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

V. Ramaswami, David Poole
AT&T Labs, 180 Park Avenue, Florham Park, New Jersey 07932 {vramaswami@att.com, dpoole@att.com}

Soohan Ahn
Department of Statistics, University of Seoul, Seoul 130-743, South Korea, sahn@uos.ac.kr

Simon Byers
AT&T Labs, 180 Park Avenue, Florham Park, New Jersey 07932, sbyers@att.com

Alan Kaplan
Unique Technical Solutions, 52 Brookside Drive, Upper Saddle River, New Jersey 07458, ae_kaplan@yahoo.com

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 and 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.

Key words: facilities: equipment planning, capacity expansion; industries: communications.

In 2001, AT&T received complaints from customers in one of its market segments that they were not getting dial tones 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 the 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 daytime 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 with a delay. Because 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 PhD dissertation (Ramaswami and Neuts 1980). That work obtained some counterintuitive 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 upstream 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 for several other areas of our business.

(In this paper, we present our work in a manner accessible to a general audience. In a more technically oriented paper (Ramaswami et al. 2003), we provide additional details.)

**Call Holding Times**

The data for 4.5 million residential calls over a week in one serving area form a distribution with a long tail and a mean of 297 seconds that significantly exceeds three minutes, the highly quoted value for average voice-call duration (Figure 1). While we could attribute these characteristics to the use of circuits by some customers for Internet dial-up, we also found that only a small fraction (six percent) of the calls were of the Internet type and needed to determine whether that small fraction could indeed cause congestion of the type suspected.

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, 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, 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

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 and Ramaswami 1999). To fit a phase-type model to the data, we employed a maximum-likelihood procedure using an expectation maximization (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, Table 1). We also tried fitting with six and seven states, with no noticeable improvement. The phase-type model provides an excellent fit to the data up to the 99th percentile (Table 1).

With a well-fitting model and powerful algorithms based on matrix-geometric methods for queues (Neuts 1981, Latouche and 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 five to eight percent of all calls in the data sets, the fraction of circuit usage due to such calls is very high (35 to 45 percent).

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) (Figure 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

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>Data</th>
<th>Exponential distribution</th>
<th>Phase-type fit</th>
<th>Residual distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>5.4</td>
<td>31.3</td>
<td>5.8</td>
<td>40.8</td>
</tr>
<tr>
<td>20</td>
<td>12.0</td>
<td>66.3</td>
<td>12.7</td>
<td>116.1</td>
</tr>
<tr>
<td>30</td>
<td>21.0</td>
<td>105.9</td>
<td>21.2</td>
<td>238.0</td>
</tr>
<tr>
<td>40</td>
<td>32.4</td>
<td>151.7</td>
<td>32.2</td>
<td>423.7</td>
</tr>
<tr>
<td>50</td>
<td>48.0</td>
<td>205.9</td>
<td>47.7</td>
<td>703.0</td>
</tr>
<tr>
<td>60</td>
<td>72.6</td>
<td>272.1</td>
<td>72.2</td>
<td>1,124.6</td>
</tr>
<tr>
<td>70</td>
<td>120.6</td>
<td>357.6</td>
<td>120.2</td>
<td>1,809.7</td>
</tr>
<tr>
<td>80</td>
<td>232.8</td>
<td>478.0</td>
<td>235.9</td>
<td>3,285.3</td>
</tr>
<tr>
<td>90</td>
<td>601.8</td>
<td>683.9</td>
<td>597.6</td>
<td>7,028.2</td>
</tr>
<tr>
<td>95</td>
<td>1,237.8</td>
<td>889.7</td>
<td>1,268.0</td>
<td>11,073.8</td>
</tr>
<tr>
<td>99</td>
<td>3,952.2</td>
<td>1,367.7</td>
<td>4,035.2</td>
<td>20,489.2</td>
</tr>
<tr>
<td>99.5</td>
<td>6,074.4</td>
<td>1,573.6</td>
<td>7,168.4</td>
<td>24,470.7</td>
</tr>
</tbody>
</table>

Table 1: In this table, the columns are the percentiles for the observed holding times, the holding times (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. The exponential distribution is a bad fit, the phase-type model provides an excellent fit, and the computed residual percentiles are much larger than what is observed.
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.

