12

TRAFFIC ANALYSIS

Except for station sets and their associated loops, a telephone network is composed of a variety of common equipment such as digit receivers, call processors, interstage switching links, and interoffice trunks. The amount of common equipment designed into a network is determined under an assumption that not all users of the network need service at one time. The exact amount of common equipment required is unpredictable because of the random nature of the service requests. Networks conceivably could be designed with enough common equipment to instantly service all requests except for occurrences of very rare or unanticipated peaks. However, this solution is uneconomical because much of the common equipment is unused during normal network loads. The basic goal of traffic analysis is to provide a method for determining the cost-effectiveness of various sizes and configurations of networks.

Traffic in a communications network refers to the aggregate of all user requests being serviced by the network. As far as the network is concerned, the service requests arrive randomly and usually require unpredictable service times. The first step of traffic analysis is the characterization of traffic arrivals and service times in a probabilistic framework. Then the effectiveness of a network can be evaluated in terms of how much traffic it carries under normal or average loads and how often the traffic volume exceeds the capacity of the network.

The techniques of traffic analysis can be divided into two general categories: loss systems and delay systems. The appropriate analysis category for a particular system depends on the system's treatment of overload traffic. In a loss system overload traffic is rejected without being serviced. In a delay system overload traffic is held in a queue until the facilities become available to service it. Conventional circuit switching operates as a loss system since excess traffic is blocked and not serviced without a retry on the part of the user. In some instances "lost" calls actually represent a loss of revenue to the carriers by virtue of their not being completed.

Store-and-forward message or packet switching obviously possesses the basic characteristics of a delay system. Sometimes, however, a packet-switching operation can also contain certain aspects of a loss system. Limited queue sizes and virtual circuits both imply loss operations during traffic overloads. Circuit-switching networks also incorporate certain operations of a delay nature in addition to the loss operation of the circuits themselves. For example, access to a digit receiver, an operator, or a call processor is normally controlled by a queuing process.

The basic measure of performance for a loss system is the probability of rejection (blocking probability). A delay system, on the other hand, is measured in terms of service delays. Sometimes the average delay is desired, while at other times the probability of the delay exceeding some specified value is of more interest.

Some of the analyses presented in this chapter are similar to those presented in Chapter 5 for the blocking probabilities of a switch. Chapter 5 is concerned mostly with matching loss—the probability of not being able to set up a connection through a switch under normal or average traffic volumes. This chapter, however, is mostly concerned with the probability that the number of active sources exceeds some specified value. Typically, the specified value is the number of trunk circuits in a route.

12.1 TRAFFIC CHARACTERIZATION

Because of the random nature of network traffic, the following analyses involve certain fundamentals of probability theory and stochastic processes. In this treatment only the most basic assumptions and results of traffic analysis are presented. The intent is to provide an indication of how to apply results of traffic analysis, not to delve deeply into analytical formulations. However, a few basic derivations are presented to acquaint the user with assumptions in the models so they can be appropriately applied.

In the realm of applied mathematics, where these subjects are treated more formally, blocking probability analyses are referred to as congestion theory and delay analyses are referred to as queuing theory. These topics are also commonly referred to as traffic flow analysis. In a circuit-switched network, the "flow" of messages is not so much of a concern as are the holding times of common equipment. A circuit-switched network establishes an end-to-end circuit involving various network facilities (transmission links and switching stages) that are held for the duration of a call. From a network point of view, it is the holding of these resources that is important, not the flow of information within individual circuits.

On the other hand, message-switching and packet-switching networks are directly concerned with the actual flow of information, since in these systems traffic on the transmission links is directly related to the activity of the sources.

As mentioned in Chapter 7, circuit switching does involve certain aspects of traffic flow in the process of setting up a connection. Connect requests flow from the sources to the destinations acquiring, holding, and releasing certain resources in the process. As was discussed, controlling the flow of connect requests during network overloads is a vital function of network management.

The unpredictable nature of communications traffic arises as a result of two underlying random processes: call arrivals and holding times. An arrival from any particular user is generally assumed to occur purely by chance and be totally independent of arrivals from other users. Thus the number of arrivals during any particular time interval is indeterminate. In most cases holding times are also distributed randomly. In some applications this element of randomness can be removed by assuming constant holding times (e.g., fixed-length packets). In either case the traffic load presented to a network is fundamentally dependent on both the frequency of arrivals and the average holding time for each arrival. Figure 12.1 depicts a representative situation in which both the arrivals and the holding times of 20 different sources are unpredictable. The bottom of the figure depicts activity of each individual source while the top displays the instantaneous total of all activity. If we assume that the 20 sources are to be connected to a trunk group, the activity curve displays the number of circuits in use at any particular time. Notice that the maximum number of circuits in use at any one time is 16 and the average utilization is a little under 11 circuits. In general terms, the trunks are referred to as servers, and a trunk group is a server group.

Traffic Measurements

One measure of network capacity is the volume of traffic carried over a period of time. Traffic volume is essentially the sum of all holding times carried during the interval. The traffic volume represented in Figure 12.1 is the area under the activity curve (approximately 84 call minutes).

