Inherent Load-Balancing in Step-By-Step Switching Systems

By M. M. BUCHNER, JR., and S. R. NEAL

(Manuscript received June 19, 1970)

Questions have been raised over the years regarding the validity of the traffic-engineering tables used by the Bell System and others for the graded-multiple trunk groups within the step-by-step switching system. The tables indicate that the capacity of a graded multiple is increased when it is imbedded in the step-by-step system, the increase becoming larger as the number of switches connected to the graded multiple decreases. The increase in capacity supposedly occurs because of a "finite-source effect."

In this paper, we investigate in detail the flow of traffic through the step-by-step system to determine the validity of the tables. By a combination of simulation and analysis, we show that the increase in capacity arises not from a finite-source effect but from an inherent load-balancing that results from the clever manner in which the switches and trunks are interconnected. We conclude that, when used for engineering in the presence of day-to-day variations, the tables are adequate when the gradings are used with the large selector groups. For the minimum number of selectors, the tables estimate fairly accurately the capacity of the larger gradings, slightly overestimate the capacity for the medium-sized gradings, and overestimate the capacity of some of the smaller gradings by about 20 percent.

I. INTRODUCTION

The graded-multiple trunk groups that connect the successive switching stages of the step-by-step switching system* are commonly engineered according to a set of tables furnished by American Telephone and Telegraph Company.[1] Because the tables, which were prepared some years ago, are based upon certain approximations and a good deal of engineering judgment and because the arguments used to

* It is assumed that the reader is familiar with the step-by-step system. For those who are not, a summary of the system is given in Appendix A.

justify the tables are not convincing, their validity has been questioned.[†]

Specifically, the tables indicate that the traffic capacity of a graded multiple is increased when it is imbedded in the step-by-step system and that the increase becomes more significant as the number of switches connected to the grading decreases. The common "explanations" for the increase in capacity usually relate in some imprecise manner to a "finite-source effect."

Several common incorrect arguments[2–4] that relate to the capacity of the graded multiples in the step-by-step system are summarized in Appendix C. It is important to realize the nature of these arguments in order to appreciate the uncertainty that has persisted concerning the validity of the tables and to view the present study in the proper perspective.

To clarify the issues involved, we examine in detail the flow of traffic through the step-by-step system. We demonstrate that the capacity of a graded multiple is increased when it is imbedded in the step-by-step system and that the increase becomes larger as the number of switches connected to the grading decreases. We show that the increase in capacity results from an "inherent load-balancing"[*] caused by the way in which the switches and trunks are interconnected rather than from a finite-source effect. By a combination of simulation and analysis, we conclude that the accuracy of the tables depends upon the size of the grading and the number of selectors.

II. INHERENT LOAD-BALANCING

In this section, we show qualitatively why the capacity of a graded multiple increases when traffic is offered to the graded multiple through the line finders and first selectors.

Two fully equipped line-finder groups and four first-selector half-shelves are pictured in Fig. 1. For clarity, the line-finder to first-selector connections are shown only for one line-finder group. The graded multiple in Fig. 1 is connected to a particular level of the first selectors (we refer to the level as "our level"). The graded multiple is used by calls directed to our level; calls to other levels and, thus, to other destinations use trunk groups that are not shown.

[†] There are indications that part of the uncertainty surrounding the tables is due to an incomplete knowledge of the mathematical models used to generate them. Since this information is not readily available, some of the history of the development of the tables is presented in Appendix B.

[*] A phrase suggested by E. E. Sellars, American Telephone and Telegraph Company.

Assume that the five calls indicated by dashed lines in Fig. 1 are in progress. Now, suppose that a subscriber served by the top line-finder group originates a call. Of the 16 idle line finders, only one can direct the call into the top selector half-shelf, i.e., the relatively congested part of the system. However, there are 15 line finders that will direct the call into one of the lower selector half-shelves, i.e., the relatively uncongested part of the system. The important point is that the call is much more likely to be routed through the relatively

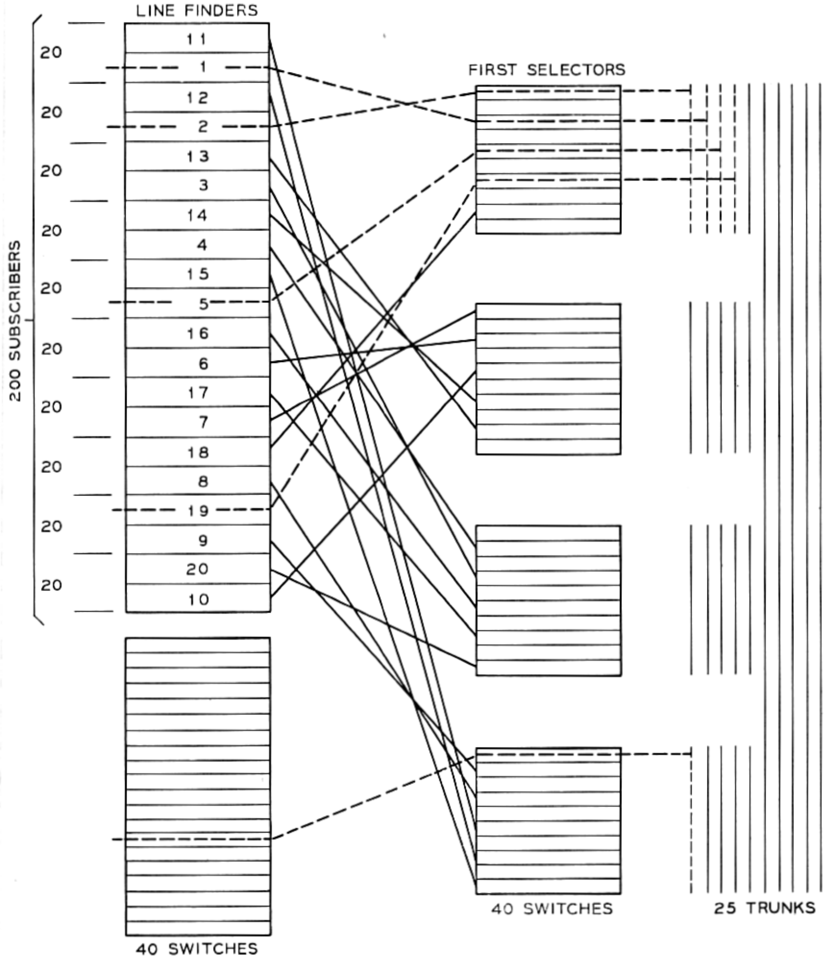

Fig. 1—Inherent load-balancing.

free part of the system than through the relatively busy part. The bias toward the less occupied selector half-shelves increases as the imbalance in occupancy increases. Therefore, we have the desirable situation wherein new arrivals tend to be offered to the less occupied first-choice subgroups. This results in a more uniform filling of the grading and a corresponding reduction in the chance that a call is blocked at one first-choice subgroup of the grading when there are idle trunks in some other subgroup. That is, the capacity of the grading is increased. The term "inherent load-balancing" is used to refer to this effect.

III. MATHEMATICAL MODELS

We wish to measure quantitatively the increase in grading capacity that results from the inherent load-balancing and to determine the validity of the tables.

3.1 *Complete Model*

For the complete model, the step-by-step system is considered in detail. Traffic is offered to the graded multiple through the complex interactions of the line finders and first selectors. Included in the model are all physical characteristics of the switching system deemed pertinent to the determination of the capacity of the grading. (These characteristics are specified in Appendix A.) The following assumptions are made concerning the traffic.

(*i*) Each line-finder group is fully equipped (i.e., contains 20 switches) and serves 200 subscribers.* Requests for service from each group of 200 subscribers are approximated by a Poisson process. The separate processes for the several groups are assumed to be independent with the same mean.

(*ii*) The 200 subscribers served by a line-finder group are divided into 10 subgroups of 20 subscribers each. For the ith subgroup of the jth line-finder group, let $b_{i,j}$ denote the number of busy subscribers. Given that the next arrival occurs in the jth line-finder group, the probability that the next arrival occurs in the ith subgroup is

$$\frac{20 - b_{i,j}}{200 - \sum_{i=1}^{10} b_{i,j}}.$$

[*] In practice, the number of working lines is not greater than 194. However, we assume 200 lines both because of the resulting numerical convenience and because of the negligible effect upon our results.

(*iii*) The line finders operate as described in Section A.2. Calls blocked at the line finders are delayed. An important aspect of the complete model is that the actual line-finder to first-selector wiring patterns are used. Therefore, traffic is distributed over the first selectors exactly as in the physical system.

