translation records which are also recorded in the permanent memory. These records give, for each directory number, the network terminal at which it appears and all class of service information.

Special Services

The application of electronic techniques to telephone switching make possible new services; services which have been impractical heretofore due to either economic or technical reasons. One example of what could be done is abbreviated dialing where one dials a preliminary "one" followed by a single digit. Each of the ten possible digits represents a preselected, frequently called, telephone number. You might choose to have a 1 represent a nearby friend; 2, your office; 3, a relative living thousands of miles away, etc. The system would recognize the type of call and, using its photographic plates, translate your 1X into the actual telephone number, whether it be a local call or not. The translations would, of course, have been previously been recorded in the system. This and many other services are made technically feasible by the use of a stored program and the use of electronic memory.

Conclusion

The introduction of electronic techniques into telephone switching represents a major change in the art. Both the new types of devices and new types of telephone system organization offer important advantages. Perhaps the most important result will be the increased flexibility and new services that this will make possible.

Traffic Aspects of Communications Switching Systems

JOSEPH A. BADER†

TO DESIGN a communications switching system which provides a satisfactory grade of service at minimum cost, an understanding of the nature of the offered traffic is necessary.

For many years, telephone traffic engineers have been studying the problem of providing sufficient equipment to meet time-varying demands at a given level of service. The experience gained in these studies and the analytical results derived may prove valuable in the planning and use of modern digital data processing equipment for real-time applications.

Major Components of Traffic Variation

In order to design a switching system to meet a specified grade of service, an estimate of the average traffic offered to the switching system must be made. The reliability of this estimate depends on the magnitude of the components of variation present.

Variations such as seasonal, day-to-day, and hour-to-hour are not easily susceptible to an analytic approach. They are usually determined empirically for each exchange or area. Typical seasonal, daily, and hourly traffic patterns are shown in Figs. 1 to 3. These patterns are usually stable so that in choosing the average busy season, busy day, busy hour traffic as a base for engineering, we can be reasonably assured that the busy hour traffic offered during the rest of the year does not greatly exceed our engineered estimate. Of course, events occur such as earthquakes, snow storms, disasters, etc., which provide a common cause for call origination. This results in traffic "peaking" well above the engineered level with a resulting degradation of service. Under such extreme conditions, special overload-control procedures are usually initiated which tend to spread the peaked demand over a longer period of time.

The remaining component of variation is the instantaneous variation of the number of calls offered per unit time during the busy hour. Fig. 4 shows the variation in the number of calls offered per 24-second interval for a 10-minute measurement period. From these data, a frequency distribution giving the fraction of 24-second intervals containing x call originations was constructed (Fig. 5). If calls arrive at random, then the distribution per unit time should be Poissonian. To test the data for randomness of call arrivals, a graph of the Poisson and the sample distribution was constructed (Fig. 5). By inspection, the agreement appears to be quite close. It is evidence of this kind which lends assurance to the assumption of random-call input which is basic to most traffic theory. Accordingly, we can compute the magnitude of the instantaneous variation of offered traffic as the first step in determining the engineered capacity of the system.

Holding-Time Variations

The next step is to investigate the length of time required to serve the offered calls. The service time, generally called the holding time, ranges from a fraction of a second for certain switching equipments to several minutes

† Bell Telephone Labs., Inc., New York, N.Y.

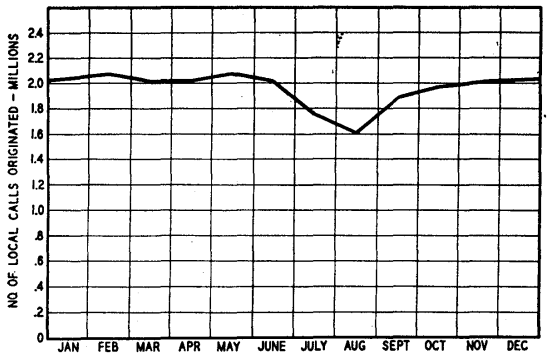

Fig. 1—Seasonal variation in daily local calls in Boston, 1934.