A more useful measure of traffic is the traffic intensity (also called traffic flow). Traffic intensity is obtained by dividing the traffic volume by the length of time during which it is measured. Thus traffic intensity represents the average activity during a period of time (10.5 in Figure 12.1). Although traffic intensity is fundamentally dimensionless (time divided by time), it is usually expressed in units of erlangs, after the Danish pioneer traffic theorist A. K. Erlang, or in terms of hundred (century) call seconds per hour (CCS). The relationship between erlangs and CCS units can be derived by observing that there are 3600 sec in an hour:

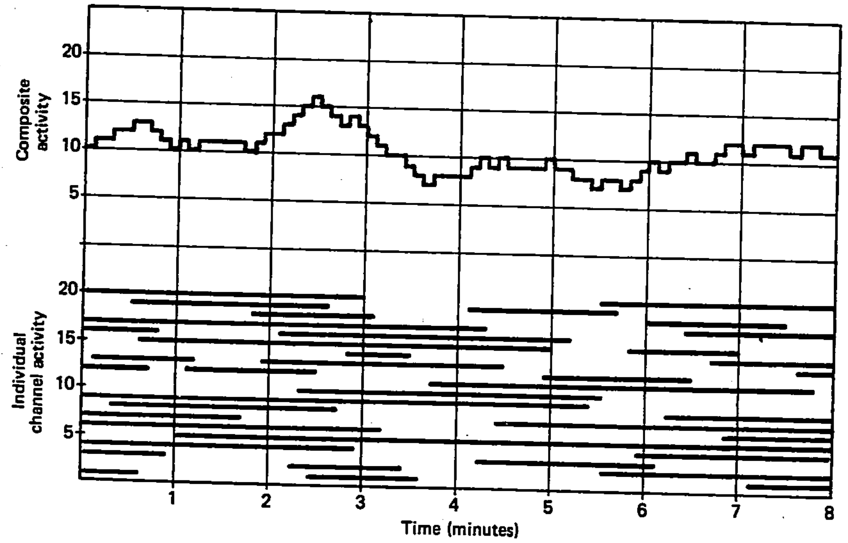

Figure 12.1 Activity profile of network traffic (all calls carried).

$$1 \text{ erlang} = 36 \text{ CCS}$$

The maximum capacity of a single server (channel) is 1 erlang, which is to say that the server is always busy. Thus the maximum capacity in erlangs of a group of servers is merely equal to the number of servers. Because traffic in a loss system experiences infinite blocking probabilities when the traffic intensity is equal to the number of servers, the average activity is necessarily less than the number of servers. Similarly, delay systems operate at less than full capacity, on average, because infinite delays occur when the average load approaches the number of servers.

Two important parameters used to characterize traffic are the average arrival rate λ and the average holding time t_m. If the traffic intensity λ is expressed in erlangs, then

$$A = \lambda t_m \tag{12.1}$$

where λ and t_m are expressed in like units of time (e.g., calls per second and seconds per call, respectively).

Notice that traffic intensity is only a measure of average utilization during a time period and does not reflect the relationship between arrivals and holding times. That is, many short calls can produce the same traffic intensity as a few long ones. In many of the analyses that follow the results are dependent only on the traffic intensity. In some cases, however, the results are also dependent on the individual arrival patterns and holding time distributions.

Public telephone networks are typically analyzed in terms of the average activity during the busiest hour of a day. The use of busy-hour traffic measurements to design and analyze telephone networks represents a compromise between designing for the overall average utilization (which includes virtually unused nighttime hours) and designing for short-duration peaks that may occur by chance or as a result of TV commercial breaks, radio call-in contests, and so on.

Busy-hour traffic measurements indicate that an individual residential telephone is typically in use between 5 and 10% of the busy hour. Thus each telephone represents a traffic load of between 0.05 and 0.10 erlangs. The average holding time is between 3 and 4 min, indicating that a typical telephone is involved in one or two phone calls during the busy hour.

Business telephones usually produce loading patterns different from residential phones. First, a business phone is generally utilized more heavily. Second, the busy hour of business traffic is often different from the busy hour of residential traffic. Figure 12.2 shows a typical hourly variation for both sources of traffic. The trunks of a telephone network are sometimes designed to take advantage of variations in calling patterns from different offices. Toll connecting trunks from residential areas are often busiest during evening hours, and trunks from business areas are obviously busiest during midmorning or midafternoon. Traffic engineering depends not only on overall traffic volume but also on time–volume traffic patterns within the network.

A certain amount of care must be exercised when determining the total traffic load of a system from the loading of individual lines or trunks. For example, since two telephones are involved in each connection, the total load on a switching system is exactly one-half the total of all traffic on the lines connected to the switch. In addition, it may be important to include certain setup and release times into the average holding times of some common equipment. A 10-sec setup time is not particularly significant for a 4-min voice call but can actually dominate the holding time of equipment used for short data messages. Common equipment setup times also become more significant in the presence of voice traffic overloads. A greater percentage of the overall load is represented by call attempts since they increase at a faster rate than completions.

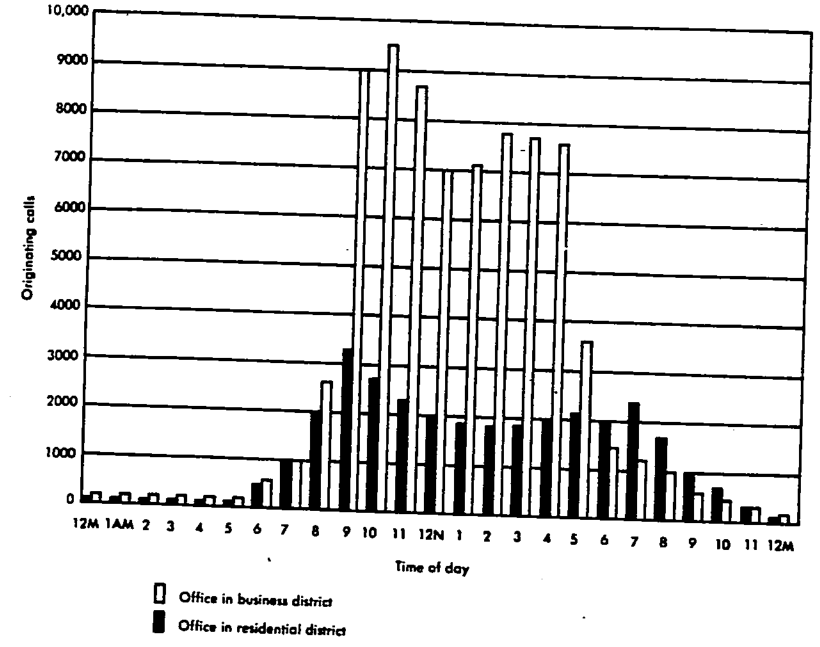

Figure 12.2 Traffic volume dependence on time of day.

