Heuristic Remarks and Mathematical Problems Regarding the Theory of Connecting Systems

By V. E. BENEŠ

(Manuscript received February 1, 1962)

A connecting system consists of a set of terminals, a control unit for processing call information, and a connecting network. Together, these three elements provide communication, e.g., supply telephone service, among the various terminals. In this paper we present a comprehensive view of the theory of connecting systems, an appraisal of its current status, and some suggestions for further progress.

The existing probabilistic theory is reviewed and criticized. The basic features of connecting systems, such as structure, random behavior, complexity, and performance, are discussed in a nontechnical way, and the chief difficulties that beset the construction of a theory of traffic in large systems are described. It is then pointed out that despite their great complexity, connecting systems have a definite structure which can be very useful in analyzing their performance. A natural division of the subject into combinatory, probabilistic, and variational problems is drawn, and is illustrated by discussing a simple problem of each type in detail.

I. INTRODUCTION

Mass communication long ago spread beyond the manual central office and assumed a nationwide character; it is presently becoming world-wide in extent. Many of the world's telephones already form the terminals of one enormous switching system. The scale, cost, and importance of the system make imperative a comprehensive theoretical understanding of such global systems.

Nevertheless, a lack of knowledge about the combinatory and probabilistic properties of large switching systems is still a major lacuna in the art of mass communication. It is a fact of experience that each time a new switching system is planned, its designers ask once again some of

the perennial unanswered questions about connecting network design and system operation: How does one compute the probabilities of loss and of delay? What method of routing is best? What features make some networks more efficient than others? Etc.

The present paper is an informal discussion of problems in the theory of traffic flow and congestion in connecting systems (called traffic theory, or congestion theory, for short). The comments to be made are prefatory, tutorial, and illustrative. They are intended as background for several papers of a more technical nature; one of these papers[1] appears in this issue, and the remaining three[2,3,4] are to appear later. In these papers, topics touched on in the present work are considered in greater depth and detail. Together, the papers are an attempt to describe a comprehensive point of view towards the subject of connecting systems. I believe that this point of view will be useful in constructing a general theory of connecting networks and switching systems. What follows is then in part a prospectus for research to be reported on in the future.

My concern in this paper is with some of the physical bases and principal problems, with the fundamentals and difficulties, of the subject. I wish to emphasize some important properties and distinctions on which a systematic approach may be based. I am making a plea for a much more general, abstract, and systematic approach to large-scale congestion problems than has been envisaged heretofore.

Naturally, it is impossible to explore all the consequences of such a comprehensive approach in one paper; I do not pretend to have solved even some of the basic problems of the theory. I am only saying "Look, perhaps these observations will help provide a general approach."

Examples and simple problems appear in the text as illustrations of the principal points made. For tutorial purposes, I have chosen particularly simple and clear illustrations, which may seem trivial to cognoscenti of traffic theory. Nevertheless, it has been my experience in talking with engineers that the comprehensive view here presented is sufficiently new to warrant clear, simple examples. More complex problems do not belong in an introductory work; they are to appear in later papers.

II. SUMMARY

In Section III we give a historical sketch of traffic theory, which is followed by a critique of existing theories in Section IV. The general properties of switching systems are discussed in Section V. The performance of switching systems and desiderata for a theory of congestion are considered in Section VI and Section VII, respectively. Sections V to VII are heuristic and nonmathematical in character. Mathematical

models are considered in a general way in Section VIII, while Section IX concerns itself with some of the basic difficulties and questions that arise in constructing a theory of traffic in a large-scale system.

In Sections X and XI we show that, despite their great complexity, connecting systems actually have a definite structure which can be very useful in analyzing their performance. This usefulness is exemplified by four specific instances in Section XII. In Section XIII we make a general division of the subject into combinatory, probabilistic, and variational problems. The remaining sections, Sections XIV to XVI, are devoted to illustrating this division by working out a simple problem of each type in full detail.

III. HISTORICAL SKETCH

We shall not attempt to canvass systematically the literature of congestion theory. For the interested reader, the best single theoretical reference on the theory of probability in connecting systems is undoubtedly the treatise of R. Syski;[5] the historical development of the subject has been described in papers by L. Kosten[6] and R. I. Wilkinson.[7] Nevertheless, we include a brief account of previous work in order to substantiate our critique (Section IV) of present theories of traffic in connecting systems.

The first contributions to traffic theory appeared almost simultaneously in Europe and in the United States, during the early years of the 20th century. In America, G. T. Blood of the American Telephone and Telegraph Company had observed as early as 1898 a close agreement between the terms of a binomial expansion and the results of observations on the distribution of busy calls.* In 1903, M. C. Rorty used the normal approximation to the binomial distribution in a theoretical attack on trunking problems, and in 1908 E. C. Molina improved Rorty's work by his[8] (or Poisson's) approximation to the binomial distribution.

In Europe, the Danish mathematician A. K. Erlang, from 1909 to 1918, laid the foundations of the first dynamic theory of telephone traffic, which is in general use today. Perhaps influenced by statistical mechanics, Erlang introduced the notion of statistical equilibrium, and used it as a theoretical basis for deriving his now well-known loss and delay formulae. An account of Erlang's work is given by Jensen.[9]

From 1918 to 1939 traffic theory developed in many directions that are (on retrospect) closely allied to specific problems that arose in the design of the automatic telephone systems that were coming into use, and in

* Blood's unrecorded work was reported by E. C. Molina and described by R. I. Wilkinson.[7]

related queueing systems. We mention only a few topics: T. Engset[10] introduced the notion of a finite number of sources of traffic, G. F. O'Dell[11] published a classical paper on gradings, C. D. Crommelin[12] studied constant holding-time delay systems with many servers, E. C. Molina[13] made contributions to trunking theory. F. Pollaczek[14] and A. I. Khinchin[15] studied the queue with one server, and derived the delay distribution that bears their linked names. Pollaczek has also solved single-handedly many other difficult loss and delay problems. All these important contributions are concerned with congestion in specific parts of connecting systems. During this period, T. C. Fry wrote the first systematic and comprehensive book[16] on applied probability; this book devoted a chapter to telephone traffic, and appeared in 1928.

Between 1939 and 1948 there developed an increasing awareness (among workers in traffic theory) that the mathematical bases of traffic theory were closely related to the modern theory of stochastic processes initiated by A. N. Kolmogorov[17] in 1933. In particular, Erlang's idea of statistical equilibrium was identified with the stationary measure of a Markov process (or more generally with a semigroup of transition probability operators). Also, C. Palm[18] stressed the importance of recurrent processes, and W. Feller[19] that of birth-and-death processes, to traffic theory. However, particular problems continued to form the bulk of the new literature. Palm[18] made a penetrating theoretical analysis of traffic fluctuations, and L. Kosten studied such topics as retrials for lost calls,[20] and error in measurements of loss probability.[21]

The introduction of crossbar switching and common control of connecting networks in 1938 (see Ref. 22) was accompanied by a new kind of problem: calculating the loss due to *mismatching of available links* (rather than to unavailability of trunks). The first comprehensive treatment of loss in such systems was given by C. Jacobaeus[23]; his theory is adequate for practical purposes, but is based on assumed *a priori* distributions for the state of the system. R. Fortet[24] has also made contributions to this topic in the spirit of Jacobaeus' approach. A less satisfactory method for the same problems based only on the possible paths for a call has been developed (independently) by C. Y. Lee[25] and P. Le Gall.[26]

The statistical equilibrium approach to congestion in crossbar systems is rendered extremely arduous by the large number of possible states. The difficulties in this method have been faced with some success by K. Lundkvist[27] and A. Elldin[28]. However, no practically feasible approach exists at present that simultaneously includes both the concept of statistical equilibrium and the structure of the connecting network. *A fortiori*, no approach exists that also includes the effect of the common control equipment that places calls in the network.