(*iv*) Arrivals go to our level with probability p_ℓ , i.e., p_ℓ is the proportion of the calls arriving at the first selectors that require a trunk in our graded multiple.

(*v*) Holding times are independent and identically distributed according to a negative-exponential distribution.

(*vi*) Calls blocked at the graded multiple leave the system immediately and do not return. Calls directed to other levels are never blocked and remain in the system for one holding time.

3.2 *Approximate Model*

The complexity of the complete model arises both from the interactions occurring in the subscriber to line-finder network and from the interconnections between line finders and first selectors. To reduce complexity (and simulation computing time), an approximate model has been developed. In the model, the line finders and the line-finder to first-selector interconnections are modeled by assuming that an arrival seizes a selector at random from the group of idle first selectors.

The model is illustrated in Fig. 2. Suppose that the five calls indicated by dashed lines are in progress when a new call arrives. The probability that the call is served by a selector in the top half-shelf is 5/35 whereas the probability that the call is served by a selector in, say, the second half-shelf is 10/35. Because the arrival tends to be directed to the relatively uncongested part of the system, the model provides a good characterization of the inherent load-balancing.

The following assumptions are made concerning the traffic in the approximate model.

(*i*) All subscribers who can originate calls to our grading are considered as one large group. Requests for service from the group of subscribers are approximated by a Poisson process.

(*ii*) An arrival has full access to the first selectors and seizes a selector at random from the group of idle first selectors.

(*iii*) Assumptions (*iv*) and (*v*) from the complete model are also used here.

(*iv*) Calls blocked at the selectors or at the graded multiple leave the system immediately and do not return. Calls directed to

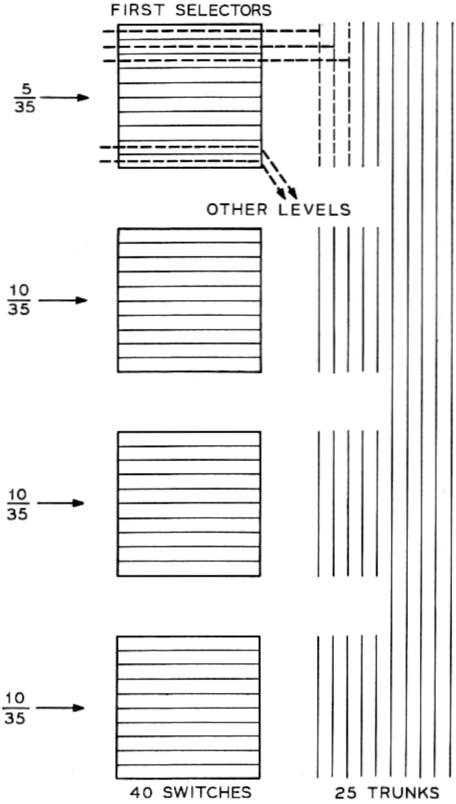

Fig. 2—Approximate model.

other levels are never blocked and remain in the system for one holding time.

3.3 *Isolated Model*

We wish to measure the increase in capacity that results from imbedding a graded multiple in the step-by-step system. Thus it is necessary to determine the capacity of the graded multiple when all traffic effects that arise because of the line finders and first selectors are ignored. This is called the isolated model. The load-loss relations for the isolated model can be computed by means of the equivalent-random method.[5] The following assumptions are made concerning the traffic.

(*i*) Calls arrive at the grading according to a Poisson process.

(*ii*) Arrivals are uniformly distributed over the first-choice subgroups.

(*iii*) Calls blocked at the graded multiple leave the system immediately and do not return.

(*iv*) Holding times are independent and identically distributed according to a negative-exponential distribution.

IV. NUMERICAL RESULTS AND EVALUATION OF TABLES

It is difficult to analyze a graded multiple even without the complexity of the step-by-step system. The approximations most often employed (such as the equivalent-random method[5]) are not directly applicable when the inherent load-balancing affects the capacity of the grading. Thus, two random-walk simulations were constructed, one using the complete model and one using the approximate model.

Considerable effort was devoted to determining the number of calls that should be processed at various loads to achieve a reasonably homogeneous coefficient of variation for the blocking probability. By a detailed statistical analysis of the simulation results, it was shown that, for a load a, the desired homogeneity is obtained by processing $1500/[p_\ell B(a)]$ calls where $B(a)$ is the blocking probability. The resulting coefficient of variation fell in the range of four to five percent. Since the number of calls processed (and, thus, the computing time) increases rapidly as p_ℓ becomes small, our results are for $p_\ell \geqq 0.5$.

Because the computing time is sensitive to p_ℓ, it would be desirable to show that meaningful results can be obtained when the multilevel aspects of the selectors are ignored, i.e., $p_\ell = 1$. In particular, it would be very useful if satisfactory estimates of blocking could be obtained from the approximate model with $p_\ell = 1$. This idea was tried successfully; the details are given in Section 4.1.

One purpose of this study is to determine the effect of the number of selectors connected to a graded multiple. Thus, two extreme situations are considered: the maximum and minimum number of selectors used with each grading. The maximum is always 320 selectors. The minimum is determined by the requirement that there be at least ten selectors for each first-choice subgroup.

For the graded multiples considered below, the tables in Ref. 1 were consulted to obtain the load-loss relations presently in use. The load-loss relations for the isolated model were obtained from the results in Ref. 6.

4.1 *Comparison of the Complete and Approximate Models*

The models are compared for both a small and a large graded multiple. In Figs. 3 and 4, results are presented for a 25-trunk and 45-trunk

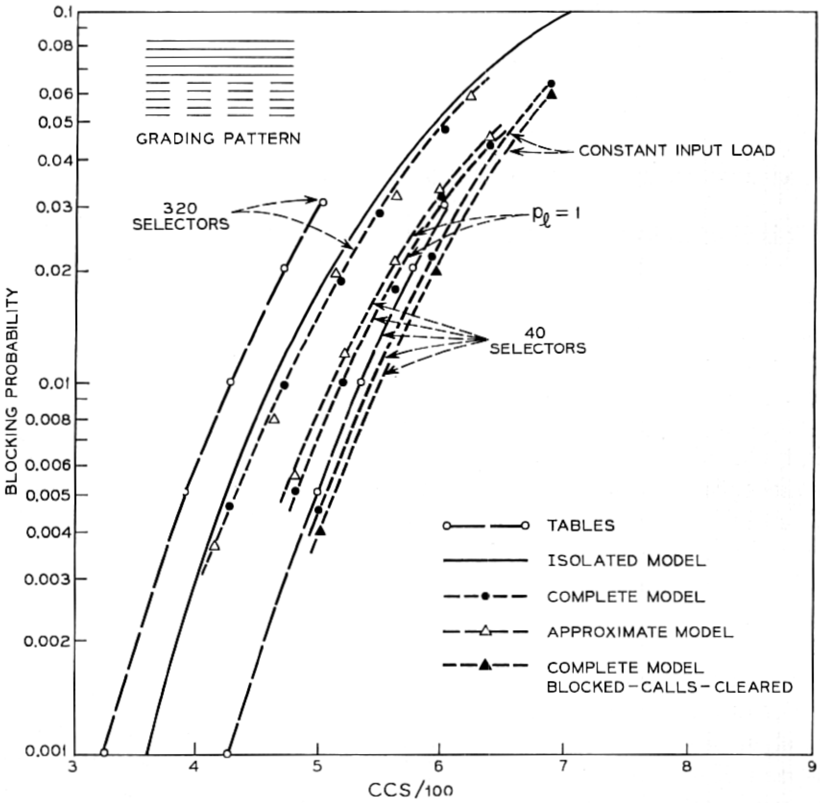

Fig. 3—Load-loss relations for a 25-trunk graded multiple.

graded multiple. Let us begin by describing how the four curves for 40 selectors in Fig. 3 (indicated by dashed lines) were obtained from the complete and approximate models: subsequently we shall comment on the significance of the curves. The top two curves, labeled $p_\ell = 1$, were generated by letting $p_\ell = 1$ in both the complete model and the approximate model. The next dashed curve was obtained by using the complete model when the load offered to the switching system was held constant at 983 CCS (27.3 erlangs) and the load offered to the grading was varied by changing p_ℓ* (six to seven percent of the calls were delayed at the line finders). A similar curve was obtained using

* Changing the load to the grading by varying p_ℓ keeps the load (and the blocking) approximately constant at the line finders. This seems to be the most realistic way to generate load-loss relations for the gradings.

the approximate model but, because the curve fell ever so slightly to the right of the $p_\ell = 1$ curve, it is not shown.