An important distinction to be made when discussing traffic in a communications network is the difference between the offered traffic and the carried traffic. The offered traffic is the total traffic that would be carried by a network capable of servicing all requests as they arise. Since economics generally precludes designing a network to immediately carry the maximum offered traffic, a small percentage of offered traffic typically experiences network blocking or delay. When the blocked calls are rejected by the network, the mode of operation is referred to as blocked calls cleared or lost calls cleared. In essence, blocked calls are assumed to disappear and never return. This assumption is most appropriate for trunk groups with alternate routes. In this case a blocked call is normally serviced by another trunk group and does not, in fact, return.

The carried traffic of a loss system is always less than the offered traffic. A delay system, on the other hand, does not reject blocked calls but holds them until the necessary facilities are available. With the assumption that the long-term average of offered traffic is less than the capacity of the network, a delay system carries all offered traffic. If the number of requests that can be waiting for service is limited, however, a delay system also takes on properties of a loss system. For example, if the queue for holding blocked arrivals is finite, requests arriving when the queue is full are cleared.

12.1.1 Arrival Distributions

The most fundamental assumption of classical traffic analysis is that call arrivals are independent. That is, an arrival from one source is unrelated to an arrival from any other source. Even though this assumption may be invalid in some instances, it has general usefulness for most applications. In those cases where call arrivals tend to be correlated, useful results can still be obtained by modifying a random arrival analysis. In this manner the random arrival assumption provides a mathematical formulation that can be adjusted to produce approximate solutions to problems that are otherwise mathematically intractable.

Negative Exponential Interarrival Times

Designate the average call arrival rate from a large group of independent sources (subscriber lines) as λ. Use the following assumptions:

1. Only one arrival can occur in any sufficiently small interval.
2. The probability of an arrival in any sufficiently small interval is directly proportional to the length of the interval. (The probability of an arrival is $\lambda\,\Delta t$, where Δt is the interval length.)
3. The probability of an arrival in any particular interval is independent of what has occurred in other intervals.

It is straightforward [1] to show that the probability distribution of interarrival times is

$$P_0(\lambda t) = e^{-\lambda t} \qquad (12.2)$$

Equation 12.2 defines the probability that no arrivals occur in a randomly selected interval t. This is identical to the probability that t seconds elapse from one arrival to the next.

Example 12.1. Assuming each of 10,000 subscriber lines originate one call per hour, how often do two calls arrive with less than 0.01 sec between them?

Solution. The average arrival rate is

$$\lambda = 3600/10{,}000 = 2.78 \text{ arrivals/sec}$$

From Equation 12.2, the probability of no arrival in a 0.01-sec interval is

$$P_0(0.0278) = e^{-0.0278} = 0.973$$

Thus 2.7% of the arrivals occur within 0.01 sec of the previous arrival. Since the arrival rate is 2.78 arrivals per second, the rate of occurrence of interarrival times less than 0.01 sec is

$$2.78 \times 0.027 = 0.075 \text{ times/sec}$$

The first two assumptions made in deriving the negative exponential arrival distribution can be intuitively justified for most applications. The third assumption, however, implies certain aspects of the sources that cannot always be supported. First, certain events, such as television commercial breaks, might stimulate the sources to place their calls at nearly the same time. In this case the negative exponential distribution may still hold but for a much higher calling rate during the commercial.

A more subtle implication of the independent arrival assumption involves the number of sources, not just their calling patterns. When the probability of an arrival in any small time interval is independent of other arrivals, it implies that the number of sources available to generate requests is constant. If a number of arrivals occur immediately before any subinterval in question, some of the sources become busy and cannot generate requests. The effect of busy sources is to reduce the average arrival rate. Thus the interarrival times are always somewhat larger than what Equation 12.2 predicts them to be. The only time the arrival rate is truly independent of source activity is when an infinite number of sources exist.

If the number of sources is large and their average activity is relatively low, busy sources do not appreciably reduce the arrival rate. For example, consider an end office that services 10,000 subscribers with 0.1 erlang of activity each. Normally, there are 1000 active links and 9000 subscribers available to generate new arrivals. If the number of active subscribers increases by an unlikely 50% to 1500 active lines, the number of idle subscribers reduces to 8500, a change of only 5.6%. Thus the arrival rate is relatively constant over a wide range of source activity. Whenever the arrival rate is fairly constant for the entire range of normal source activity, an infinite source assumption is justified.

Actually, some effects of finite sources have already been discussed in Chapter 5 when analyzing blocking probabilities of a switch. It is pointed out that Lee graph analyses overestimate the blocking probability because, if some number of interstage links in a group are known to be busy, the remaining links in the group are less likely to be busy. A Jacobaeus analysis produces a more rigorous and accurate solution to the blocking probability, particularly when space expansion is used. Accurate analyses of interarrival times for finite sources are also possible. These are included in the blocking analyses to follow.

Poisson Arrival Distribution

Equation 12.2 merely provides a means of determining the distribution of interarrival times. It does not, by itself, provide the generally more desirable information of how many arrivals can be expected to occur in some arbitrary time interval. Using the same assumptions presented, however, the probability of j arrivals in an interval t can be determined [1] as

$$P_j(\lambda t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \qquad (12.3)$$

Equation 12.3 is the well-known Poisson probability law. Notice that when $j = 0$, the probability of no arrivals in an interval t is $P_0(t)$, as obtained in Equation 12.2.

Again, Equation 12.3 assumes arrivals are independent and occur at a given average rate λ, irrespective of the number of arrivals occurring just prior to an interval in question. Thus the Poisson probability distribution should only be used for arrivals from a large number of independent sources.