IV. CRITIQUE

In comparison with the highly sophisticated communications systems that are being built, the models and assumptions on which theoretical studies are based are often crude and fragmentary, almost more indicative of our ignorance than of the properties of systems. It may be argued that such a harsh appraisal of the condition of traffic theory is unjustified, and is disproved by the practical successes of current engineering methods. However, it is not the efficacy of these methods, but their theoretical basis and scope, that we are questioning. Who knows to what extent present systems are "overdesigned"?

To be sure, measures of performance, loss and delay formulas, and routing methods are in daily use. Still, only in very special cases have they been investigated, let alone analyzed and understood in the full context of the system to which they are applied. Although the published literature on telephone traffic alone is vast, and many models and problems have been considered, the existing theories tend to be incomplete and oversimplified, applicable to at most a small portion of a system. Useful comprehensive models are needed; to date, only individual pieces of systems have been treated with theoretical justice. As R. Syski remarks on p. 611 of Ref. 5: "At the present stage of development . . . the theoretical analysis of the [telephone] exchange as a whole has not been attempted." The general theory of switching systems now consists of some apparently unrelated theorems, hundreds of models and formulas for relatively simple parts of systems, and much practical lore associated with specific systems. It will stay in this condition until sufficient theoretical underpinning is provided to unify the subject. We believe that this sad "state of the theory" is due largely to these three factors:

(*i*) The large scale, and consequent inherent difficulty of the problems.

(*ii*) The absence of a widely accepted framework of concepts in which problems could be couched and solved.

(*iii*) The lack of emphasis on and success with the combinatorial aspects of the problems.

More generally, many of the basic mathematical properties of connecting networks and switching systems have either never been studied, or, if studied, have not been digested, advertised, and disseminated for engineering use. As a result, the design and complexity of systems has consistently run ahead of the analysis of their performance.

V. GENERAL PROPERTIES OF CONNECTING SYSTEMS

We start by discussing some universal properties of connecting systems from the point of view of congestion, without reference to definite mathe-

matical models for their operation. Specifically, we describe, in a nontechnical way, (*i*) the general nature and outstanding features of connecting systems, (*ii*) the principal kinds of congestion that interest engineers, and (*iii*) some of the difficulties and desiderata in both the theory and practice of large-scale switching. No mathematical abstractions are used at first. Some observations made may seem obvious or trivial; nevertheless, they are necessary for the general understanding that we desire. On these observations, we shall base a systematic division of the theory into three kinds of problems, *combinatory*, *probabilistic*, and *variational*.

By a *connecting system* we shall mean a physical communication system consisting of (*i*) a set of terminals, (*ii*) control units which process requests for connection (usually between pairs of terminals), and (*iii*) a connecting network through which the connections are effected. The system is to be conceived as operating in the following manner: (1) calls (or requests for connection) between pairs of idle terminals arise; (2) requests are processed by a control unit, and desired connections are completed, if possible, in the connecting network; (3) calls exist in the network until communication ends; (4) terminals return to an idle condition when a call terminates. (Naturally, the arising requests may "defect" at any point during the process of connection.)

The gross structure of a connecting system is depicted in Fig. 1. Most modern connecting systems follow this basic pattern. Particularly important examples are telephone central offices, toll centers, telegraph networks, teletypewriter systems, and the many military communications systems.

All the examples cited share three important properties. These are (*i*) great combinatorial *complexity*, (*ii*) definite geometrical or other *structure*, and (*iii*) *randomness* of many of the events in the operating system.

It is obvious that many connecting systems are highly complicated.

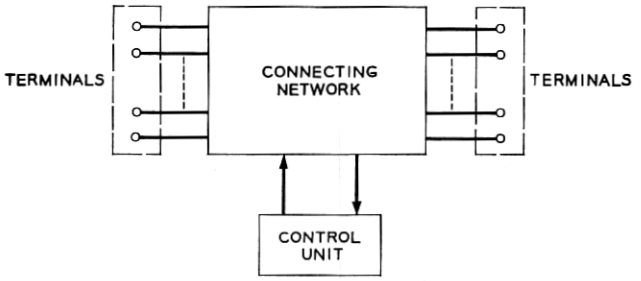

Fig. 1 — Connecting system.

Both the control unit and the connecting network contain thousands of parts which may (together) assume millions of combinations. That is, the system can be in any one of millions of possible "states." These numbers are increased when several switching centers are considered together as a unit, as in toll switching. Our purpose in calling attention to this complexity is to suggest that it calls for theoretical methods that, like those of statistical mechanics, are especially designed to distill important facts from masses of detail.

It is less often realized, however, that this complexity is accompanied by definite mathematical structure, and is frequently alleviated by many symmetries. The control unit and the connecting network always have a specific combinatory, geometric, and topological character, on which the performance of the system closely depends.

By imputing randomness to the systems of interest we do not imply that their operation is unpredictable; we mean only that the best way of describing this operation is by use of probability theory. It is not practical, even though it might be possible in principle, to predict the operation of a switching system by means of differential equations in the way that the flight of a rocket is predicted. However, differential equations have been used for many years to describe, not the motion of an actual system, but the changes in the *likelihoods* or *probabilities* of its possible states. Such equations govern the flow or change of probabilities and averages associated with the system, not the detailed time behavior of the system itself. It is in this weaker sense of assigning likelihood to various events that we can predict the behavior of switching systems, a fact first emphasized by A. K. Erlang's pioneering work on telephone traffic.[9] For instance, certain features (such as average loads offered and carried) of telephone traffic that are predictable in this weaker sense form the basis on which toll trunking routes are engineered.

We now turn to examples of the structure of connecting networks and of control units. The basic features of the connecting network for the No. 5 crossbar system are shown in a simplified form in Fig. 2. The network has two sides, one for subscribers' lines and the other for trunks. Small squares represent rectangular *crossbar switches*, capable of connecting any inlet terminal to any outlet terminal. These switches are arranged in groups called *frames*, either line link frames for subscribers' lines, or (on the other side) trunk line frames for trunks. Frames are indicated in Fig. 2 by large dashed squares enclosing four small squares; dots indicate repetition. The pattern of links which interconnect the switches is shown by solid lines between small squares. At most one link connects any pair of switches.