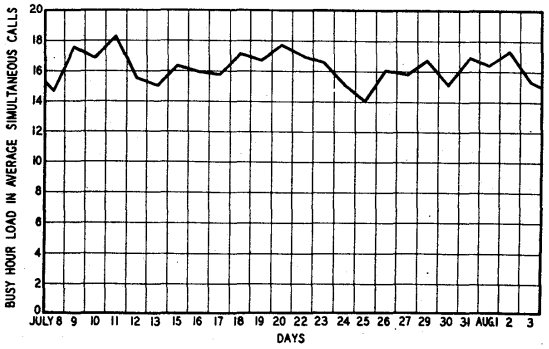

Fig. 2—Day-to-day busy hour variation in load, Newark, 1918.

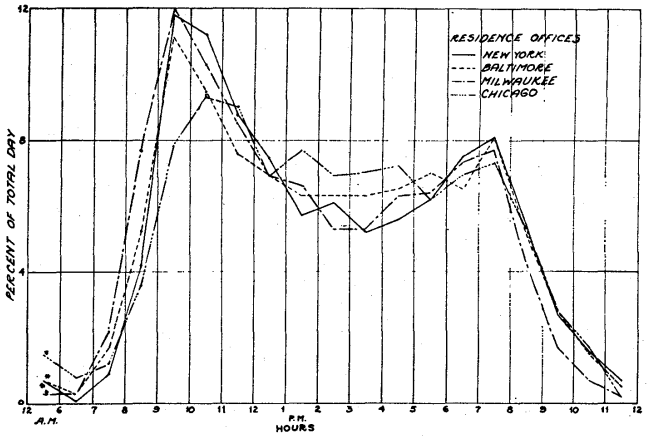

Fig. 3—Hourly traffic distribution. (*) values shown are per cents of total traffic handled in 12 MIDNIGHT-6 A.M. period.

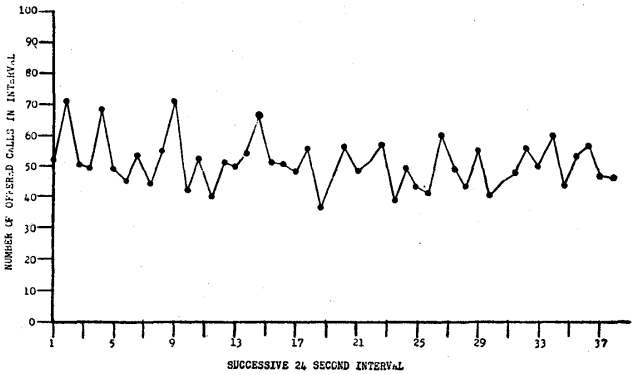

Fig. 4—Variation in number of calls offered per 24-second interval, Asbury Park, 1957.

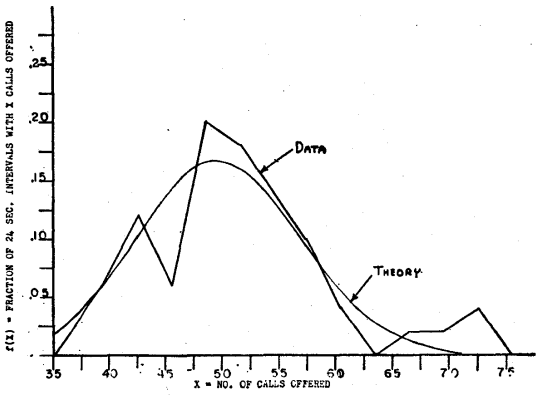

Fig. 5—Comparison of Poisson theory to measured number of calls offered.