Equation 12.3 defines the probability of experiencing exactly j arrivals in t seconds. Usually there is more interest in determining the probability of j or more arrivals in t seconds:

$$\begin{aligned} P_{\geq j}(\lambda t) &= \sum_{i=j}^{\infty} P_i(\lambda t) \\ &= 1 - \sum_{i=0}^{j-1} P_i(\lambda t) \\ &= 1 - P_{<j}(\lambda t) \qquad (12.4) \end{aligned}$$

where $P_i(\lambda t)$ is defined in Equation 12.3.

Example 12.2. Given a message-switching node that normally experiences four arrivals per minute, what is the probability that eight or more arrivals occur in an arbitrarily chosen 30-sec interval?

Solution. The average number of arrivals in a 30-sec interval is

$$\lambda t = 4 \times \frac{30}{60} = 2$$

The probability of eight or more arrivals (when the average is 2) is

$$\begin{aligned} P_{\geq 8}(2) &= \sum_{i=8}^{\infty} P_i(2) \\ &= 1 - \sum_{i=0}^{7} P_i(2) \\ &= 1 - e^{-2}\left(1 + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \ldots + \frac{2^7}{7!}\right) \\ &= 0.0011 \end{aligned}$$

Example 12.3. What is the probability that a 1000-bit data block experiences exactly four errors while being transmitted over a transmission link with a bit error rate (BER) of 10^{-5}?

Solution. Assuming independent errors (a questionable assumption on many transmission links), we can obtain the probability of exactly four errors directly from the Poisson distribution. The average number of errors (arrivals) $\lambda t = 10^3 \times 10^{-5} = 0.01$. Thus

$$\text{prob(4errors)} = P_4(0.01) = \frac{(0.01)^4}{4!} e^{-0.01} = 4.125 \times 10^{-10}$$

An alternative solution can be obtained from the binomial probability law:

$$\begin{aligned} \text{prob(4errors)} &= \binom{1000}{4} p^4 (1-p)^{996} \\ &= 4.101 \times 10^{-10} \qquad \text{where } p = 10^{-5} \end{aligned}$$

As can be seen, the two solutions of Example 12.3 are nearly identical. The closeness of the two answers reflects the fact that the Poisson probability distribution is often derived as a limiting case of a binomial probability distribution. Because it is easier to calculate, a Poisson distribution is often used as an approximation to a binomial distribution.

12.1.2 Holding Time Distributions

The second factor of traffic intensity as specified in Equation 12.1 is the average holding time t_m. In some cases the average of the holding times is all that needs to be known about holding times to determine blocking probabilities in a loss system or delays in a delay system. In other cases it is necessary to know the probability distribution of the holding times to obtain the desired results. This section describes the two most commonly assumed holding time distributions: constant holding times and exponential holding times.

Constant Holding Times

Although constant holding times cannot be assumed for conventional voice conversations, it is a reasonable assumption for such activities as per-call call processing requirements, interoffice address signaling, operator assistance, and recorded message playback. Furthermore, constant holding times are obviously valid for transmission times in fixed-length packet networks.

When constant holding time messages are in effect, it is straightforward to use Equation 12.3 to determine the probability distribution of active channels. Assume, for the time being, that all requests are serviced. Then the probability of j channels being busy at any particular time is merely the probability that j arrivals occurred in the time interval of length t_m immediately preceding the instant in question. Since the average number of active circuits over all time is the traffic intensity $A = \lambda t_m$, the probability of j circuits being busy is dependent only on the traffic intensity:

$$P_j(\lambda t_m) = P_j(A)$$

$$= \frac{A^j}{j!}e^{-A} \qquad (12.5)$$

where $\lambda =$ arrival rate
$t_m =$ constant holding time
$A =$ traffic intensity (erlangs)

Exponential Holding Times

The most commonly assumed holding time distribution for conventional telephone conversations is the exponential holding time distribution:

$$P(>t) = e^{-t/t_m} \qquad (12.6)$$

where t_m is the average holding time. Equation 12.6 specifies the probability that a holding time exceeds the value t. This relationship can be derived from a few simple assumptions concerning the nature of the call termination process. Its basic justification, however, lies in the fact that observations of actual voice conversations exhibit a remarkably close correspondence to an exponential distribution.

The exponential distribution possesses the curious property that the probability of a termination is independent of how long a call has been in progress. That is, no matter how long a call has been in existence, the probability of it lasting another t seconds is defined by Equation 12.6. In this sense exponential holding times represent the most random process possible. Not even knowledge of how long a call has been in progress provides any information as to when the call will terminate.

Combining a Poisson arrival process with an exponential holding time process to obtain the probability distribution of active circuits is more complicated than it was for constant holding times because calls can last indefinitely. The final result, however, proves to be dependent on only the average holding time. Thus Equation 12.5 is valid for exponential holding times as well as for constant holding times (or any holding time distribution). Equation 12.5 is therefore repeated for emphasis: The probability of j circuits being busy at any particular instant, assuming a Poisson arrival process and that all requests are serviced immediately, is

$$P_j(A) = \frac{A^j}{j!}e^{-A} \qquad (12.7)$$

where A is the traffic intensity in erlangs. This result is true for any distribution of holding times.

Example 12.4. Assume that a trunk group has enough channels to immediately carry all of the traffic offered to it by a Poisson process with an arrival rate of one call per minute. Assume that the average holding time is 2 min. What percentage of the total traffic is carried by the first five circuits, and how much traffic is carried by all remaining circuits? (Assume that the traffic is always packed into the lowest numbered circuits.)

Solution. The traffic intensity (offered load) of the system is $A = 1 \times 2 = 2$ erlangs. The traffic intensity carried by i active circuits is exactly i erlangs.