The bottom curve was also obtained by holding the load offered to the system constant at 983 CCS. However, in this case the complete model was modified so that the line finders operated on a blocked-calls-cleared basis; i.e., if all line finders in a line-finder group are busy when a call arrives at the group, the blocked call is immediately cleared (about two percent of the calls were blocked at the line finders). The system actually would operate between the latter two curves because, although blocked calls are delayed, defections can occur.

Consider the cause of the disparity in the curves. Because there are ten selectors per first-choice subgroup, no calls can be blocked at the line finders when $p_\ell = 1$. Therefore, the only difference in the models

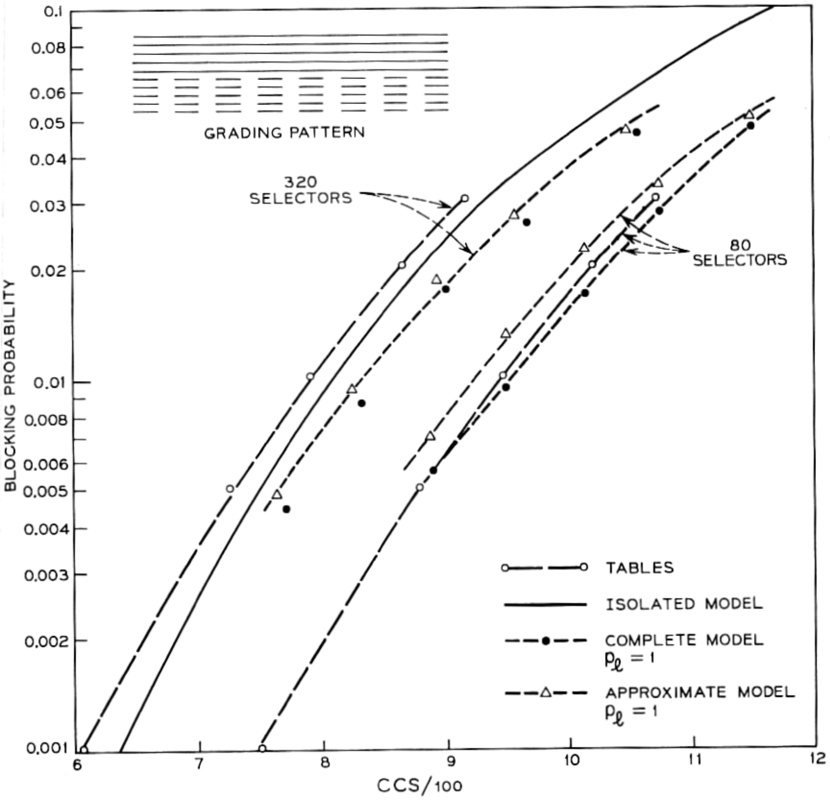

Fig. 4—Load-loss relations for a 45-trunk graded multiple.

is the representation of the inherent load-balancing, i.e., in complete detail versus our approximation. The capacity predicted by the approximate model is about 2.5 percent less than the capacity obtained from the complete model. We conclude that, in this case, the approximate model is a good characterization of the inherent load-balancing and provides a useful upper bound for the load-loss relation.

Recall that, in the approximate model, arrivals have full access to the 40 selectors. In the complete model, requests for service are directed to one of the two line-finder groups, thereby providing only partial access. It follows that, when $p_\ell < 1$, the traffic offered to the grading is somewhat smoother in the complete model than in the approximate model. The difference in smoothing appears to be the reason the complete model deviates from the $p_\ell = 1$ curve when $p_\ell < 1$ whereas the approximate model is relatively insensitive to changes in p_ℓ .

Furthermore, the two curves for constant input-load differ because blocked calls receive different treatment in the blocked-calls-cleared characterization of the line finders than in the blocked-calls-delayed operation. However, in both cases of constant input-load, the blocking at the line finders is somewhat higher than Bell System design and the deviation from the $p_\ell = 1$ curve is, therefore, somewhat exaggerated.

For 320 selectors, observe that the complete and approximate models give almost identical results for $p_\ell = 1$. It is expensive to evaluate completely the effects of smoothing because, in order to achieve significant smoothing at the line finders, p_ℓ must be small. One test was run with an input load to the switching system of 5500 CCS and $p_\ell = 0.12$ (only 0.3 percent of the calls were delayed at the line finders). In this case, the smoothing did not change the blocking probability from the $p_\ell = 1$ case.

In Fig. 4, the models are compared for 80 and 320 selectors connected to a 45-trunk grading. As above, the approximate model underestimates the capacity by about 2.5 percent for 80 selectors but, for 320 selectors, the two models are in good agreement.

We conclude that the approximate model provides a good characterization of the inherent load-balancing. For the minimum number of selectors, the approximate model underestimates the single-hour capacity by about 2.5 percent but, for larger numbers of selectors, it is in almost exact agreement with the complete model. When the originating traffic is smoothed somewhat by the line finders, the approximate model predicts blocking probabilities that are slightly higher than those obtained with the complete model. Thus, for the range of blocking at the line finders encountered in Bell System designs, the approximate model with $p_t = 1$ furnishes a reasonable upper bound for the load-loss relations. Consequently, we feel secure in using the approximate model for testing the other cases of interest.

4.2 *Results and Evaluation of Tables*

The validity of the tables is determined in this section. The approach is to begin with the small trunk groups and progress to the larger graded multiples. In Appendix D we present analytical results which provide useful relations between several of the system parameters. The results are obtained for the approximate model with $p_t = 1$ under the assumption that the graded-multiple gain is independent of the gain due to the inherent load-balancing.

Using the results of Section II, we see that the inherent load-balancing occurs only when a trunk group has at least two first-choice subgroups: the line-finder to first-selector interconnections do not enhance capacity in step-by-step trunk groups of sizes one through ten trunks (analytical justification for this result is provided in Appendix D). The tables agree with this point for nine or fewer trunks (where standard Poisson* capacities are used) but not for the case of ten trunks. For ten trunks, the tables underestimate single-hour capacity by as much as ten percent for 320 selectors, are fairly accurate for 40 selectors, but, for 20 selectors, overestimate single-hour capacity by varying amounts ranging from seven percent at three percent blocking to 22 percent at 0.1 percent blocking.

Next, consider the smaller gradings with two first-choice subgroups, i.e., 11 through 19 trunks. The results in Appendix D imply that the capacity of a graded multiple of N trunks is bounded above by the capacity of a full-access group of N trunks having Poisson input. However, for 20 selectors, the tables show a capacity that, to varying degrees, exceeds the capacity of the corresponding full-access groups: the discrepancy is greatest for 11 trunks and decreases with increasing trunk-size until, for 19 trunks, the tables show a capacity that is very close to the capacity of the 19-trunk full-access group.

Figures 5 and 6 give results for gradings of 11 and 19 trunks, respectively. In each figure, the load-loss relation for the isolated viewpoint and the Erlang-B load-loss curve represent upper and lower bounds,

* R. I. Wilkinson[7] has shown that the blocked-calls-cleared assumption (i.e., the Erlang-B load-loss relation) gives good estimates of single-hour losses for full-access trunk groups. However, when typical day-to-day variations are included, the increased average loss is better approximated by the blocked-calls-held assumption (i.e., the Poisson load-loss relations).

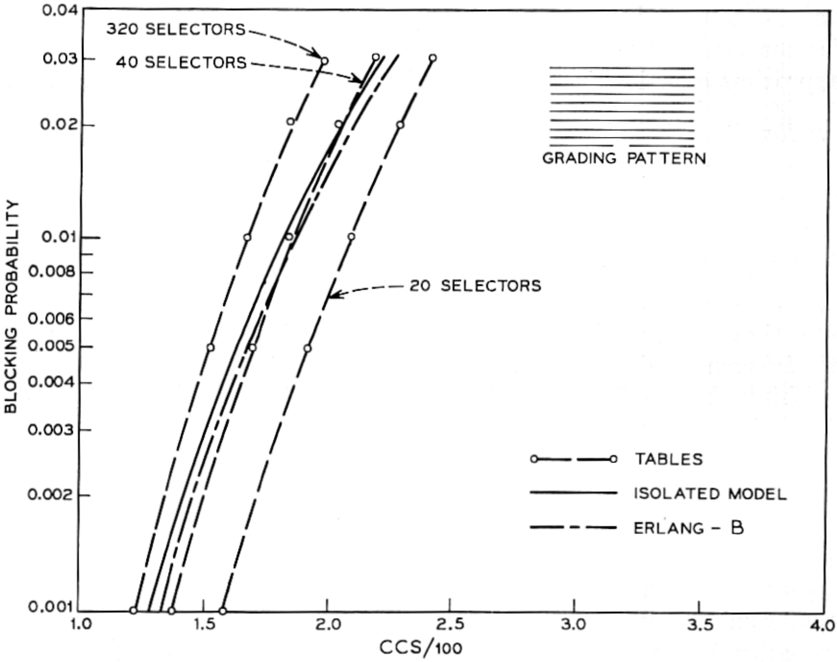

Fig. 5—Load-loss relations for the 11-trunk graded multiple.

respectively, for the actual load-loss relations. Because these bounds are close for 11 trunks, there can be little inherent load-balancing. The tables underestimate by about ten percent the single-hour capacity of the 11-trunk grading for 320 selectors, are nearly correct for 40 selectors, but overestimate by about 20 percent the single-hour capacity for 20 selectors. The results in Fig. 6 for the 19-trunk grading indicate that the tables are accurate for 20 selectors (this 19-trunk grading is only used with 20 selectors).