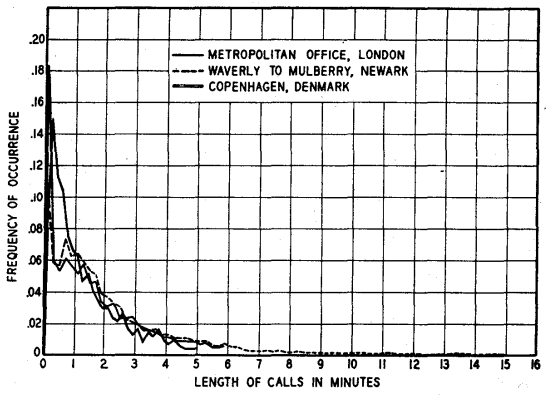

Fig. 6.

for talking paths. To expedite theoretical treatment, one of the following two assumptions is usually made with regard to call lengths or holding times: 1) holding times are distributed according to the exponential distribution, or 2) holding times are constant. The assumption of exponentially varying holding times is very well confirmed where local-call conversation and control time are examined. Fig. 6 shows the distribution of such times as measured in Newark, N.J., in a past traffic study. The agreement between data and theory is obviously good. Based on evidence of this kind, an exponentially varying holding time is assumed for equipment held during the entire conversation.

Constant holding time is exhibited by equipments whose function is to perform rapid switching operations prior to the beginning of conversations. For example, in one type of telephone switching system, a so-called marker performs the function of connecting the subscriber's line to an outgoing trunk. The marker then releases and is immediately available to another subscriber. The time required for this function is essentially constant.

Traffic Usage

Having determined the number of offered calls and the server's holding time, we can define a third parameter called traffic load, or usage. This is the product of the number of calls and the average length of each call. Thus, if a system is offered 1000 calls in the busy hour, each of average length 100 seconds, the offered load is 100,000 call seconds. In dealing with a system which is offered a large number of calls per hour, it is more convenient to deal with call hours per hour. In our example, then, the offered load would be 100,000/3600 or 27.8 call hours per hour. The traffic unit of one call hour per hour has been named the erlang in honor of A. K. Erlang, Danish mathematician, who pioneered in traffic theory. The offered load expressed in erlangs represents the average number of simultaneous calls that would be in progress if sufficient servers were always available. We can see this from Fig. 7 which shows the variation in the number of calls present on a trunk group between two central offices during the busy hour. During this hour, 246 calls were carried. Greatest number of calls in progress was 16 and this occurred four times during the hour. Average number of simultaneous calls in progress was 9.5. This means that on the average during this hour, the trunk group carried 9.5 call hours.

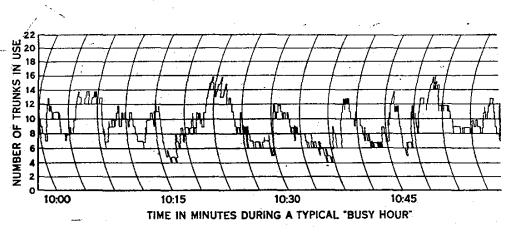

Fig. 7.

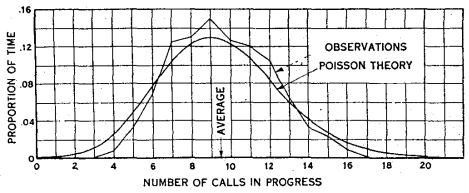

Fig. 8.—A comparison of theory with the observations in the hour above.

As indicated previously, the Poisson distribution correctly describes the variation in the number of random-call arrivals per unit time. Under certain assumption, however, the Poisson also describes correctly the probability of finding at a random instant a given number of calls on a group of servers. Fig. 8 shows the Poisson with average 9.5 and an empirical distribution of the number of calls in existence each 30-second interval taken from the data of Fig 7. In this case, the Poisson closely predicts the proportion of time x calls that are present; this is also clearly the fraction of calls arriving and finding x calls ahead of them in the system. If x exceeds the number of servers provided, newly arriving calls will fail in obtaining immediate service. The proportion of such failures is a commonly used criterion for the adequacy of service. Modification of the assumptions underlying the Poisson gives rise to other formulations. In particular, the assumption regarding the behavior of calls failing to find an idle server immediately is of primary interest.