Hence the traffic carried by the first five circuits can be determined as follows:

$$A_5 = 1P_1(2) + 2P_2(2) + 3P_3(2) + 4P_4(2) + 5P_5(2)$$

$$= e^{-2}\left(2 + \frac{2\times 2^2}{2!} + \frac{3\times 2^3}{3!} + \frac{4\times 2^4}{4!} + \frac{5\times 2^5}{5!}\right)$$

$$= 1.89 \text{ erlangs}$$

All of the remaining circuits carry

$$2 - 1.89 = 0.11 \text{ erlang}$$

The result of Example 12.4 demonstrates the principle of diminishing returns as the capacity of a system is increased to carry greater and greater percentages of the offered traffic. The first five circuits in Example 12.4 carry 94.5% of the traffic while all remaining circuits carry only 5.5% of the traffic. If there are 100 sources, 95 extra circuits are needed to carry the 5.5%.

12.2 LOSS SYSTEMS

Example 12.4 provides an indication of the blocking probabilities that arise when the number of servers (circuits) is less than the maximum possible traffic load (number of sources). The example demonstrates that 94.5% of the traffic is carried by only five circuits. The implication is that the blocking probability, if only five circuits are available to carry the traffic, is 5.5%. Actually, Example 12.4 is carefully worded to indicate that all of the offered traffic is carried but that only the traffic carried by the first five circuits is of interest. There is a subtle but important distinction between the probability that six or more circuits are busy (as can be obtained from Equation 12.7) and the blocking probability that arises when only five circuits exist.

The basic reason for the discrepancy is indicated in Figure 12.3, which depicts the same traffic pattern arising from 20 sources as is shown previously in Figure 12.1. Figure 12.3, however, assumes that only 13 circuits are available to carry the traffic. Thus the three arrivals at $t = 2.2$, 2.3, and 2.4 min are blocked and assumed to have left the system. The total amount of traffic volume lost is indicated by the shaded area, which is the difference between all traffic being serviced as it arrives and traffic being carried by a blocked calls cleared system with 13 circuits. The most important feature to notice in Figure 12.3 is that the call arriving at $t = 2.8$ is not blocked, even though the original profile indicates that it arrives when all 13 circuits are busy. The reason it is not blocked is that the previously blocked calls left the system and therefore reduced the congestion for subsequent arrivals. Hence the percentage of time that the original traffic profile is at or above 13 is not the same as the blocking probability when only 13 circuits are available.

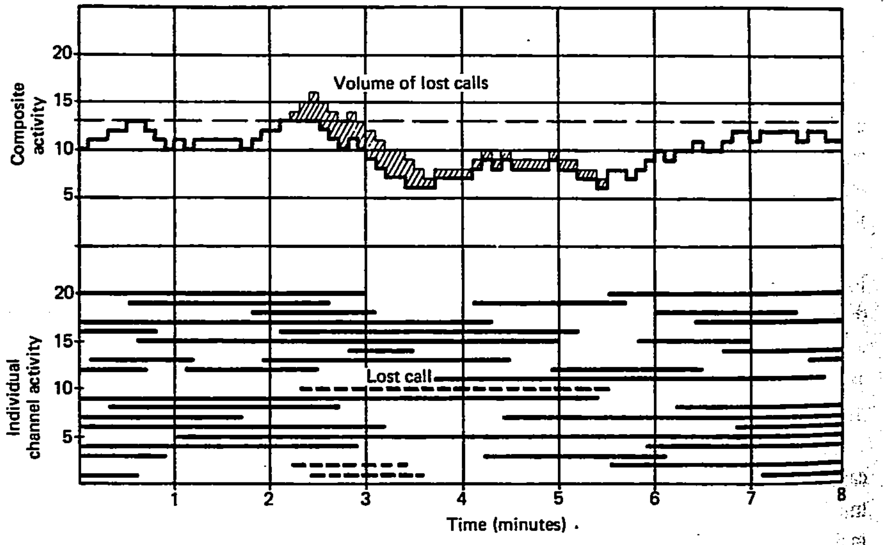

Figure 12.3 Activity profile of blocked calls cleared (13 channels).

12.2.1 Lost Calls Cleared

The first person to account fully and accurately for the effect of cleared calls in the calculation of blocking probabilities was A. K. Erlang in 1917. In this section we discuss Erlang's most often used result: his formulation of the blocking probability for a lost calls cleared system with Poisson arrivals. Recall that the Poisson arrival assumption implies infinite sources. This result is variously referred to as Erlang's formula of the first kind, $E_{1,N}(A)$; the Erlang-B formula; or Erlang's loss formula.

A fundamental aspect of Erlang's formulation, and a key contribution to modern stochastic process theory, is the concept of statistical equilibrium. Basically, statistical equilibrium implies that the probability of a system's being in a particular state (number of busy circuits in a trunk group) is independent of the time at which the system is examined. For a system to be in statistical equilibrium, a long time must pass (several average holding times) from when the system is in a known state until it is again examined. For example, when a trunk group first begins to accept traffic, it has no busy circuits. For a short time thereafter, the system is most likely to have only a few busy circuits. As time passes, however, the system reaches equilibrium. At this point the most likely state of the system is to have $A = \lambda t_m$ busy circuits.

When in equilibrium, a system is as likely to have an arrival as it is to have a termination. If the number of active circuits happens to increase above the average A, departures become more likely than arrivals. Similarly, if the number of active circuits happens to drop below A, an arrival is more likely than a departure. Thus if a system is perturbed by chance from its average state, it tends to return.