Now, consider some representative larger gradings. Results for 19-trunk and 25-trunk gradings each with four first-choice subgroups are given in Figs. 7 and 3, respectively. Similarly, results for 37-trunk and 45-trunk gradings each with eight first-choice subgroups are shown in Figs. 8 and 4, respectively. In each case, the tables are fairly close to the simulation results for the minimum number of selectors but, to varying degrees, are conservative for the 320-selector case. These comparisons are made on the basis of single-hour capacity.

Because the inherent load-balancing increases as the number of first-choice subgroups increases, one might expect the capacity of gradings with 12 subgroups to be considerably higher than the tables indicate. To check this point, a grading of 45 trunks having 12 first-choice subgroups was tested (see Fig. 9). The grading is used with either 120 or 240 selectors. The simulation results indicate that the tables do underestimate capacity by about ten percent.

Because "grade of service" for the gradings in the step-by-step system means average service in the busy-season busy hours, the tables are generally used in situations where the effects of day-to-day variations are included. From Ref. 7 we observe that the effects of day-to-day variations cause a decrease in capacity of approximately seven percent. Consequently, from the above remarks concerning the single-hour capacities of the gradings, it follows that the accuracy of the tables for engineering in the presence of day-to-day variations also depends upon the size of the gradings and the number of selectors. Specifically, we find that the tables are fairly accurate when the

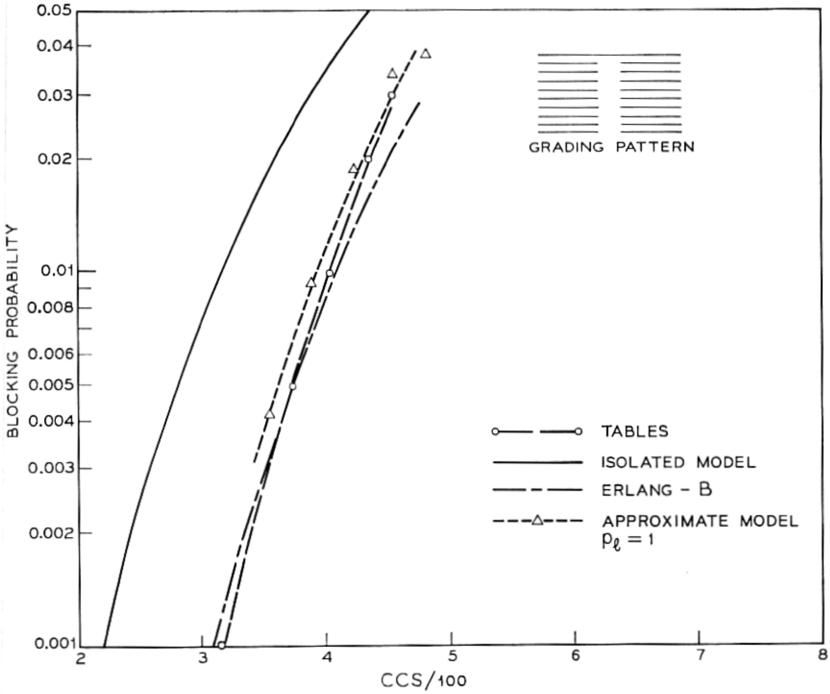

Fig. 6—Load-loss relations for a 19-trunk graded multiple with 20 selectors.

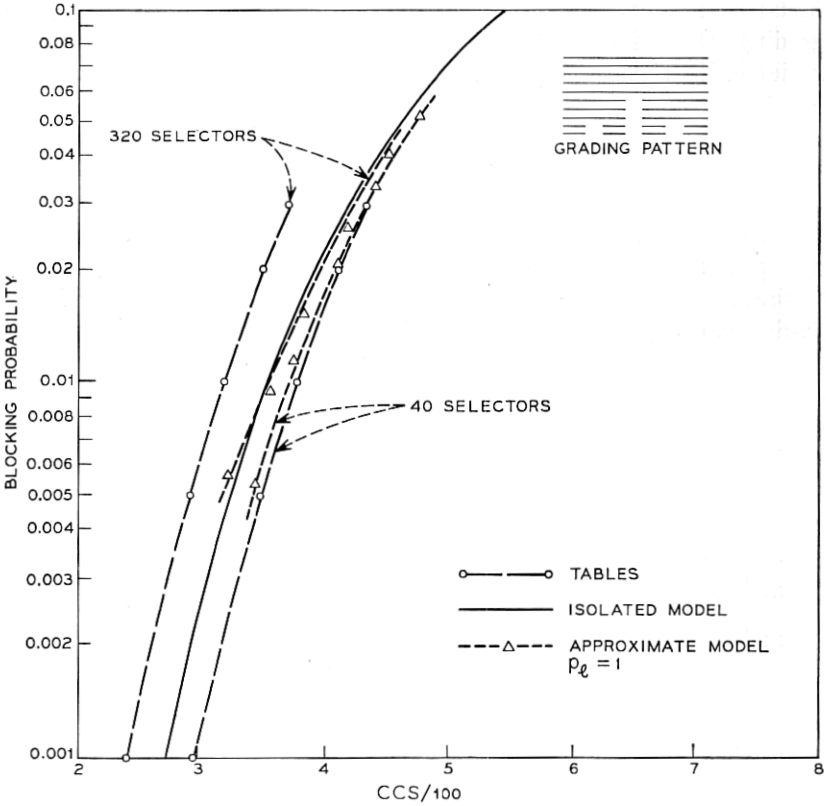

Fig. 7—Load-loss relations for a 19-trunk graded multiple.

gradings are used with the large selector groups. For the minimum number of selectors, the tables estimate fairly accurately the capacity of the larger gradings, slightly overestimate the capacity of the medium-sized gradings, and overestimate the capacity of some of the smaller gradings by about 20 percent. The tables are correct for trunk groups of one through nine trunks but, for ten trunks, they incorrectly indicate that the capacity varies with the number of selectors connected to the trunk group.

We conclude that the capacity of a step-by-step graded multiple does increase as the number of selectors connected to the grading decreases. The reason seems to be that, with fewer selectors, an imbalance in the loading of the grading results in a larger probability that a new arrival is directed to the relatively idle areas of the grading.

The simulations also indicate that the inherent load-balancing is more pronounced in the larger gradings having many first-choice subgroups. Analytical results in Appendix D imply that the inherent load-balancing does increase as the number of first-choice subgroups increases, provided the number of selectors per subgroup is held constant. The tables reflect this phenomenon too.