Defection Ratio

The behavior of calls which fail to find an idle server immediately can be expressed in terms of the defection ratio j. This is the ratio of the rate at which waiting calls abandon before being served, to the rate at which they are served. j, of course, can assume values from 0 to infinity. However, in Telephone Traffic Engineering it is usual to find only three different values of j assumed. These values and their physical interpretation are as follows. 1) $j = 0$ corresponds to the case in which unserved calls wait indefinitely for service. In telephone traffic parlance, this is called the "lost calls delayed" assumption. 2) $j = 1$ corresponds to the case in which calls wait no longer than their holding time and then abandon. If an idle server becomes available, a call seizes the server and uses it for the remaining part of its holding time. This is the "lost calls held" assumption. 3) $j =$ infinity corresponds to the case in which calls are not willing to wait at all for servers and immediately abandon. This is the "lost calls cleared" assumption.

Loss Engineering

A group of servers engineered solely on the basis of "expected proportion of calls which fail to receive immediate service" is said to be engineered on a "loss" basis, and formulas used to predict the proportion of calls failing to find an idle server immediately are called "loss" formulas. Loss formulas and their underlying assumptions are listed in Fig. 9. The assumption of "infinite sources" is, of course, never quite realized in practice but where the rate of arrival of calls is nearly independent of the number of calls momentarily being served, this assumption can be used with confidence. The list in Fig. 9 is by no means complete. However, the formulas tabulated are those most widely used for engineering telephone switching equipment. Graphs of these loss formulas are shown in Figs. 10 to 12. By means of these curves, the traffic engineer is able to solve a wide range of "loss-engineering" problems. The following examples demonstrate the use of these curves.

Example 1

How many trunks should be provided at P.01 service if the load offered is 10 erlangs?

Solution: Past experience indicates that for trunk groups with no provision made to reroute overflow calls,

c = Number of Full Access Trunks

a = Load Submitted in Average Simultaneous Calls

$$= \frac{\text{(Number of Calls per Hour)(Average Holding Time in Seconds)}}{3600}$$

j	Lost Calls Assumption	Usual Designation of Formula	Frequency Distributions, $f(x)$		Probability of Delay, P
			When $x \leqslant c$	When $x \geqslant c$	
0	"Delayed"	Erlang "C"	$\dfrac{\frac{a^x e^{-a}}{\underline{/x}}}{1-P(c,a)+\frac{a^c e^{-a}}{\underline{/c}}\cdot\frac{c}{c-a}}$	$\dfrac{\frac{a^x e^{-a}}{\underline{/c}\,c^{x-c}}}{1-P(c,a)+\frac{a^c e^{-a}}{\underline{/c}}\cdot\frac{c}{c-a}}$	$C(c,a)=\dfrac{\frac{a^c e^{-a}}{\underline{/c}}\cdot\frac{c}{c-a}}{1-P(c,a)+\frac{a^c e^{-a}}{\underline{/c}}\cdot\frac{c}{c-a}}$
1	"Held"	Poisson	$\frac{a^x e^{-a}}{\underline{/x}}$	$\frac{a^x e^{-a}}{\underline{/x}}$	$P(c,a)=\sum_{x=c}^{\infty}\frac{a^x e^{-a}}{\underline{/x}}$
∞	"Cleared"	Erlang "B"	$\dfrac{\frac{a^x e^{-a}}{\underline{/x}}}{1-P(c+1,a)}$	0	$B(c,a)=\dfrac{\frac{a^c e^{-a}}{\underline{/c}}}{1-P(c+1,a)}$

Fig. 9—Familiar telephone-traffic formulas assuming infinite sources.

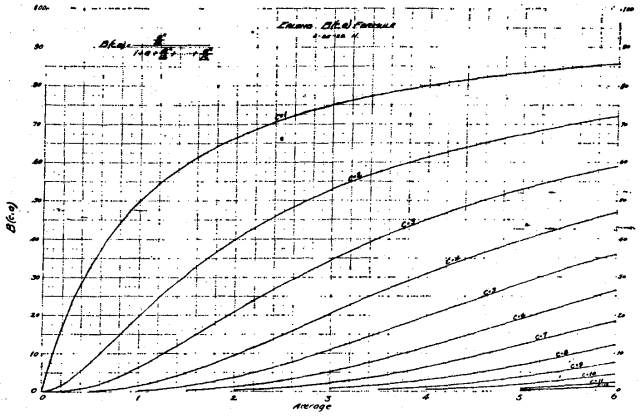

Fig. 10—Erlang "B" load vs loss curves.