the distribution of voice-call durations (Figure 5). We 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

Although our data sets showed only a small percentage (five to eight percent) of calls to be of the Internet type, a large portion (35 to 45 percent) of circuit usage was for 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 would 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 percent 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 insignificant; we need to weight this fraction against appropriate differences in durations. Recognition that a large fraction of circuits can be held by the longer type of calls immediately caused us to realize that the chance is high that once a circuit is grabbed by a call, it could be held for a very long time. To make this precise, we considered residual holding times 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 $\mu$ 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 and 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. The predicted residual holding times are stochastically much larger than total holding times (Figure 6, Table 1).

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

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 (Figure 4) both in magnitude and distributional properties.
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 percent—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.

We interpret the residual service-time distribution as representing the remaining holding times of 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, that 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.

These observations explain 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 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 one percent blocking rate is only about 64 percent for a 24-circuit group, while it is about 86 percent 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, 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. These findings show 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 (because 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 a long period of time (Figure 10). This is so for the following reason: If blocking rates of two systems over the long haul are 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 in which controls can be effective in that, once they are exercised, the system will return to a stable state and remain there for a long time.

**Congestion Controls**

We have observed that when congestion occurs many circuits are likely to be held by ISP dial-up connections that last for a long time. Thus, we need to control ISP calls, and to do so we must first be able to identify such call attempts (Figure 11).
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, which ensures frequent release of circuits and their availability for providing dial tone and digit reception. We do this without reserving some specific set of $T$ circuits, which 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 because preempted callers return as the youngest in the system, they will be unlikely to be hit again. We validated these intuitive considerations with our model computations and simulations.

**Performance Results**

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 three 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 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, some customers face blocking even after getting a dial tone, because

<table>
<thead>
<tr>
<th>Erlang</th>
<th>No control</th>
<th>$T = 1$ and $K = 26$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No DT</td>
<td>No DT</td>
</tr>
<tr>
<td></td>
<td>Reject ISP</td>
<td>Reject ISP</td>
</tr>
<tr>
<td>76.8</td>
<td>0.0045</td>
<td>0.00004</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0005</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0043</td>
</tr>
<tr>
<td>86.4</td>
<td>0.0282</td>
<td>0.00019</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0049</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0155</td>
</tr>
<tr>
<td>96.0</td>
<td>0.0772</td>
<td>0.00057</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0210</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0374</td>
</tr>
<tr>
<td>105.6</td>
<td>0.1374</td>
<td>0.00136</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0531</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0727</td>
</tr>
</tbody>
</table>

Table 2: We show 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 $\rho$ values 0.8, 0.9, 1.0, and 1.1.
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 preempted midstream. For the various load levels considered, these probabilities are 0.058, 0.165, 0.255, and 0.303. 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 than ISP calls.

Without control, the system performance degrades drastically as the load increases (Table 3). 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 fraction blocked with no 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 and 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 because 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 require system development and modification. We developed approaches that convert ISP connections transparently to packet mode (in which connections are shared by calls that 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 exist 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 three percent of the US 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 and 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 AT&T with an increased ability to maintain high levels of call completion rates and circuit usage. These results, 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 on the order of $15 million per year in transit costs to other carriers through more efficient routing of such calls.

The nontrivial 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 Hossein Eslambolchi, president of AT&T Global Networking Technology Services:

AT&T prides itself as a leader in telecommunications research, and this work exemplifies that leadership. Its nontrivial, subtle, and counterintuitive 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 operations research is not only a science of better but also a science of the safer and more secure.

Appendix

Queueing Models and 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) and 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 \( (N + 3)^3 \) 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 \( (N + p + 2)^2 \), 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 and 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 or 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 tridiagonal matrices with embedded tridiagonal blocks), and one can use efficient algorithms based on standard matrix analytic methods (Latouche and Ramaswami 1999).

Once we have determined the steady-state distribution 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 and 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 premption 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:
\[
\frac{\sum_{v+d+s>N-T, d \geq K} \pi(v, d, s) \mu_k p_v}{\sum_{v, d, s} \pi(v, d, s) \mu_k p_v},
\]
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, \( \mu_k \) 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.

References