Notice that the simulations imply that the inherent load-balancing is more pronounced in the 45-trunk grading than the 37-trunk multiple even though both gradings have eight first-choice subgroups. In Appendix D, it is also shown that for a fixed number of first-choice subgroups and selectors, the inherent load-balancing increases as the number of trunks in these subgroups increases. Since the 37-trunk grading effectively has fewer trunks in the first-choice subgroups, the

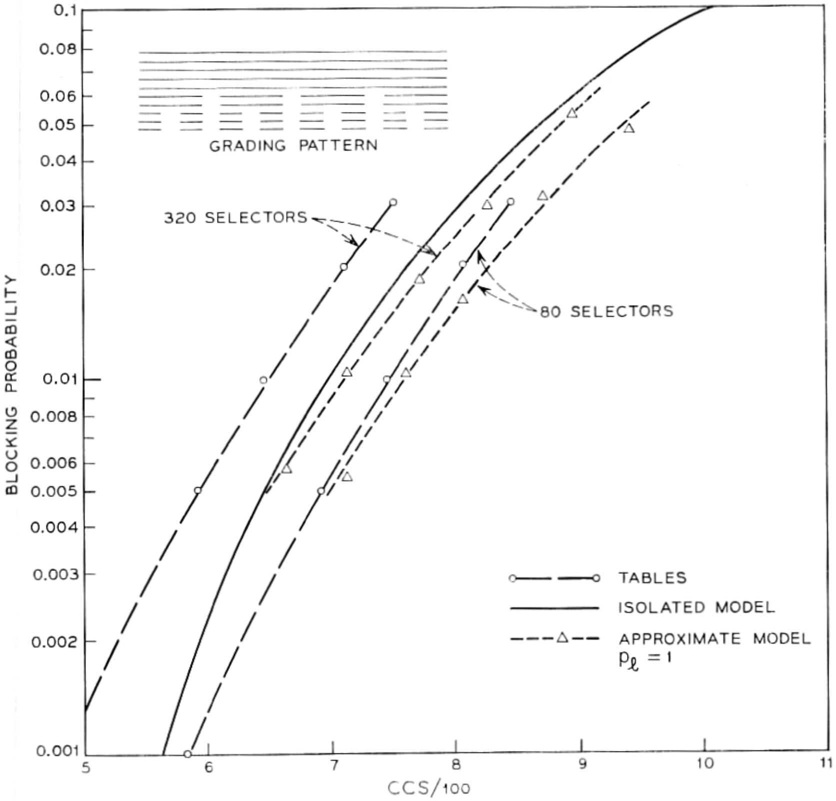

Fig. 8—Load-loss relations for a 37-trunk graded multiple.

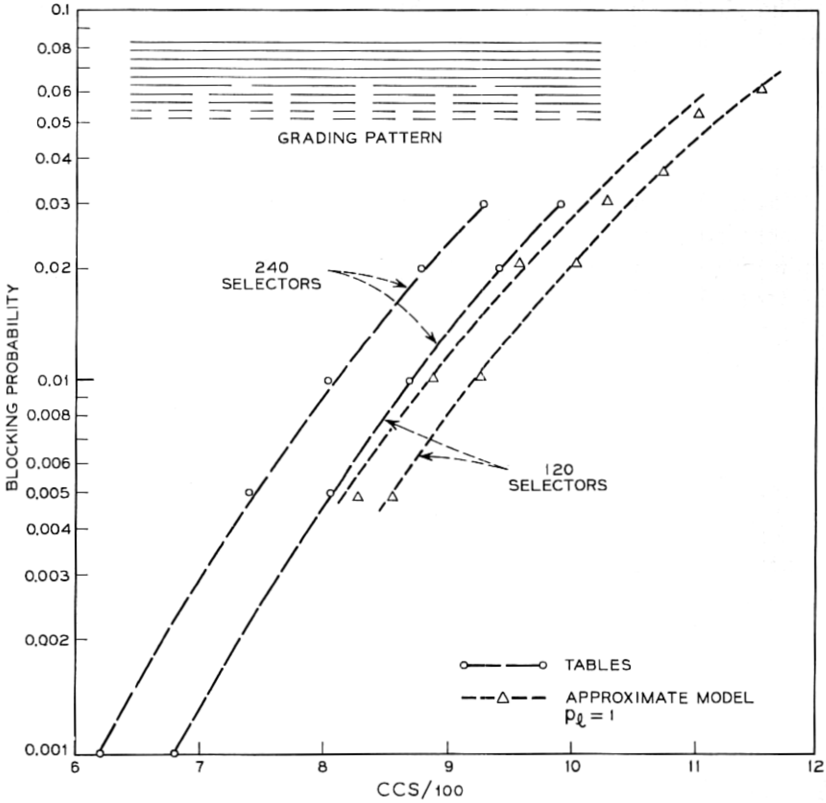

Fig. 9—Load-loss relations for a 45-trunk graded multiple.

simulation results are consistent. The tables also agree with this conclusion.

V. CONCLUSIONS

We have shown that the capacity of a graded multiple is increased when it is imbedded in the step-by-step system. The increase occurs because of the inherent load-balancing that arises from the interconnections between subscribers, line finders, and first selectors. Furthermore, the increase becomes larger as the number of selectors connected to the grading decreases.

The accuracy of the tables depends upon the size of the gradings and the number of selectors. Specifically, we find that, when used for engineering in the presence of day-to-day variations, the tables are reasonably accurate when the gradings are used with the large selector groups. For the minimum number of selectors, the tables estimate fairly accurately the capacity of larger gradings, slightly overestimate the capacity of the medium-sized gradings, and overestimate the capacity of some of the smaller gradings by about 20 percent. The tables are correct for trunk groups of one through nine trunks but, for ten trunks, they incorrectly indicate that the capacity varies with the number of selectors connected to the trunk group.

Revised tables incorporating the improvements achievable by the methods described herein are being generated. The revised tables consist of two sections, one for single-hour engineering and one which includes the effects of day-to-day variations. The second section is of particular importance because "grade of service" for the graded multiples in the step-by-step system means average service in the busy-season busy hours.

This study considered the graded multiples that are connected to the first stage of selectors. However, the connections between selector stages are arranged so that there is also inherent load-balancing for the gradings which appear at subsequent stages. Consequently, if the smoothing effect of the preceding selector stages is negligible, the results presented herein are also valid for graded multiples at all stages in the distribution network. Thus, our results furnish upper bounds for the blocking on subsequent stages.

VI. ACKNOWLEDGMENTS

R. V. Laue conducted the statistical studies leading to the quantitative measure of the coefficient of variation of the blocking probability. We are grateful to R. I. Wilkinson and E. E. Sellars for reading an early version of this paper and making several suggestions for improvement. The historical data presented in Appendix B were obtained from documents furnished by Wilkinson. He also brought to our attention the useful information in Ref. 7. We gratefully acknowledge the numerous conversations with M. F. Morse which helped us to understand the pertinent physical characteristics of the step-by-step switching system.

APPENDIX A

Traffic Flow in the Step-by-Step Switching System

We give a summary of the operation of the step-by-step switching system. An attempt is made to show how the physical design of the switching system influences traffic flow. Because the size and layout of a step-by-step system can vary, we describe one particular configuration to illustrate the necessary points. The material in this appendix has been gathered from many sources. A large portion has been obtained from Refs. 3 and 8 and is presented without further specific reference.

A.1 *General Summary*

The step-by-step switch is used in three different modes, as: a line finder, a selector, and a connector. Each subscriber line appears on the terminal bank of several line finders. The exact number of line finders is determined by traffic considerations. When the calling party goes off-hook, one of the line finders serving the subscriber is chosen to hunt for and find the requesting line.* Each line finder is permanently connected to a first selector.† Thus, when the line finder locates the requesting line, the system is ready to receive the first dialed digit and dial tone is returned.

In response to the first digit, the first selector steps vertically to the appropriate level and then hunts horizontally across the terminals on the level for an idle trunk to the second selector stage. Upon locating an idle trunk, the first selector stops. The subscriber is thereby connected to a second selector through the line finder, the first selector, and the trunk from the first selector to the second selector. The second digit, which determines the movement of the second selector, can now be supplied by the subscriber. The remaining selector stages operate similarly with each selector switch controlled by a different digit.

After passing through the required number of selectors,‡ the call reaches the connector switch. The next to the last digit causes the connector to step vertically to the appropriate level and the last digit causes the connector to move horizontally across the level to the appropriate terminal. If the line is free, ringing current is applied. If the line is busy, busy tone is returned. During the establishment of a connection, it is possible that a selector may find all available trunks busy. When this occurs, the selector returns the reorder tone.

* In Section A.2, we specify which line finder is chosen.

† This applies to a basic step-by-step system not equipped with *Touch-Tone®* service or common control.

‡ Because seven digits are the standard number dialed and because such a large selection is frequently not required, some of the digits are not needed to complete the call. The step-by-step system can perform "digit absorption." Thus, the number of selector stages is variable and is determined by local considerations.

A.2 *Line Finders*

In the more recent step-by-step offices, 200-point line-finder switches* are used. Hence, the subscriber lines are divided into groups of 200 lines.† Each group of lines is served by a group of at most 20 line finders.‡

For a line finder to have access to 200 subscriber lines, each of the 100 switch positions serves two lines. A relay circuit associated with the line finder determines which subscriber line is calling. The 200 lines are divided into ten groups of 20 subscribers and are arranged on the switch banks in a slipped multiple, as shown in Fig. 10. In Fig. 10, each rectangle represents a group of 20 subscriber-lines. For example, the particular group of 20 lines represented by the shaded rectangle appears on level 5 of switch 1, level 4 of switch 2, . . . , level 1 of switch 5, level 10 of switch 6, . . . , and level 6 of switch 20. Thus, each group of 20 subscriber lines will terminate on the first level of two switches, the second level of two switches, . . . , and the tenth level of two switches. In order to minimize hunting time, the line finder chosen to service a request is the available line finder with the requesting line on the lowest level. Because each subscriber line appears on two line finders for each level, a priority is established to determine which of the two line finders is used first. In some systems, the preferred line finder is that with the lowest number; in other systems, the preference alternates on successive calls.