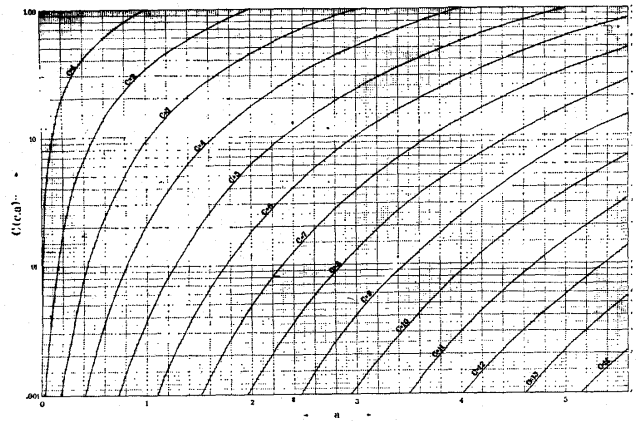

Fig. 11—Erlang "C" load vs loss curves.

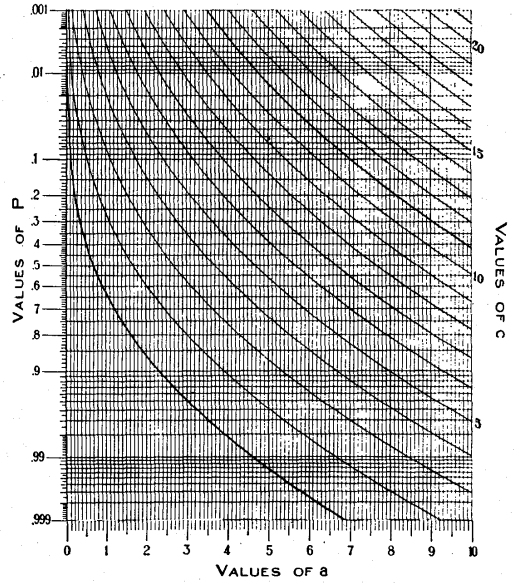

Fig. 12—Poisson load vs loss curves.

the lost calls held assumption applies reasonably well. Therefore, the Poisson loss formula is chosen here. From Fig. 12 we see that 19 trunks are required.

Example 2

Measurement indicates that a group of 10 dial-pulse registers are giving P.03 service. How many registers must be added in order to give P.01 service? Assume holding times are exponential and lost calls are delayed.

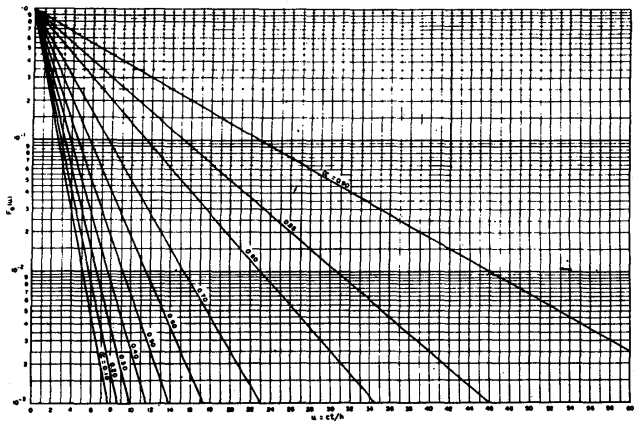

Fig. 13—Probability of delays exceeded with exponential calls served in order of arrival.