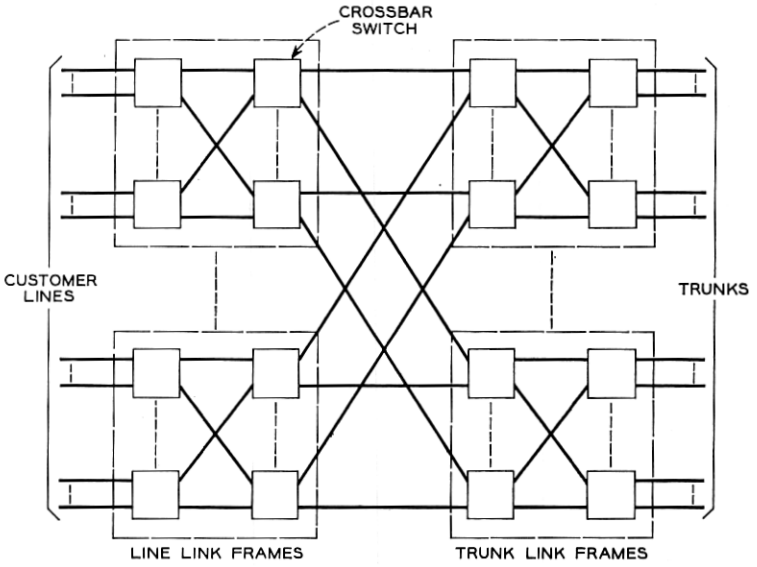

Fig. 2 — Basic No. 5 crossbar network.

As a second example of a connecting network, consider the three-stage Clos network (see Ref. 29) depicted in Fig. 3. The interpretation of this figure is the same as that of Fig. 2: small squares stand for crossbar switches, and lines between them represent links. Each call can be put into the network in m ways, one for each of the m switches in the middle column. This network has the property that if $m \geqq 2n - 1$, it is non-blocking.

A control unit consists of parts that are arranged in a manner reflecting their function, and are determined by the operations necessary to establish a connection, and by the philosophy of design and the tech-

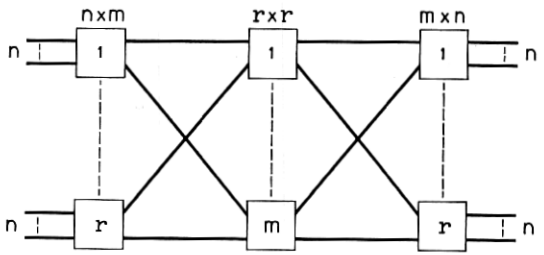

Fig. 3 — Clos three-stage network.

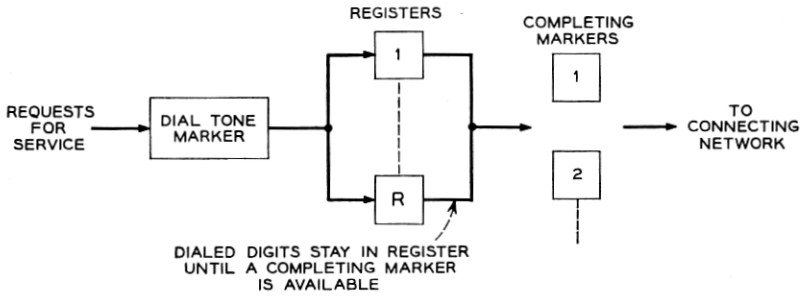

Fig. 4 — Simple control unit.

nology that are basic to the system. To establish a connection, the control unit must do some or all of the following: (*i*) identify the calling party or terminal, (*ii*) find out who the called party is, and (*iii*) complete the connection. Three examples will be considered, in order of increasing complexity and modernity.

A simple example of the structure of a control unit is given in Fig. 4. The unit consists of a dial-tone marker which assigns and connects available idle registers to subscribers for dialing. The dialed digits remain in the register until a completing marker (one of possibly several) removes them and uses them to complete the call. The calls, or requests for connection, may be thought of as arriving from the left, and proceeding through the diagram from left to right. There may be a delay in obtaining dial tone, a delay in securing the services of a completing marker, or a circuit-busy delay (or rejection) in the network. It should be observed that the switching equipment necessary for connecting subscribers to registers, or registers to completing markers, is left out of account in this model.

A second example is obtained from the first by inserting a buffer memory between the registers and the markers as shown in Fig. 5. (One

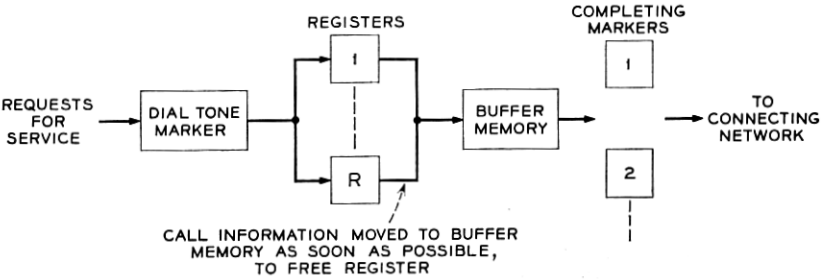

Fig. 5 — Control unit with buffer memory.

can argue that registers are expensive special-purpose units and should not be used for storing call information when cheap memory is available.) When dialing is finished, the call information is forthwith transferred to the buffer memory, there to wait for a completing marker without preempting a register. The markers and registers are now effectively isolated, so that delays in completing calls do not cause delays in obtaining dial tone. Again, traffic is viewed as moving from left to right.

The high speeds possible with electronic circuits have led to new configurations and problems (for control units and networks) which have not yet received much attention in congestion theory. Although it performs the same functions, the control unit of a modern electronic central office usually has an organization differing from that of the examples of Figs. 4 and 5, which are characteristic of electromechanical systems. Four principal reasons for this contrast are:

(*i*) The electronic office relies heavily on a large digital memory to aid in processing calls and (in time division systems) to keep track of calls in progress; electromechanical systems, on the other hand, are based largely on "wired-in" memory.

(*ii*) In the electronic office, processing a given call usually requires several consultations of the digital memory; thus, the flow of traffic in the control unit is re-entrant and not unidirectional as in Figs. 4 and 5.

(*iii*) The speed of electronic components often makes it possible to perform only one operation at a time; thus, a single unit may be (alternately) part of a dial-tone marker, part of a register, part of a completing marker, etc., depending on the details of organization of the control unit.

(*iv*) The replacement of "wired-in" memory, whose stored information is immediately available, by an electronic memory which has to be consulted, creates problems analogous to the problem of connecting completing markers to registers in the No. 5 crossbar system: special access units are needed. Subunits of the control unit, such as dial-tone markers, completing markers, senders, etc., must take turns in using the access circuit to the digital memory.