For our study, we model the selection rules for the line finders as follows. The 200 subscribers served by a line-finder group are subdivided into ten groups of 20 subscribers. For each subgroup, first- and second-choice line finders are specified as shown in Fig. 11 (here, the line finder with the lowest number is used first). For example, the subgroup indicated by an asterisk in Fig. 11 uses line finder 5 as the first choice and line finder 15 as the second choice. If line finders 5 and 15 are both busy, the system then searches for an idle line finder in the order 4, 14, 3, 13, 2, 12, . . . , 6, 16.

When all line finders are busy and new requests for service arrive, the system acts as a delay system with defections; i.e., the requests

* We restrict our attention to 200-point line finders, although 100-point line finders are used in some offices.

† Although 200-point line finders can serve 200 lines, in practice the number of working lines is not greater than 194. However, we assume that the subscribers are divided into groups of 200 both because of the resulting numerical convenience and because the assumption has a negligible effect upon our results.

‡ For minor classes of lines with high calling rates, it is possible to use a line-finder group of 30 switches.

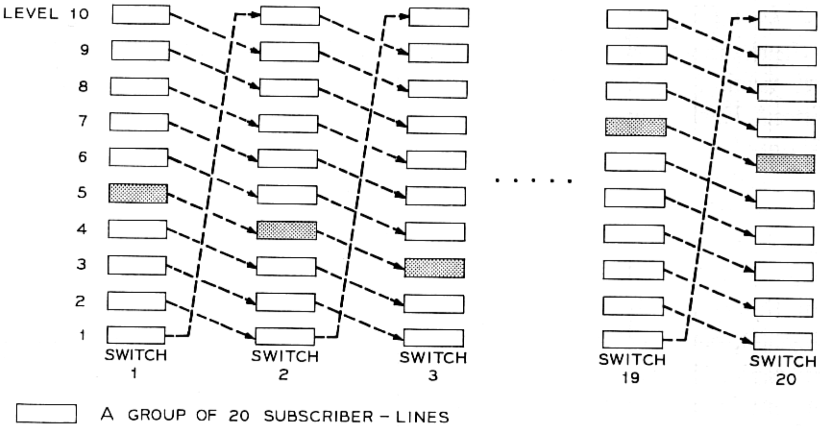

Fig. 10—Slipped multiple on terminal banks of line-finder switches.

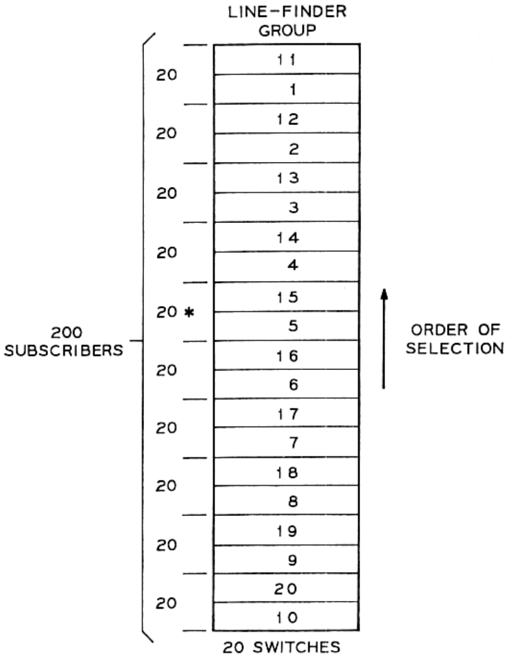

Fig. 11—Line-finder selection sequence.

join a queue and wait as long as necessary for dial tone unless the subscriber tires of waiting and hangs up. (For our study, we assume that there are no defections and no retrials.)

The order of service for requests in the queue is rather interesting. The queue is a backlog of requests for line finders to hunt for the requesting lines. Thus, when a line finder becomes free, it will immediately begin to hunt. The requesting line that is served is the one that appears lowest on the terminal bank of the hunting line finder.

A.3 *First Selectors*

The selector switches are mounted on frames with a capacity of 320 switches and are arranged on shelves with a capacity of 20 selectors per shelf. Each shelf is divided into two half-shelves of ten switches each. The corresponding terminals of all selectors on a half-shelf are permanently multipled together. It would be possible to provide a small independent trunk-group from each level of each half-shelf of first selectors to the appropriate second-selector switches. However, it is more efficient to arrange trunks in a graded multiple than in small independent trunk-groups.[9] Accordingly, the first selectors for a particular class of service are divided into groups of 20, 40, 60, 80, 120, 160, 240 or 320 switches depending upon the traffic and the number of switches in the class. On each level, the trunks from a group of first selectors to the second selector stage are arranged in a graded multiple. The graded multiple used on a particular level is determined by the amount of traffic directed to the level.

If a call is offered to a graded multiple and an idle trunk is not available, the reorder tone is returned. Thus, calls blocked at the graded multiples are cleared. Although there is actually a six- to ten-second delay, we assume that blocked calls are cleared immediately.

A.4 *Line-Finder to First-Selector Interconnections*

Each line finder is permanently connected to a unique first selector. In order to distribute the traffic evenly over the first selector stage, the links from the line finders are distributed over the first selectors in fixed patterns. The patterns attempt to connect a set of line-finder groups to a set of first-selector half-shelves such that the traffic from a line-finder group is uniformly distributed over a number of selector shelves.

Two fully equipped line-finder groups and four first-selector half-shelves are shown in Fig. 1. For clarity, the line-finder to first-selector connection pattern is shown only for one line-finder group. In Sections III and IV, we show that the patterns and the selection rules for the line finders are extremely important in establishing the traffic capacity of the graded multiples that connect the first selectors to the second selectors.

APPENDIX B

History of the Development of the Tables

The early step-by-step systems did not use graded multiples. Consequently, corresponding analysis of trunk capacity concerned itself only with a consideration of the effect of the switching apparatus located at the input of a trunk group. Since the selectors only had ten terminals, large trunk groups were split into smaller subgroups not exceeding ten trunks each. The original mathematical efforts on this problem were carried out by E. C. Molina[10,11] in the early twenties. His results are known as the splits-in-the-multiple theory.

The original step-by-step systems used a primary-secondary switching arrangement instead of line-finder groups. Molina was aware of the importance of the various switch interconnections. However, it appears that he may not have been quite certain of their actual effect since he presented three different mathematical models for the engineers to consider. After the construction of load-loss relations for each of the three models, it was decided empirically that the (calls distributed) collectively-at-random model was most appropriate.

The principal feature of the collectively-at-random model is that the busy selectors are distributed over the group of all selectors according to a hypergeometric distribution. More precisely, assume that the outgoing trunk group is split into g groups of c trunks which are connected to s selectors; that is, there are cg trunks and sg selectors. Now, let M denote the (random) number of busy selectors in the first subgroup, and N the number in the remaining $(g - 1)c$ subgroups. Then, the collectively-at-random model assumes that

$$P\{M = m, N = n \mid M + N = m + n\} = \frac{\binom{s}{m}\binom{(g-1)s}{n}}{\binom{gs}{m+n}}.$$

Some time later (in 1941), Wilkinson observed that the load-loss relations furnished by the splits-in-the-multiple theory were well approximated by the relations resulting from the assumption that the selectors were independent traffic sources. That is, for engineering purposes, the step-by-step system could be viewed as a system with finite-source input, the number of sources being taken as the number of selectors. Thus, the increase in trunk capacity (due to the primary-secondary to selector interconnections and explained by the splits-in-the-multiple theory) came to be known as the finite-source effect. (This terminology may have caused some of the confusion cited in the Introduction and Appendix C.)

Graded multiples were placed in use in step-by-step systems during the late twenties. The original step-by-step gradings used a maximum of 19 trunks. This number was increased to 37 in 1941, and to the present level of 45 in 1949.