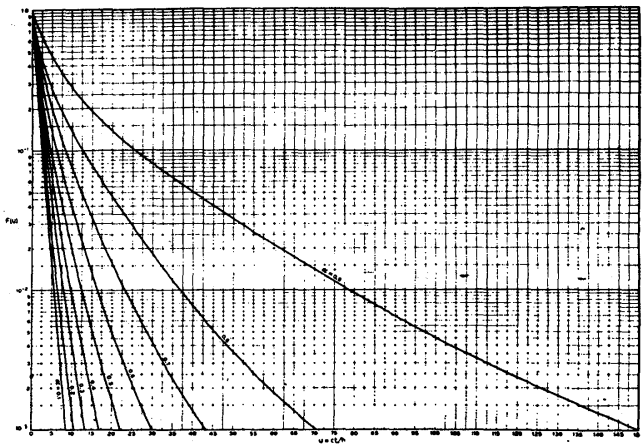

Fig. 14—Probability of delays exceeded with delayed exponential calls served in random order.

Solution: Fig. 11 shows that for 10 registers operating at P.03, the offered load is 4.84 erlangs. In order to give P.01 service, 11.3 registers are needed. Therefore, add two registers.

Delay Engineering

In many instances the traffic engineer is concerned with the length of delay as well as the probability of a delay. (Speed of dial-tone service is a typical example.) Theoretical formulations have been derived under the assumption of lost calls delayed for both exponential and constant holding times. In general, the length of a delay depends on 1) the offered load, 2) the number of servers, 3) the distribution of holding times, and 4) the queue discipline.

Theoretical descriptions of delay have been worked out for three queue disciplines. They are 1) "first come, first served," that is, waiting calls are served in the order in which they arrive; 2) "last come, first served"; and 3) "random service," in which waiting calls are served at random. Since the "last come, first served" queue discipline maximizes delays, it is never realized in telephone switching systems. There are, however, systems in which the other two disciplines are realized. Graphs of delay distributions for the queue disciplines 1) and 2) are shown in Figs. 13 and 14. The graphs show the conditioned distribution of delays. That is, they yield the probability of a delay greater than t *given* that a call is delayed at all. In order to get the unconditioned probability of a delay greater than t, we multiply by the probability of a delay C (c, a) shown in Fig. 11. It should be noted that the average delay is the same for each discipline. From these curves, a traffic engineer can determine the number of equipments required to provide a given grade of delay service. The following examples demonstrate the use of these curves.

Example 3

A certain computer has a steering circuit which directs single, incoming pulses to an idle arithmetic unit. Pulses arrive at random and each pulse is acted on by the arithmetic unit for an average of 5 μsec. Pulses which arrive when all units are busy are held in a storage unit. Waiting pulses are served in the order of their arrival. The holding time of the steering circuit itself may be ignored. What is the capacity of the computer in pulses per hour if five arithmetic units are supplied and no more than 3 per cent of the delayed pulses are to be delayed longer than 10 μsec?

Solution: From Fig. 13, we find the occupancy, that is, the ratio of the offered load to the number of servers, of each arithmetic unit, is alpha = 0.65 so that $a = 3.25$ erlangs, since $a = \frac{N\bar{t}}{3600}$, we have $N = \frac{11700}{5 \times 10^{-6}} = 2.34 \times 10^{9}$ pulses per hour.

Example 4

A single toll booth is provided at the entrance to a bridge. Cars arrive at the toll booth at random at a rate of 0.15 per second. The time required to collect a toll is exponential with an average of 5 seconds; the booth is located such that cars cannot defect from the waiting line. What fraction of the cars are delayed more than 30 seconds in reaching the toll booth?

Solution: The load in erlangs offered to the single toll booth is $a = 0.15 \times 5 = 0.75$ erlangs.

$$\text{The occupancy alpha} = \frac{a}{c} = 0.75.$$

From Fig. 13 we find that $P(>6) = 0.23$ so that 23 per cent of the delayed cars are delayed more than 30 seconds. Multiplying by $C(1.75) = 0.75$ we have 19 per cent of all cars delayed more than 30 seconds.