Fig. 6 depicts a (hypothetical) control unit for an electronic switching system built entirely around a memory which stores all information on the current status of calls. The control unit consists of various special-purpose units such as a sender, a receiver, a completing marker, a dial-tone marker, and registers. Each of the listed units can operate independently of and simultaneously with the others; however, they compete for (take turns at, possibly with priorities) the access circuit to the memory. Each unit depends on the memory to give it a new assignment, to file the results of the last one, or both. Every operation of a special-

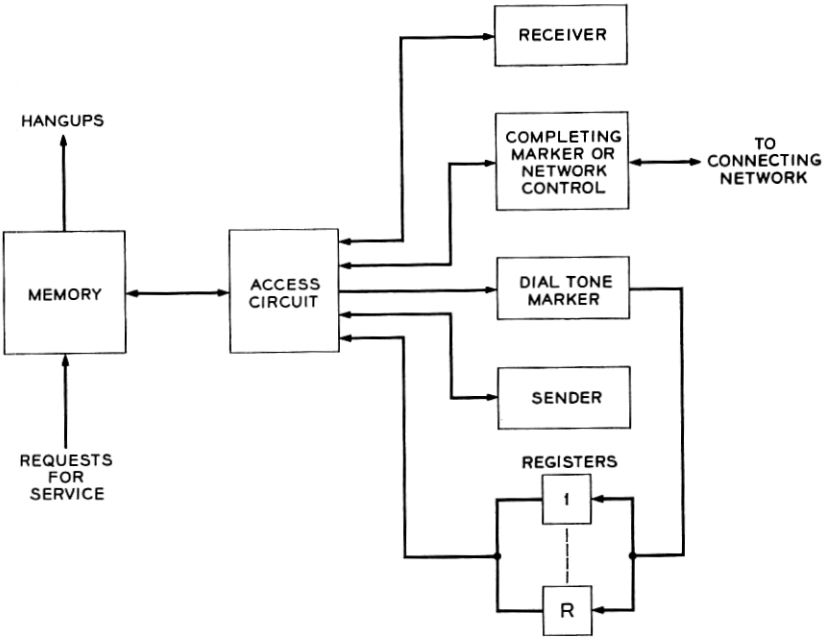

Fig. 6 — Block diagram of electronic control unit.

purpose unit requires access to the memory, either to obtain data from it, or to file data in it, or both. The memory contains several classes of calls: those waiting for dial tone, those waiting for a completing marker, those actually in progress in the connecting network, etc.

VI. PERFORMANCE OF SWITCHING SYSTEMS

In general, the gross or average features of switching systems are both more accurately predictable and more economically important than the specific details. The average load carried by a trunk group is usually more easily predicted than the condition of a particular trunk; and the "all trunks busy" condition of the group is of greater concern to the telephone administration than the busy condition of a single trunk.

From the point of view of economics and traffic engineering, only certain average features of the behavior of a system (used as measures of performance) are important. These few quantities of interest depend on the multitude of details of "fine structure" in the control unit and the connecting network. Although the intricate details give rise to the important averages, the details themselves are of relatively little interest.

In the rest of this paper, we shall repeatedly contrast the few average quantities that are of engineering interest with the many millions of detailed features and properties (of connecting systems) on which the averages are based. The central problem in the theory of connecting systems is to understand how the interesting quantities arise from the details, and to calculate them.

We shall start our discussion of the contrasting roles of averaged features and details by considering some of the different kinds of congestion that interest engineers, and in addition some associated measures for the performance of systems.

Congestion is said to occur in a connecting system when a requested connection cannot be completed immediately. By "immediately" we mean, of course, not "instantaneously", but "as fast as control equipment, assumed available, can do its work". The time it takes to complete a call contributes to congestion only if it keeps other calls from being completed at the normal rate. That a call cannot be completed immediately (in this sense) may be due to facts of three kinds: (*i*) certain necessary units of switching equipment (like trunks, or markers) are all busy; (*ii*) there are available units, but they occur in an unusable combination, or "fail to match"; (*iii*) congestion has occurred previously, and other requests are awaiting completion.

In telephone traffic theory, requests for connections which encounter congestion are traditionally termed *lost calls*. This terminology is used whether the request is refused (and never completed), or merely delayed (and completed later). Switching systems differ in the *disposition* of lost calls, i.e., in what is done with requests which encounter congestion. There are in theory two principal ways of disposing of lost calls. In the first way, termed "lost calls cleared", the request is denied and leaves the system; this way of dealing with lost calls naturally gives rise to the *proportion of requests denied*, or the probability of blocking or loss, as a measure of performance. The second way of disposing of lost calls is termed "lost calls delayed", and consists in delaying the request until equipment becomes available for completing the connection; associated with this is the *probability of delay* in excess of a specified time t, as a measure of performance.

On the simplified account of the last paragraph we must impose at least two qualifications. First, whether a request suffers blocking or delay (or both!) may depend on the condition of the system at times shortly after the request is made; second, the completion of a request usually involves a sequence of steps, any one of which may expose the request to delay or loss. For example, a request may encounter delay in obtaining

dial tone, delay in securing the services of a completing marker, and delay or blocking in the attempted completion of the desired connection through the connecting network.

We conclude this section by briefly considering what general features of connecting systems are particularly relevant to their performance as measured (for example) by probabilities of blocking or delay, or by average loads carried, offered, or both. Now, a connecting system has two principal parts, the control unit and the connecting network; the features of the system that are relevant to performance are conveniently distinguished according to whether they are features of the control or of the network. This distinction is fundamental because the performance of the control is largely determined by the speed and number of the various sub-units comprising it, while the performance of the network is largely dependent on what combinations of calls can be in progress simultaneously.

The control unit is basically a data processing system: it collects information about desired connections, digests it, makes routing decisions, and issues orders for completing requested calls in the connecting network. Its capacity is measured, e.g., by the number of customers who can be dialing simultaneously, or by the number of calls which are being completed in the network at the same time. Its performance is described by the probability distributions of delay before receiving dial tone, and of delay after completion of dialing until the desired connection is completed.

For a simple model of a control unit (such as depicted in Fig. 4), the features pertinent to performance are: (*i*) the calling rate, (*ii*) the number of registers for dialing, and (*iii*) the speed and number of completing markers. In the case of the prototype electronic control unit (depicted in Fig. 6) some additional features appear: (*iv*) the speed of the access circuit to the memory, (*v*) the order of priority of the functions being performed, the discipline of access to various services, and the competition for access among marker, dial tone marker, sender, etc., (*vi*) the presence of re-entrant traffic (every call must "use" the access circuit at least twice), and (*vii*) the number and arrangement of the various functions which are going on simultaneously.

The connecting network, in contrast to the control unit, determines what calls can be in progress, rather than how fast they can be put up. Its configuration determines what combinations of terminals can be connected simultaneously together. For example, if $m \geqq n$, the Clos network of Fig. 3 has the property of *rearrangeability*: any preassigned set of calls can be simultaneously connected. The No. 5 network of Fig. 2 does not

have this property: the number of calls between a line link frame and a trunk link frame is limited by the number of links between those two frames. Such combinatory properties of the structure of the connecting network play a determining role in estimating the cost and the performance (probability of blocking) of the network. If the structure is too simple, very few calls can be in progress at a given time and blocking is high; if it is extensive and complex, it may indeed provide for many large groups of simultaneous calls in progress, and so a low probability of blocking, but the network itself may be prohibitively expensive to build and to control.