A mathematical basis for engineering graded multiples was furnished by Molina some time around 1925. His graded-multiple work is still referred to as the no-holes-in-the-multiple theory. The name arises from the assumption that calls on common trunks are transferred to individuals as soon as an individual trunk becomes idle. Since such a grading would have a higher capacity than is actually achievable, Molina suggested using only one-half the predicted increase in capacity (over a full-access 10-trunk group). In 1927, a data analysis of gradings in a panel system indicated that one-half gain was a little low,[9] but the difference was not great enough to cause much concern. In 1931, Wilkinson[9] published a set of engineering curves (which he called hump-back curves) giving the increase in capacity for several classes of simple graded-multiples. These curves became the basis for engineering graded multiples in the Bell System.

It appears that the original step-by-step selector-multiple tables were constructed in 1925 by P. P. Coggins and J. R. Ferguson of the AT&T Company. These tables (for gradings up to 19 trunks) used the splits-in-the-multiple theory to estimate the increase in capacity due to primary-secondary interconnections. A graded-multiple gain was obtained from "Graded Random 'B' " curves developed by Ferguson in 1924. The results were then combined to give a total gain in capacity over a full-access ten-trunk group having Poisson input.

When the size of the gradings was increased to 37 (in 1941), Wilkinson's half-gain hump-back curves were used to obtain the graded-multiple gain. At this time, a "finite-source gain" was estimated by curves which were apparently constructed from a combination of splits-in-the-multiple theory, binomial capacity and some smoothing "based on engineering judgment."

Although one can raise several objections to the preceding approach, the results were reasonably good. Moreover, the material in Section IV indicates that the original work is still fairly good for trunk engineering in present step-by-step systems, even though line finders are now used in place of a primary-secondary switching system. The results in Appendix D show that the main reason the original work is still adequate can be traced to the fact that assumption two in the approximate model (Section 3.2) is implicitly accounted for in Molina's collectively-at-random model.

APPENDIX C

Common Incorrect Arguments

We present several common arguments that supposedly explain the capacities of the graded multiples in step-by-step systems.

C.1 *Finite-Source Effect*

From the finite-source effect viewpoint, the increase in the capacity of a grading is attributed to the finite number of selectors located before the trunk group. Since originating traffic must pass through the selectors prior to reaching a trunk group, it is reasoned that the traffic to the group will be similar to that from a finite number of independent traffic-sources.

Below, we give two statements of the argument based on the finite-source effect. The first is taken from Ref. 2, page 15; the second, from Ref. 3 (in slightly modified form).

(*i*) Limited source is the constrictive effect of forcing traffic to enter a trunk group through a limited number of sources, in our case, selectors. The maximum effect of limited sources is illustrated where 10 selectors have access to 10 trunks. No matter how much traffic is offered, the trunks themselves cannot be the cause of any blocking; and their capacity is not 149 CCS (for P.01 grade-of-service using blocked-calls-held assumption) but 360 CCS, which represents 100-percent usage of each trunk. If 11 selectors have access to these 10 trunks, this maximum trunk capacity would no longer be obtained since, with 10 calls in progress through the system, the eleventh selector may offer a call resulting in a blockage. However, only one of the 11 selectors can produce this condition at any one time, hence, the capacity, though not 360, is still much greater than 149. And so as we increase the number of selectors to exceed the number of trunks more and more, the capacity of the group of 10 trunks approaches more closely the basic unlimited source capacity of 149 CCS.

(*ii*) It will be found that 10 trunks with varying numbers of selectors feeding traffic thereto will carry traffic as follows: 320 selectors on 10 trunks, 149 CCS; · · ·; 40 selectors on 10 trunks, 164 CCS. The apparent increase in capacity as the number of selectors (traffic sources) is reduced is the gain that results from the "limited source" effect. That is, as each trunk becomes busy, the remaining sources are reduced by one, reducing to this extent the probability of all remaining trunks becoming busy.

There is considerable dissent about the accuracy of these statements. The most common counter-argument claims that the traffic offered to the trunk group does not decrease gradually as the selectors become busy but is concentrated through the idle selectors, i.e., the entire load is shifted to the idle selectors. As a result, the traffic capacity of the trunk group does not decrease gradually as the number of selectors increases from ten. Instead, the capacity should change drastically as one goes from ten to eleven selectors and, thereafter, remain approximately constant as the number of selectors increases from eleven. Of course, one is free to speculate how the finite-source effect argument might apply to a graded multiple.

Notice that both the argument and the counter-argument (and all variations on these themes) differ substantially from the correct viewpoint based on inherent load-balancing (as described in Section II). Also, the arguments are very imprecisely stated and can vary considerably depending on how the missing details are chosen. The underlying models bear little resemblance to the step-by-step system and, thus, one cannot point to a few specific defects.

C.2 *The Effect of Clipping*

Now, we present the arguments of those who believe the finite-source effect to be negligible. Consider a graded multiple on a particular selector-level. The arrivals to the grading come from half-shelves of selectors. Assuming that calls for the entire system arrive according to a Poisson process,* one might assume that the arrivals from a half-

* Only high-usage gradings (first-choice routes) are considered in this study.

shelf of selectors also constitute a Poisson process until all the selectors in the half-shelf become busy (due to calls on the several levels), at which time no new arrivals can occur. That is, clipping takes place, leading to the term "clipped Poisson process."

R. R. Mina[4] has examined the effect of clipping. Specifically, he considered the following situation:

(*i*) Exactly one selector half-shelf (ten selectors) is connected to each first-choice subgroup.
(*ii*) Calls arrive at each selector half-shelf according to a Poisson process. The several processes are independent with the same mean.
(*iii*) All arrivals go to the level under consideration (i.e., the multilevel aspect of the system is ignored).
(*iv*) Holding times are independent and identically distributed according to a negative-exponential distribution.

Using this model, Mina has shown (using a simulation) that grading capacity is not significantly enhanced by the clipping. Because Mina believes that his model adequately represents the step-by-step system, he further asserts [Ref. 4, page 12]:

> This simulation shows that the effect of the limitation of the number of input switches is insignificant and that the increase in the traffic carrying capacity of graded trunk tables used in the U.S.A. is not realistic.
>
> Furthermore, the structure of the AT&T tables suggests that small gradings, since they can be formed with fewer groups, are more efficient than larger gradings. This is obviously not true as it is apparent that large gradings are more efficient than small gradings.

Also [Ref. 4, page 24],

> The principle of limited sources, on which the AT&T tables are based, is not realistic and underestimates the quantity of trunks.

Mina's comments would be valid if the input to a grading were adequately approximated by a clipped Poisson process. However, such a model overlooks a significant aspect of step-by-step systems: the inherent load-balancing that results from the subscriber to line-finder network and the interconnection of line finders and first selectors [see Section II].

APPENDIX D

Inherent Load-Balancing

The purpose of this appendix is to relate Molina's collectively-at-random model[10,11] to the approximate model with $p_\ell = 1$ presented in Section 3.2. To accomplish this task, the graded-multiple gain is assumed to be independent of the inherent load-balancing. We model the system as follows:

> A service system $\mathcal{S}$ has $c_1 + \cdots + c_g$ servers separated into g independent groups $G_1, \cdots, G_g$. The group G_i contains c_i servers, and associated with G_i is a switch-group S_i containing s_i switches, $s_i \geqq c_i$: a switch in S_i is allotted to each customer being served in G_i.
>
> Customers arrive at $\mathcal{S}$ according to a Poisson process with intensity a. If n_i denotes the number of busy servers in G_i, $i = 1, \cdots, g$, at an arrival epoch, the customer is sent to G_k with probability
>
> $$\frac{s_k - n_k}{(s_1 + \cdots + s_g) - (n_1 + \cdots + n_g)}, \qquad k = 1, \cdots, g.$$
>
> If a customer arrives to find all positions occupied in all switch-groups, he leaves the system and does not return. If a customer is directed to G_i when at least one of the c_i servers is available, a server is selected and service commences immediately. An arrival occurring when all c_i servers are busy leaves the system and does not return.
>
> All service times are assumed to be independent and identically distributed according to a negative-exponential distribution having unit mean.