Although Erlang's elegant formulation is not particularly complicated, it is not presented here because we are mostly interested in application of the results. The interested reader is invited to see reference [2] or [3] for a derivation of the result:

$$B = E_{1,N}(A) = \frac{A^N}{N! \sum_{i=0}^{N} (A^i/i!)} \qquad (12.8)$$

where N = number of servers (channels)

A = offered traffic intensity, λt_m (erlangs)

Equation 12.8 specifies the probability of blocking for a system with random arrivals from an infinite source and arbitrary holding time distributions. The blocking probability of Equation 12.8 is plotted in Figure 12.4 as a function of offered traffic intensity for various numbers of channels. An often more useful presentation of Erlang's results is provided in Figure 12.5, which presents the output channel utilization for various blocking probabilities and numbers of servers. The output utilization ρ represents the traffic carried by each circuit:

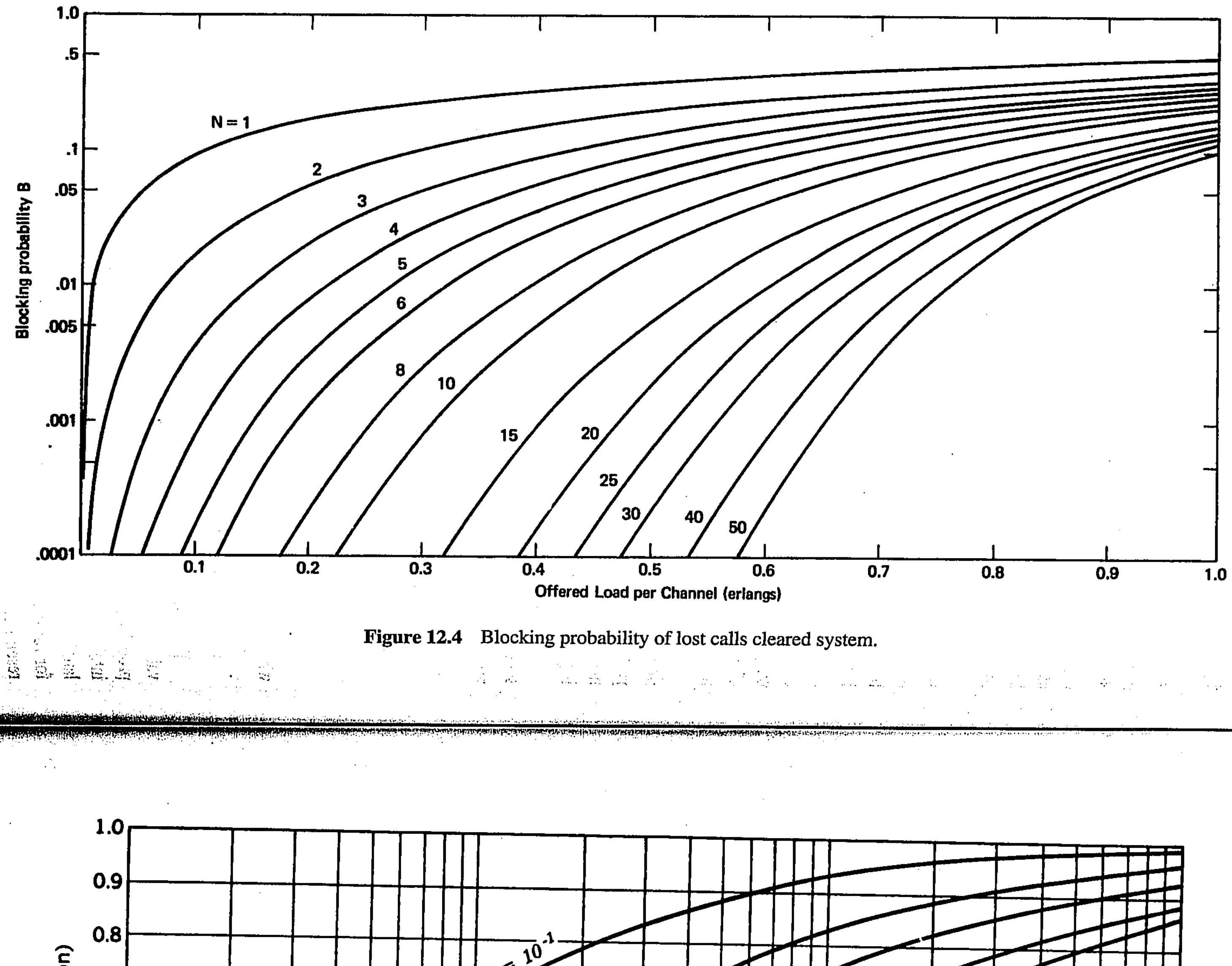

Figure 12.4 Blocking probability of lost calls cleared system.