VII. DESIDERATA

Our discussion of the three prominent features of switching systems — (*i*) great complexity, (*ii*) definite structure, and (*iii*) randomness — has exposed or suggested some of the problems and desiderata which a theory of congestion in large-scale systems must (respectively) encounter and supply. Specific statements of requirements and tasks are now given.

General desiderata can be obtained by examining the purpose served by a theory of congestion. The function of such a theory is twofold: it is (*i*) *to describe* the operation of switching systems, and (*ii*) *to predict* the performance of systems. More specifically, the descriptive function (*i*) is to provide a theoretical framework into which any system can be fitted, and which permits one to evaluate the performance of the system, e.g., to compute the chance of loss, to estimate a sampling error, or to prove a network nonblocking. The predictive function (*ii*) has logically the same structure as (*i*), but emphasizes the use of theory to make future capital out of past experience, to extrapolate behavior and thus to guide engineering practice.

More specific tasks than these appear when we list some of the activities comprised by the theory and practice of traffic engineering. A possible list is as follows:

i. Describing and analyzing mathematical models.
ii. Computing measures of performance for specific models.
iii. Studying the accuracy of traffic measurements, the effects of transients, and problems explicitly involving random behavior in time.
iv. Comparing networks, control systems, methods of routing, etc.
v. Using traffic data to verify empirically the assumptions of theories.
vi. Making predictions and estimates for engineering use.

On the basis of this list, and of our previous discussions of complexity,

randomness, gross features, and details, we can say that a satisfactory theory of congestion must meet the following requirements:

i. It must be sufficiently general to apply to any system.

ii. It must yield computational procedures for system evaluation and prediction of performance, based on masses of detail. These procedures must be at once feasible and sufficiently accurate, and if approximations are made, their effect must be analyzable.

iii. It must encompass all the three basic elements simultaneously, viz., the random traffic, the control unit, and the connecting network.

VIII. MATHEMATICAL MODELS

We shall now consider what mathematical structures are appropriate theoretical descriptions of operating connecting systems. The discussion will provide an intuitive picture of an operating system, and will help to motivate a natural division of our subject into *combinatory*, *probabilistic*, and *variational* problems.

By a *state* we shall mean a partial or complete description of the condition (of the system under study) in point of (*i*) busy or idle network links, crosspoints, and terminals, and (*ii*) idle or busy control units or parts thereof. Complete, highly detailed descriptions correspond to fine-grained states specified by the condition of every crosspoint, link, or other unit in the system, in absolute detail. Incomplete descriptions correspond to coarse-grained states, or to equivalence classes of fine-grained states.

During operation, the connecting system can pass through any permitted sequence of its states. Each time a new call arises, or some phase of the processing of a call by the control unit is finished, or a call ends, the system changes its fine-grained state. These changes do not usually occur at predetermined epochs of time, nor in any prescribed sequence; they take place more or less at random. At any particular time, it is likely that some terminals, links, and parts of the control unit are idle, that various requested calls are being processed, and that certain calls are in progress in the connecting network.

The last paragraph suggests the following intuitive account of an operating switching system: it is a kind of dynamical system that describes a random trajectory in a set of states. Such an intuitive notion can be made mathematically precise in many ways. Any one precise version is a *mathematical model* for the operation of the switching system. In constructing such a model, it is neither necessary nor desirable always to use the most detailed (the fine-grained, or microscopic) states; often a partial

description in terms of coarse-grained states suffices, and is less difficult to study. Indeed, in building a model it is to some extent possible to choose the set of states to suit special purposes. One can, for instance, control the amount of information included in the state so as to strike a balance between excessive detail and insufficient attention to relevant factors. It is possible to make the notion of state more or less complete so as to achieve certain (desired) mathematical properties (such as the Markov property, or a suitable combinatory structure) which simplify the analysis of the random trajectory. Finally, one can add supplementary variables analogous to counter readings or cumulative measurements, and obtain their statistical properties.

The abstract entity appropriate for describing the random behavior of a switching system is a *stochastic process*. For our present heuristic purposes, we can define a stochastic process as follows: by a *possible history* of the system we mean a function of time taking values in the chosen set of states; a stochastic process is then a collection Ω of possible histories of the system in time, with the property that many (presumably interesting) subsets A of Ω have numerical probabilities $\Pr\{A\}$ associated with them. The probability $\Pr\{A\}$ of the set A of possible histories is interpreted as the chance or likelihood that the actual history of the system be one of the histories from the set A. Models of this kind furnish information because desired quantities can be calculated from the basic probabilities $\Pr\{A\}$.

IX. FUNDAMENTAL DIFFICULTIES AND QUESTIONS

The systematic use of mathematical models (such as stochastic processes) in congestion theory and engineering has been largely limited to small pieces of systems like single-server queues, groups of trunks with full access, etc. More complex models of systems involving connecting networks have hardly been touched by theory. This limitation has been due almost entirely to the large number of states such models require, and to the complex structure of the transitions (changes of state) that can occur. In short, the essential characteristics (of large-scale connecting systems) themselves generate the basic difficulties of the theory.

In most congestion problems, it is easy enough to construct (say) a Markov process that is a probabilistic model of the system of interest. But it is difficult, because of the large number of states and the complexity of the structure, to obtain either analytic results or fast, reliable simulation procedures. This circumstance has been a major obstacle to progress in the congestion theory of large systems. One of its consequences has been that in some cases, models known to be poor repre-

sentations of systems have been used merely because they were mathematically amenable, and no other tractable models were available. Even overlooking such extremes, it is fair to state that, to date, problems of analysis and computation have limited the amount of detail embodied in the notion of state for models of switching systems. Every effort has been made to keep the number of states in models small, and their complexity low.

Having exposed some basic properties of and theoretical problems arising from congestion in connecting systems, let us acknowledge that an operating, large-scale connecting system cannot be done full theoretical justice except by a stochastic model with an astronomical number of states and a very complicated structure of possible transitions. At this point, let us try to take a synoptic view of the subject, and ask some general questions whose discussion might indicate new approaches and emphases. Let us, in the current idiom, lean back in our chairs, make a (n) (agonizing?) reappraisal, and draw ourselves the "big picture."*

The following three questions seem (to this writer) to be pertinent, and are taken up in the next sections:

i. What is the value of mathematical models that have a very detailed notion of state?

ii. Is it possible to make explicit theoretical use of the very properties of connecting systems that appear to be most troublesome? How can the two principal difficulties (large number of states, complex structure of changes) be turned into positive advantages?

iii. What features of connecting systems are especially relevant to the mathematical analysis of system operation?

We do not pretend to provide iron-clad answers to these questions. We try to give a helpful discussion of relevant matters, illustrated by examples.