In order to gain some insight into the problem, the case $g = 2$ is examined first. Assume that the system is in statistical equilibrium, let N_i denote the number of busy servers in G_i at an arbitrary instant, and define

$$p(n_1, n_2) = P\{N_1 = n_1, N_2 = n_2\}.$$

The following relations determine p. For $0 \leqq n_i \leqq c_i - 1$,

$$\begin{aligned}(a + n_1 + n_2)&p(n_1, n_2)\\ &= a\frac{s_1 - n_1 + 1}{s_1 + s_2 - n_1 - n_2 + 1}p(n_1 - 1, n_2)\\ &\quad + a\frac{s_2 - n_2 + 1}{s_1 + s_2 - n_1 - n_2 + 1}p(n_1, n_2 - 1)\\ &\quad + (n_1 + 1)p(n_1 + 1, n_2) + (n_2 + 1)p(n_1, n_2 + 1). \qquad (1)\end{aligned}$$

If $0 \leqq n_1 \leqq c_1 - 1$,

$$\left(a \frac{s_1 - n_1}{s_1 + s_2 - n_1 - c_2} + n_1 + c_2\right) p(n_1, c_2)$$

$$= a \frac{s_1 - n_1 + 1}{s_1 + s_2 - n_1 - c_2 + 1} p(n_1 - 1, c_2)$$

$$+ a \frac{s_2 - c_2 + 1}{s_1 + s_2 - n_1 - c_2 + 1} p(n_1, c_2 - 1)$$

$$+ (n_1 + 1) p(n_1 + 1, c_2). \tag{2}$$

An expression similar to (2) can be written for $0 \leqq n_2 \leqq c_2 - 1$. Finally, for $n_1 = \mathbf{c}_1$, $n_2 = \mathbf{c}_2$,

$$(c_1 + c_2) p(c_1, c_2) = a \frac{s_1 - c_1 + 1}{s_1 + s_2 - c_1 - c_2 + 1} p(c_1 - 1, c_2)$$

$$+ a \frac{s_2 - c_2 + 1}{s_1 + s_2 - c_1 - c_2 + 1} p(c_1, c_2 - 1). \tag{3}$$

Moreover,

$$\sum_{n_1=0}^{c_1} \sum_{n_2=0}^{c_2} p(n_1, n_2) = 1. \tag{4}$$

Equations (1) through (3) can be solved by using the definition of conditional probability to write

$$p(n_1, n_2)$$

$$= P\{N_1 = n_1, N_2 = n_2 \mid N_1 + N_2 = n_1 + n_2\}$$

$$\cdot P\{N_1 + N_2 = n_1 + n_2\}.$$

Now, taking note of the way the arrivals are distributed over G_1 and G_2, one might guess that

$$P\{N_1 = n_1, N_2 = n_2 \mid N_1 + N_2 = n_1 + n_2\} = \frac{\binom{s_1}{n_1}\binom{s_2}{n_2}}{\binom{s_1 + s_2}{n_1 + n_2}}, \tag{5}$$

and try

$$P\{N_1 + N_2 = n\} = k \frac{a^n}{n!},$$

where k is an appropriate constant. Direct substitution shows that

$$p(n_1\,,n_2) = p(0,0)\,\frac{\binom{s_1}{n_1}\binom{s_2}{n_2}}{\binom{s_1+s_2}{n_1+n_2}}\,\frac{a^{n_1+n_2}}{(n_1+n_2)!} \tag{6}$$

satisfies equations (1) through (3). Equations (4) and (6) can be used to determine $p(0,0)$. Since a solution to equations (1) through (4) is unique, the system state probabilities are determined. It is interesting that the distribution (6) can be viewed as a modification of a product of two Engset distributions $k_i\binom{s_i}{n_i}a^{n_i}$, $i = 1, 2$.

If necessary, these state probabilities can be generated numerically from the relations

$$p(n_1+1, n_2) = \frac{a}{n_1+1}\,\frac{s_1-n_1}{s_1+s_2-n_1-n_2}\,p(n_1\,,n_2)$$

and

$$p(n_1\,,n_2+1) = \frac{a}{n_2+1}\,\frac{s_2-n_2}{s_1+s_2-n_1-n_2}\,p(n_1\,,n_2),$$

which follow from (6).

The call congestion is given by

$$B(a;c_1\,,c_2) = (s_1-c_1)\sum_{n_2=0}^{c_2}\frac{p(c_1\,,n_2)}{s_1+s_2-c_1-n_2} + (s_2-c_2)\sum_{n_1=0}^{c_1}\frac{p(n_1\,,c_2)}{s_1+s_2-n_1-c_2}. \tag{7}$$

Perhaps the most important aspect of this model is the appearance of the hypergeometric distribution shown in equation (5). As noted in Appendix B, Molina's collectively-at-random model was characterized by this distribution. Consequently, it is no longer surprising that the tables (based on Molina's model) adequately reflect the inherent load-balancing in most of those cases tested.

There are two limiting cases of interest; namely, $s_i \to \infty$ and $s_i \to c_i$. Setting $s_1 = s_2 = s$ yields $(s_i - n_i)/(s_1 + s_2 - n_1 - n_2) \to 1/2$ as $s \to \infty$. Hence, when the number s of selectors is very large, the system operates as two independent systems with c_i servers and Poisson input of intensity $a/2$.

Equation (7) shows that there is no blockage at either service area when $s_i = c_i$, $i = 1, 2$ (since all blockage takes place at the switch groups). However, the actual capacity of the system $\mathcal{S}$ (consisting of switch groups and servers) is identical to that obtained from a full-access group of $c_1 + c_2$ servers.

Thus, when the number s_i of switching positions (selectors) exceeds c_i, the capacity of the system is bounded above by that of a full-access group of $c_1 + c_2$ servers, and below by that of two independent groups of c_1 and c_2 servers respectively; that is the system capacity increases as the number of switches decreases.

The preceding results can be extended in a straightforward fashion to any number g of service groups. Letting N_i denote the number of busy servers in G_i, $i = 1, \cdots, g$, it can be shown that

$$p(n_1, \cdots, n_g) = P\{N_i = n_i, i = 1, \cdots, g\}$$

$$= \frac{\binom{s_1}{n_1} \cdots \binom{s_g}{n_g}}{\binom{s_1 + \cdots + s_g}{n_1 + \cdots + n_g}} \frac{a^{n_1 + \cdots + n_g}}{(n_1 + \cdots + n_g)!} p(0, \cdots, 0).$$

The system capacity is bounded below by the capacity of g independent groups $G_1, \cdots, G_g$ of servers having $c_1, \cdots, c_g$ servers, respectively, and is bounded above by the capacity of a full-access group of $c_1 + \cdots + c_g$ servers.

Applying this last result to the step-by-step graded multiples, it follows that the inherent load-balancing is more pronounced in the graded multiples with the larger number of first-choice subgroups. Also, for gradings with a fixed number (at least two) of first-choice subgroups, the inherent load-balancing increases as the numbers of trunks in these subgroups increase. (It is evident that there is no inherent load-balancing in full-access trunk groups.) These phenomena were observed in the simulations described in Section IV.

It is not particularly difficult to extend the model treated above to include the multilevel aspects of the step-by-step selectors. However, there is no indication that such results would furnish additional insight into the problem.

REFERENCES*

1. "Traffic Facilities Practices," Division D, Section 4h, American Telephone and Telegraph Co., New York, N. Y., May 1951.

* Since the history of our subject is not reflected in the public record, only four of the references are in the open literature. Copies of the other references, while not generally available, are in our possession, and arrangements can be made for investigators in the teletraffic field to examine them.

2. "Interdepartmental Dial Equipment Management; Step-by-Step," Division V, Section B, New York Telephone Co., March 1968.
3. Dustin, G. E., "Step-by-Step Systems—Interswitch Trunking Arrangements and Associated Traffic Engineering Considerations," unpublished work, Bell Telephone Laboratories, 1961.
4. Mina, R. R., "Gradings," lecture given at Advanced Telephone Traffic Engineering Conference, Michigan State University, July 31, 1968.
5. Wilkinson, R. I., "Theories for Toll Traffic Engineering in the U.S.A.," B.S.T.J., *35,* No. 2 (March 1956), pp. 421–514.
6. Neal, S. R., "An Extension of the Equivalent Random Theory with Application to the Analysis of Graded Multiples Having Nonrandom Offered Traffic," unpublished work, Bell Telephone Laboratories, July 14, 1969.
7. Wilkinson, R. I., "Some Recent Developments in Trunking Theory," unpublished work, Bell Telephone Laboratories, March 9, 1965.
8. "Switching Systems," Manual for Course B-328, Communications Development Training Program, Bell Telephone Laboratories, 1964.
9. Wilkinson, R. I., "The Interconnection of Telephone Systems—Graded Multiples," B.S.T.J., *10,* No. 4 (October 1931), pp. 531–564.
10. Molina, E. C., "The Theory of Probabilities Applied to Telephone Trunking Problems," B.S.T.J., *1,* No. 2 (November 1922), pp. 69–81.
11. Coggins, P. P., "Theory of Probability Applied to Telephone Problems," New York Telephone Co., Out-of-Hours Course Notes, 1927–1928.