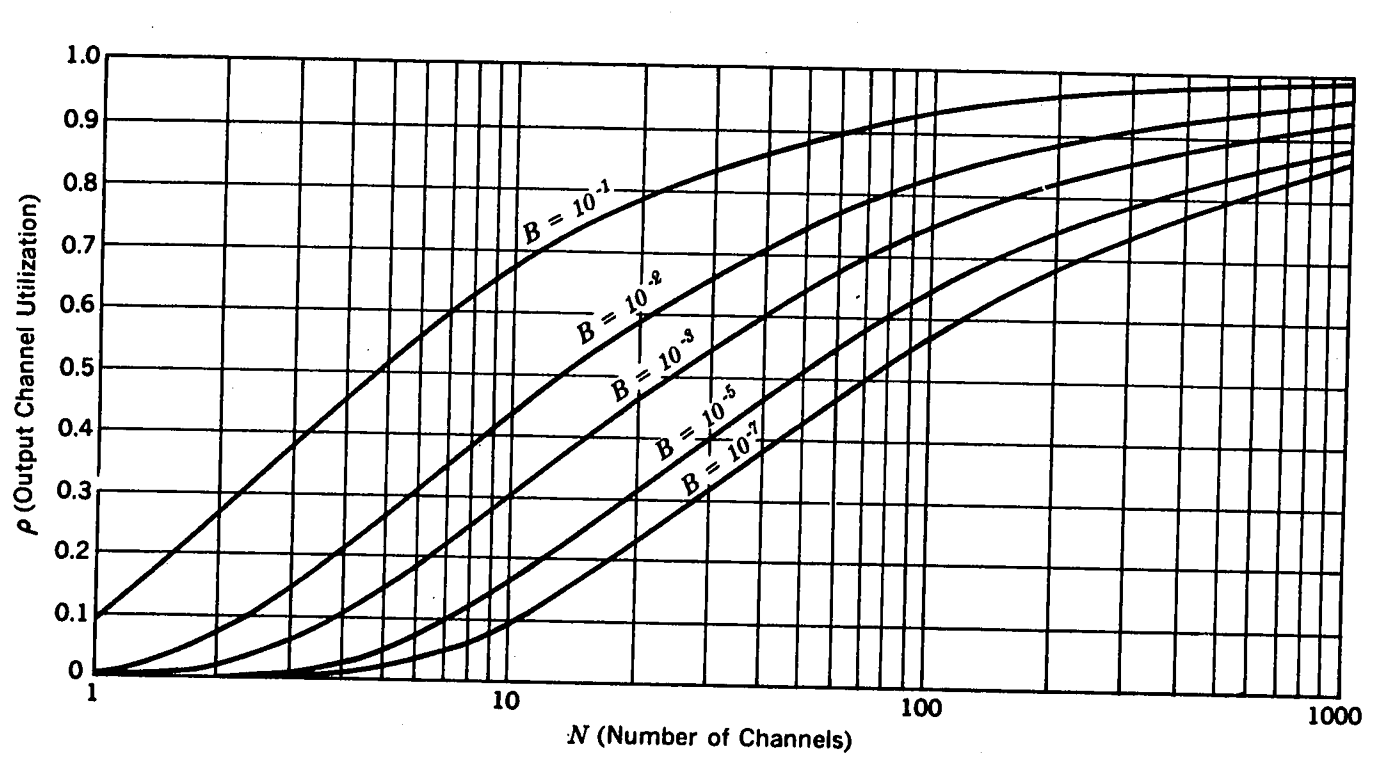

Figure 12.5 Output channel utilization of lost calls cleared system.

$$\rho = \frac{(1-B)A}{N} \tag{12.9}$$

where A = offered traffic
N = number of channels
B = blocking probability
$(1-B)A$ = carried traffic

Blocking probabilities are also provided in tabular form in Appendix D.

Example 12.5. A T1 line is to be used as a tie-line trunk group between two PBXs. How much traffic can the trunk group carry if the blocking probability is to be 0.1? What is the offered traffic intensity?

Solution. From Figure 12.5 it can be seen that the output circuit utilization for $B = 0.1$ and $N = 24$ is 0.8. Thus the carried traffic intensity is $0.8 \times 24 = 19.2$ erlangs. Since the blocking probability is 0.1, the maximum level of offered traffic is

$$A = \frac{19.2}{1-0.1} = 21.3 \text{ erlangs}$$

Example 12.6. Four clusters of data terminals are to be connected to a computer by way of leased circuits, as shown in Figure 12.6. In Figure 12.6*a* the traffic from the clusters uses separate groups of shared circuits. In Figure 12.6*b* the traffic from all clusters is concentrated onto one common group of circuits. Determine the total number of circuits required in both cases when the maximum desired blocking probability is 5%. Assume that 22 terminals are in each cluster and each terminal is active 10% of the time. (Use a blocked calls cleared analysis.)

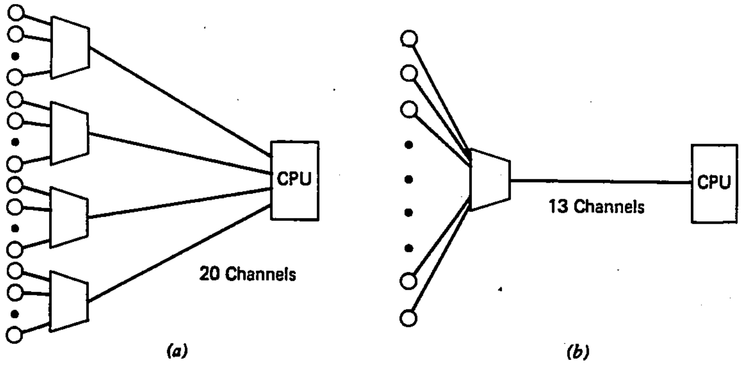

Figure 12.6 Data terminal network of Example 10.6: (*a*) four separate groups; (*b*) all traffic concentrated into one group.

Solution. The offered traffic from each cluster is $22 \times 0.1 = 2.2$ erlangs. Since the average number of active circuits is much smaller than the number of sources, an infinite source analysis can be used. Using Table D.1, the number of circuits required for B = 5% at a loading of 2.2 erlangs is 5. Thus the configuration of Figure 12.6*a* requires a total of 20 circuits.

The total offered traffic to the concentrator of the configuration of Figure 12.6*b* is $4 \times 2.2 = 8.8$ erlangs. From Table D.1, 13 circuits are required to support the given traffic load.

Example 12.6 demonstrates that consolidation of small traffic groups into one large traffic group can provide significant savings in total circuit requirements. Large groups are more efficient than multiple small groups because it is unlikely that the small groups will become overloaded at the same time (assuming independent arrivals). In effect, excess traffic in one group can use idle circuits in another group. Thus those circuits that are needed to accommodate traffic peaks but are normally idle are utilized more efficiently when the traffic is combined into one group. This feature is one of the motivations mentioned in Chapter 10 for integrating voice and data traffic into a common network. The total savings in transmission costs is most significant when the individual traffic intensities are low. Hence it is the peripheral area of a network that benefits the most by concentrating the traffic.