X. THE MERITS OF MICROSCOPIC STATES

We have raised the question: To what extent can detailed probabilistic models of the minutiae of operating switching systems (i.e., models with "microscopic" states) improve our understanding of these systems, and so our ability to engineer them? Against the value of such detailed models it can be argued that for engineering purposes only certain performance data are of interest, and that the detailed model produces a vast amount of information with no apparent practical method for reducing this information to probabilities of delay or blocking.

* Supplying those clichés whose substitution leaves the content of this last sentence invariant is left as an exercise for the reader.

Since the usefulness of mathematical models depends entirely on the desired information they can be forced to yield, it is not reasonable to dismiss detailed models *a priori*. For in truth, few if any such models have been considered, and it has not been shown that they are useless in the sense that no practical method for extracting useful quantities from these models exists.

To be sure, the congestion engineer is not as concerned with the minutiae themselves as with their effect *en masse*. But he has to base his conclusions and recommendations in *some* way on the total effects of a large number of individually trivial events. Hence, at some point in his procedure, he must take account of the large number of states and the complex structure of possible transitions of his system.

Traffic engineering practice is based on (relatively few) probabilities and averages, such as average loads, deviations about them, and blocking or delay probabilities. Any reliable theoretical estimate of these averages must be based on the combinatory and probabilistic properties of a theoretical model (stochastic process) for system operation. At worst, an approach or model that provides detailed information might yield a much-needed check point for the methods that are in current engineering use, and so increase the engineer's understanding of and confidence in these methods.

However, there is a much more general, positive sense in which attention to the details of connecting systems can contribute to theoretical progress. This is taken up in the next section.

XI. FROM DETAILS TO STRUCTURE

The prospect of solving (say) statistical equilibrium equations for models with a very detailed notion of state is discouraging indeed, although it has been faced, notably by Elldin[28] in Sweden. Nevertheless, a sanguine and useful approach (along this line) to connecting systems can be obtained by a shift of emphasis from "details" to "structure." We have emphasized that describing an operating connecting system means keeping track of numerous details, none of which is interesting in itself. We have said that the operation of such a system could be pictured as a trajectory in a very complicated set of states. We now claim that the inclusion of enough details (in the notion of state for a model) gives the set of possible states a *definite structure* that is useful because it makes possible or simplifies the analysis of the probabilistic model.

Whatever may be the value of detailed probabilistic knowledge for the immediate problems of engineering, such knowledge is useful if not essential in theoretical studies. By using a highly detailed, "microscopic"

description for the state of the system, it is possible to exploit the extensive mathematical structure (properties) that such a set of states naturally has. Indeed, the combinatory properties and geometrical structure of the set of states are two of the very few weapons available for attacking large-scale problems of traffic theory. I believe that in the past these properties and this structure have not been sufficiently exploited. They can only be put to use by a systematic application of "microscopic" states.

The three basic properties of switching systems discussed in Section V were (*i*) extreme combinatory complexity, (*ii*) definite geometrical structure, and (*iii*) randomness. The preceding paragraphs of this section can be related systematically to these properties, and elaborated into a sort of program: Instead of throwing up our hands at (*i*) in trying to do justice to (*iii*), we should realize that a detailed notion of state allows us to turn (*ii*) to our advantage in studying (*iii*). Let us then disregard the fact that there are many states, and analyze the structure of possible changes of state, to see how to capitalize on it.

For, indeed, the possible microscopic states of a particular connecting system are not arbitrary. They are rigidly determined by the combinatory and topological properties of the connecting network, and by the organization of the control unit. Such a set of possible states has a mathematical structure of its own, and this structure is relevant to the performance of the system, and to any stochastic process that represents its operation.

It can be seen quite generally that when a switching system changes its microscopic state, it can only go to a new state chosen from among a few "neighbors" of the state it is leaving. These neighbors comprise the states which can be reached from the given state by starting a new call, ending an existing call, or completing some operation in the control unit. In a large system, a state may have many such neighbors, but they will be few in comparison with the total number of microscopic states.

A striking and useful example of how details give rise to structure can be obtained by considering the possible states of a connecting network. These states can be arranged in a pattern as follows: At the bottom of the pattern we put the zero or ground state in which no calls are in progress; above this state, in a horizontal row, we place all the states which consist of exactly one call; continuing in this way, we stack up level after level of states, the kth level L_k consisting of all the states with k calls in progress.

We now construct a graph by drawing lines between states that differ from each other by exactly one call. (Such states, needless to say, are

always in successive levels of our diagram.) This graph we call the *state-diagram*. It is a natural (and standard) representation of the partial ordering $\leqq$ of the states: where x and y are states,

$$x \leqq y$$

means that y can be obtained from x by adding zero or more calls to x, or alternately, that x can be got from y by removing zero or more calls. The importance of this state-diagram lies in two facts:

i. The state diagram gives a geometrical representation of the possible states of the system. The myriad choking "details" of the connecting network have been converted into a vast geometrical structure with special properties. The operating system describes a trajectory through the state diagram, moving between levels as calls begin and end.

ii. Any stochastic process describing the operation of the connecting network is a point moving randomly on the state diagram. The motion is only between adjacent levels. New calls put into the network correspond to jumps to the next higher level; hangups correspond to jumps to the next lower level.

As a simple example, we consider the possible states of a single 2 by 2 switch. These consist of (*i*) the zero state, (*ii*) the four ways of having one call up, and (*iii*) the two ways of having two calls up. These states are depicted in Fig. 7. Fig. 8 shows the states of a 2 by 3 switch.

XII. THE RELEVANCE OF COMBINATORY AND STRUCTURAL PROPERTIES: EXAMPLES

In this section we elaborate, by discussing examples, our theme that the combinatory and structural properties of connecting systems are of

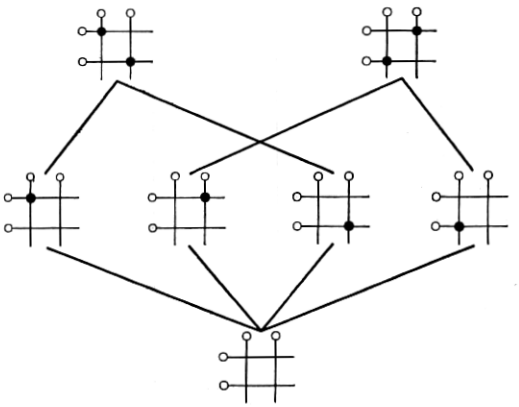

Fig. 7 — States of a 2 by 2 switch.

