. For the special case when all distributions are exponential, we can show that the probability of having to make a premature termination of an ISP call reduces to the following intuitive expression

$$\frac{\sum_{v+d+s>N-T,\ d>K} \pi(v,d,s) s\mu_s p_v}{\sum_{v,d,s} \pi(v,d,s) s\mu_s p_v},$$

where, $\pi(v,d,s)$ is the steady-state probability of finding v voice calls and d ISP calls in the system along with s circuits busy in the call setup phase, μ_s is the rate at which setups (dial tone and digit reception) complete, and p_v is the proportion of voice attempts. In the general case, we can express these quantities in terms of the steady-state probabilities and the rates specifying the phase type holding time distribution and can compute them efficiently.