The greater circuit efficiency obtained by combining traffic into large groups is often referred to as the advantage of large group sizes. This efficiency of circuit utilization is the basic motivation for hierarchical switching structures. Instead of interconnecting a large number of nodes with rather small trunk groups between each pair, it is more economical to combine all traffic from individual nodes into one large trunk group and route the traffic through a tandem switching node. Figure 12.7 contrasts a mesh versus a star network with a centralized switching node at the center. Obviously, the cost of the tandem switch becomes justified when the savings in total circuit miles is large enough.

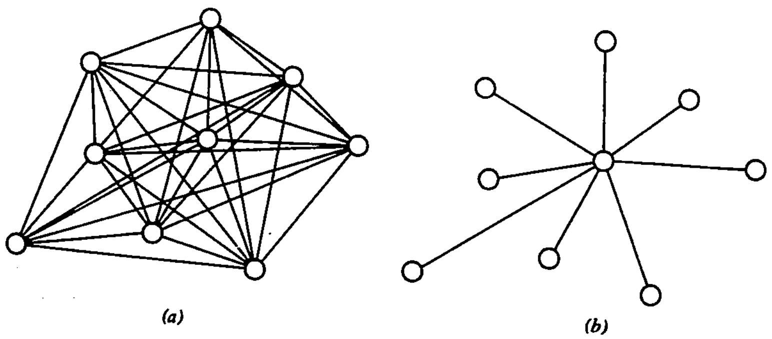

Figure 12.7 Use of tandem switching to concentrate traffic: (*a*) mesh; (*b*) star.

Example 12.7. What happens to the blocking probabilities in Figure 12.6*a* and *b* discussed in Example 12.6 when the traffic intensity increases by 50%?

Solution. If the traffic intensity of each group increases from 2.2 to 3.3 erlangs, the blocking probability of the configuration of Figure 12.6*a* increases from 5% to almost 14%.

In the configuration of Figure 12.6*b* a 50% increase in the traffic intensity causes a 400% increase in the blocking probability (from 5 to 20%).

Example 12.7 demonstrates some important considerations in network design. As indicated, blocking probabilities are very sensitive to increases in traffic intensities, particularly when the channels are heavily utilized. Because large trunk groups utilize their channels more efficiently, they are more vulnerable to traffic increases than are a number of smaller groups designed to provide the same grade of service. Furthermore, failures of equal percentages of transmission capacity affect the performance of a large group more than the performance of several small groups. In both cases the vulnerability of the large groups arises because large groups operate with less spare capacity than do multiple small groups.

A second aspect of blocking analyses demonstrated in Example 12.7 is that the calculated results are highly dependent on the accuracy of the traffic intensities. Accurate values of traffic intensities are not always available. Furthermore, even when accurate traffic measurements are obtainable, they do not provide an absolute indication of how much growth to expect. Thus only limited confidence can be attached to calculations of blocking probabilities in an absolute sense. The main value of these analyses is that they provide an objective means of comparing various network sizes and configurations. The most cost-effective design for a given grade of service is the one that should be chosen, even if the traffic statistics are hypothetical. If a network is liable to experience wildly varying traffic patterns or rapid growth, these factors must be considered when comparing design alternatives. A network with a somewhat larger initial cost may be more desirable if it can absorb or grow to accommodate unanticipated traffic volumes more easily.

12.2.2 Lost Calls Returning

In the lost calls cleared analyses just presented, it is assumed that unserviceable requests leave the system and never return. As mentioned, this assumption is most appropriate for trunk groups whose blocked requests overflow to another route and are usually serviced elsewhere. However, lost calls cleared analyses are also used in instances where blocked calls do not get serviced elsewhere. In many of these cases, blocked calls tend to return to the system in the form of retries. Some examples are subscriber concentrator systems, corporate tie lines and PBX trunks, calls to busy telephone numbers, and access to WATS lines (if DDD alternatives are not used). This section derives blocking probability relationships for lost calls cleared systems with random retries.

The following analysis involves three fundamental assumptions regarding the nature of the returning calls:

1. All blocked calls return to the system and eventually get serviced, even if multiple retries are required.
2. The elapsed times between call blocking occurrences and the generation of retries are random and statistically independent of each other. (This assumption allows the analysis to avoid complications arising when retries are correlated to each other and tend to cause recurring traffic peaks at a particular waiting time interval.)
3. The typical waiting time before retries occur is somewhat longer than the average holding time of a connection. This assumption essentially states that the system is allowed to reach statistical equilibrium before a retry occurs. Obviously, if retries occur too soon, they are very likely to encounter congestion since the system has not had a chance to "relax." In the limit, if all retries are immediate and continuous, the network operation becomes similar to a delay system discussed in later sections of this chapter. In this case, however, the system does not queue requests—the sources do so by continually "redialing."

When considered in their entirety, these assumptions characterize retries as being statistically indistinguishable from first-attempt traffic.* Hence blocked calls merely add to the first-attempt call arrival rate.

Consider a system with a first-attempt call arrival rate of λ. If a percentage B of the calls is blocked, B times λ retries will occur in the future. Of these retries, however, a percentage B will be blocked again. Continuing in this manner, the total arrival rate λ' after the system has reached statistical equilibrium can be determined as the infinite series

$$\lambda' = \lambda + B\lambda + B^2\lambda + B^3\lambda + \dots$$

$$= \frac{\lambda}{1-B} \tag{12.10}$$

where B is the blocking probability from a lost calls cleared analysis with traffic intensity $A' = \lambda' t_m$.

Equation 12.10 relates the average arrival rate λ', including the retries, to the first-attempt arrival rate and the blocking probability in terms of λ'. Thus this relationship does not provide a direct means of determining λ' or B since each is expressed in terms of the other. However, the desired result can be obtained by iterating the lost calls

*First-attempt traffic is also referred to as demand traffic: the service demands assuming all arrivals are serviced immediately. The offered traffic is the demand traffic plus the retries.