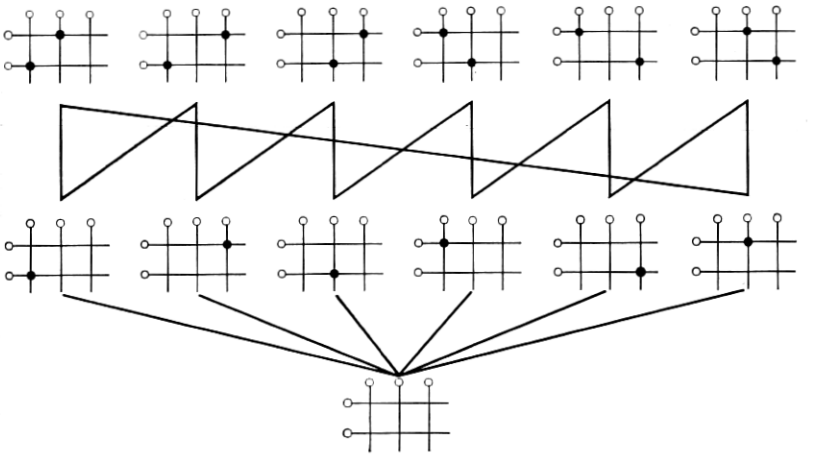

Fig. 8 — States of a 2 by 3 switch.

the greatest import (*i*) to their performance, and (*ii*) to the analysis of mathematical models of their operation. The organization of the control unit and the configuration of the connecting network largely determine the possible microscopic states of the system. Let us see what effects these features can have on problems of system analysis.

Example 1: Any connecting system has a "zero" or ground state in which all terminals and links are idle, no calls are being processed by the control unit, and the connecting network is empty. The existence of this zero state is a structural property common to all switching systems. This zero state seems most uninteresting. Nevertheless, many probabilistic models (for switching system operation) have the property that if the equilibrium probability of the zero state is known, then that of any other state can be determined in a simple way. Several specific examples of this phenomenon are worked out later in this paper, so none will be given here. (See Sections XV and XVI.)

Example 2: The relevance of combinatory properties of the connecting network to the calculation of probabilities can be vividly illustrated by reference to Clos' work on nonblocking networks (see Ref. 29). The blocking probability of a connecting network is the fraction of attempted calls that cannot be completed because no path for the call exists in the current state of the network. Until Clos' article appeared it was not generally known that, *no matter what probabilistic model was used*, an exact calculation of blocking probability for a Clos network with $m \geqq 2n - 1$ (see Fig. 3) would yield the value zero!*

* Zero, not zero factorial, which equals unity!

Example 3: Consider the class of connecting networks which have the property that in any state of the network, two idle terminals (forming an inlet-outlet pair) can be connected in at most one way. For each member of this class of networks we construct a Markov stochastic process to represent its operation under random traffic, as follows: in any state, if an inlet-outlet pair is idle, the conditional probability is $\lambda h + o(h)$ that it request connection in the next interval h, as $h \to 0$; also, an existing call terminates in the next interval h with a probability $h + o(h)$, as $h \to 0$; requests that encounter blocking are denied, and do not change the state of the system (lost calls cleared).

If X is a finite set, let $| X |$ be its cardinality, i.e., the number of elements of X, and let S be the set of all states of the network under discussion. For x in S, define

A_x = set of states accessible from x by adding a call

B_x = set of states accessible from x by removing a call

$| x |$ = number of calls in progress in state x

L_k = set of states with k calls in progress.

Note that $| B_x | = | x |$.

Let p_x be the stationary or equilibrium probability that the system is in state x. By reference to Fig. 9, it can be seen that the statistical equilibrium equations for our probabilistic model are

$$(\lambda | A_x | + | x |)p_x = \sum_{y \in A_x} p_y + \lambda \sum_{y \in B_x} p_y , \qquad x \in S.$$

Since in any state an idle pair can be connected in at most one way, no routing decisions need to be made, and the solution of this equation

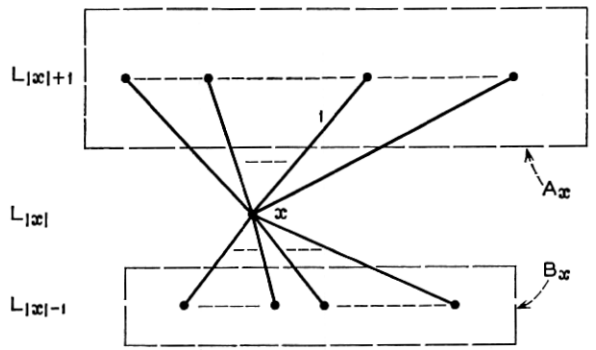

Fig. 9 — A state x, and the sets A_x, B_x in the state diagram.

(regardless of the network configuration!) is given by

$$p_x = p_0 \lambda^{|x|} \qquad x \neq 0$$

$$p_0^{-1} = 1 + \sum_{\substack{y \in S \\ y > 0}} \lambda^{|y|}$$

$$= \sum_{k \geqq 0} \lambda^k \mid L_k \mid$$

where 0 is the zero state. We have therefore shown that the simple combinatory property, that a call can be put up in at most one way, implies that the stationary probabilities of the Markov process we defined are of a simple geometric type. Note the important role played by the zero state, as discussed in Example 1.

Example 4: The Markov stochastic processes of the previous example can be used to illustrate another important point. There are many switching system models for which quantities of interest (such as the probability of blocking) can be given rigorously, without approximations, by a formula in which the distinction between system combinatorics and random customer behavior appears explicitly. In Example 3, the state probabilities $\{p_x, x \in S\}$ are completely determined by the quantities

$$\mid L_k \mid , \qquad k \geqq 0$$

i.e., by the number of states with k calls in progress, for $k \geqq 0$. For these models we can express the blocking probability as a function of the traffic parameter λ and of $\mid L_k \mid$, $k \geqq 0$. The numbers $\mid L_k \mid$ represent purely combinatory properties of the network.

The blocking probability b can be calculated as follows: b is the fraction of attempted calls that are unsuccessful, so that

$$1 - b = \frac{\text{total rate of successful attempts}}{\text{total rate of attempts}} .$$

In equilibrium, the total rate of successful attempts must equal the total rate of hang ups. The total rate of hang ups is

$$\sum_{x \in S} p_x \mid x \mid = \text{mean number of calls in progress}$$

(because the mean holding time is used as the unit of time). Let N be the number of terminals offering traffic. Since an idle inlet-outlet pair calls at a rate λ, the attempt rate in a state x is

$$\lambda \cdot (\text{number of idle pairs in a state } x) = \lambda \binom{N - 2 \mid x \mid}{2} .$$

The total rate of attempts is then

$$\lambda \sum_{x \epsilon S} p_x \binom{N - 2 \mid x \mid}{2}.$$

Hence,

$$b = 1 - \frac{\sum_{x \epsilon S} p_x \mid x \mid}{\lambda \sum_{x \epsilon S} p_x \binom{N - 2 \mid x \mid}{2}}$$

$$= 1 - \frac{\sum_{k>0}^{[N/2]} \lambda^k k \mid L_k \mid}{\lambda \sum_{k \geqq 0}^{[N/2]} \lambda^k \mid L_k \mid \binom{N - 2k}{2}},$$

where $[N/2]$ is the greatest integer less than or equal to $N/2$. This formula exhibits the blocking probability as a rational function of the calling rate λ per idle pair and as a bilinear function of the combinatory constants $\{\mid L_k \mid, k \geqq 0\}$. The degree of the denominator in λ is one more than that of the numerator, so $b \to 1$ as $\lambda \to \infty$; also, note that

$$\lim_{\lambda \to 0} b = 1 - \frac{\mid L_1 \mid}{\binom{N}{2}}.$$

This limit is greater than zero if there are calls which cannot be put up in *any* way. Finally, we observe that if the network is non-blocking, then

$$k \mid L_k \mid = \sum_{x \epsilon L_{k-1}} \binom{N - 2 \mid x \mid}{2}$$

$$= \mid L_{k-1} \mid \binom{N - 2k + 2}{2}$$

and so $b = 0$, as it should, if we interpret

$$\binom{N - 2[N/2]}{2}$$

as zero.