Example 5

A counter in a department store is manned by three clerks. The time required to serve a customer is distributed exponentially with an average of 3 minutes. Customers receive a number indicating the order of their

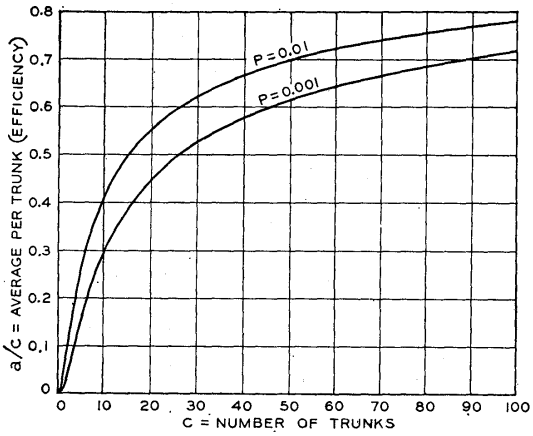

Fig. 15.

arrival so that service in the waiting line is very nearly first come, first served. Observations indicate that during the normal busy hour, 10 per cent of the delayed customers are delayed at least 6 minutes in acquiring the services of a clerk. If one clerk is added, what per cent of the customers would be delayed at least 6 minutes?

Solution: Using $P(>6)$ and Fig. 13, determine the offered load a. Since $\frac{a}{c} = 0.62$, we have $a = 1.86$. If one clerk is added, the new occupancy is alpha $= \frac{a}{c} = \frac{1.86}{4} =$ 0.47. From Fig. 13 we find that 1.40 per cent of the delayed customers would be delayed at least 6 minutes. The addition of one clerk resulted in over a 7 to 1 improvement in the per cent delayed at least 6 minutes.

Overload Performance

So far we have considered the problem of engineering equipment to accommodate an average busy hour load at a given grade of "loss" or delay service. Related to this problem is that of balancing efficiency against overload capacity. A brief discussion of this problem might be of interest.

The efficiency of a trunk group or average load per trunk is defined to be the ratio of the load carried to the number of trunks. Fig. 15 shows the relationship between efficiency and group size at engineered losses of P.01 and P.001. From the curves, it is clear that large groups are more efficient than small groups. On the other hand, the

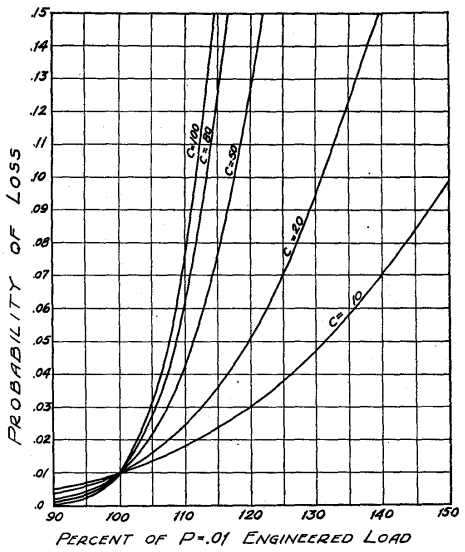

Fig. 16—Relative ability of different sizes of trunk groups to carry overloads.

higher the efficiency, the less margin available percentagewise for small overloads. Fig. 16 shows the relationship between the increase in loss from P.01 against group size when the offered load is over engineered level. The ideal balance between efficiency and overload margin depends on additional factors such as the purpose for which the system is to be used and the environment in which it is to function.

Conclusion

Some of the fundamental traffic aspects of switching systems and some formulas by which probabilities of delays and losses may be calculated have been displayed. Working curves have been shown by which many traffic-engineering problems can be solved. Examples are given which illustrate the application of the curves in practical situations.

Bibliography

[1] Brockmeyer, E., Halstrom, H. L., and Jensen, A. *The Life and Works of A. K. Erlang.* Copenhagen: Copenhagen Telephone Company, 1948.

[2] Molina, E. C. "Application of the Theory of Probability to Telephone Trunking Problems," *Bell System Technical Journal,* Vol. 6 (1927), p. 461.

[3] Fry, T. C. *Probability and Its Engineering Uses.* New York: D. Van Nostrand Company, 1928.

[4] Thorndyke, F. "Applications of Poisson's Probability Summation," *Bell System Technical Journal,* Vol. 5 (October, 1926), p. 604.