XIII. COMBINATORY, PROBABILISTIC, AND VARIATIONAL PROBLEMS

The preceding discussions have established that the ingredients going into a mathematical model of a connecting system are of two kinds.

On one hand are the combinatory and structural properties, and on the other, the probabilistic features of traffic. We emphasize the distinction between these aspects, and claim that by carefully drawing it, we can extend the general understanding of connecting systems, unify or modify existing theoretical methods, and obtain new engineering results.

Our discussion also suggests that to study stochastic processes that represent operating connecting systems, it is essential to have an extensive theory of the combinatory and topological nature of the microscopic states of such systems.

In any specific model of a connecting system, one can distinguish the combinatory from the stochastic features. However, it is also of interest to compare models of systems in an effort to determine optimal systems. These facts suggest a useful though imprecise division of the entire subject (of connecting system models) into three broad classes of problems. In order of priority, these are

i. Combinatory problems.
ii. Probabilistic problems.
iii. Variational problems.

This order of priority arises in a natural way: one needs to study combinatory problems in order to calculate probabilities; one needs both combinatory and stochastic information in order to design optimal systems.

The tripartite division just made provides a rational basis for organizing research effort. Since so many of our pronouncements have been generalities, we devote the remainder of the paper to illustrating carefully each of the three divisions (combinatory, probabilistic, variational) by working out and discussing in detail a very simple (yes, a trivial) problem from each division. These problems have been chosen for their tutorial value rather than their realism or usefulness. In discussing them, we place emphasis on furthering insight rather than solving practical problems, on exposing principles rather than providing engineering data.

XIV. A PACKING PROBLEM

It has long been suspected (and in some cases, verified experimentally) that routing calls through a connecting network "in the right way" can yield considerable improvements in performance. This procedure of routing the calls through the network is called "packing" (the calls), and the method used to choose routes is called a "packing rule." The use of the word "packing" in this context was surely suggested by an analogy with packing objects in a container. However,

the existence and description of packing rules that demonstrably improve performance (e.g., by minimizing the chance of blocking) are topics about which very little is known.

What, then, is the "right way" to route calls? It has been argued heuristically that it is better to route a call through the most heavily loaded part of the network that will still take the call. Appealing and simple as this rule is, nothing is known about it. We know of no published proof of either its optimality or its preferability over some other rule. The rule will be proven optimal for an example in Section XVI.

The question naturally arises, though, whether for a given network in which blocking can occur there exists a packing rule so cunning that by following it all blocking is avoided. Then, use of the rule makes the network nonblocking. Such a network may be termed *nonblocking in the wide sense*, while a network none of whose states has any blocked calls may be termed *nonblocking in the strict sense*.

The existence of such a rule is a purely combinatory property of the network, and so serves as an example of the first type of problem described in Section XIII. Unfortunately, *practically useful* connecting networks that are nonblocking in the wide sense are yet to be found. Since we are primarily interested in exemplifying principles, we shall be content with discussing an impractical network that is nonblocking in the wide sense. The example to be given was suggested by E. F. Moore.*

Let us first consider the three-stage connecting network depicted in Fig. 10. All switches in the middle column are 2 by 2, and there are $2n - 1$ of them, so, by a result of C. Clos,[29] the network is nonblocking. Suppose that we use the rule that an empty middle switch is not to be used unless there is no partially filled middle switch that will take the call. In other words, do not use a fresh middle switch unless you have to! In general, this rule is not quite the same as the one exhorting use of the heavily loaded switches wherever possible, because it only tells us what to avoid, but it is in the same spirit. In the case to be considered, however, a middle switch is either empty, half-full, or full; hence the two rules coincide.

We shall show that if this rule is used, then no more than $[3n/2]$ middle switches are *ever* used, where $[x]$ is the greatest integer less than or equal to x. Thus the rest, about one quarter of the middle switches, could be removed and no blocking would result if the rule were used. It can be verified by examples that if there are only $[3n/2]$ middle switches and the rule is violated, then calls can be blocked. Thus, the network of

* Private communication.

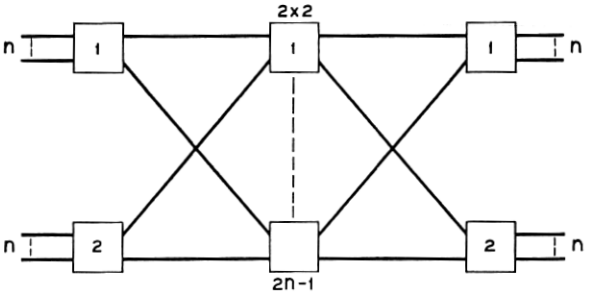

Fig. 10 — Three-stage nonblocking connecting network (Clos type).

Fig. 11 is not nonblocking in the strict sense, but is nonblocking in the wide sense.

A state x of a connecting network is called reachable (under a rule ρ) if using the rule ρ to make routing decisions does not prevent the system from reaching x from the zero state. We set

$$S(x) = \text{number of middle switches in use in state } x.$$

Let us use the diagram of Fig. 12 as a canonical representation for a 2 by 2 middle switch. The numbers at the left [top] indicate to which outer switch on the left [right] the numbered link connects. The seven possible states of a middle switch are depicted in Fig. 13, and are indexed therein by letters $a,b, \cdots, g$. A state x may then be represented (to within renaming switches and terminals) by giving seven integers $a(x), b(x), \cdots, g(x)$ where

$$a(x) = \text{number of middle switches of type } a \text{ when network is in state } x$$

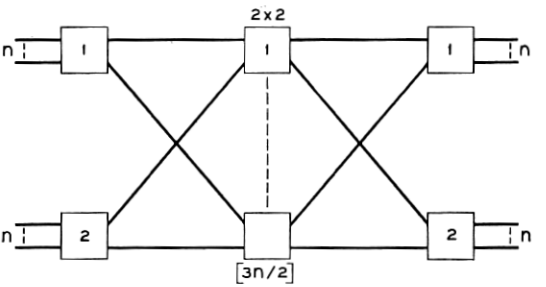

Fig. 11 — Three-stage network which is nonblocking if proper routing is used.

Fig. 12 — Representation of a 2 by 2 middle switch.

$\vdots$

$g(x)$ = number of middle switches of type g when network is in state x.

It is clear that for any state x

$$a(x) + b(x) + \cdots + g(x) = 2n - 1$$

$$b(x) + c(x) + \cdots + g(x) = S(x).$$

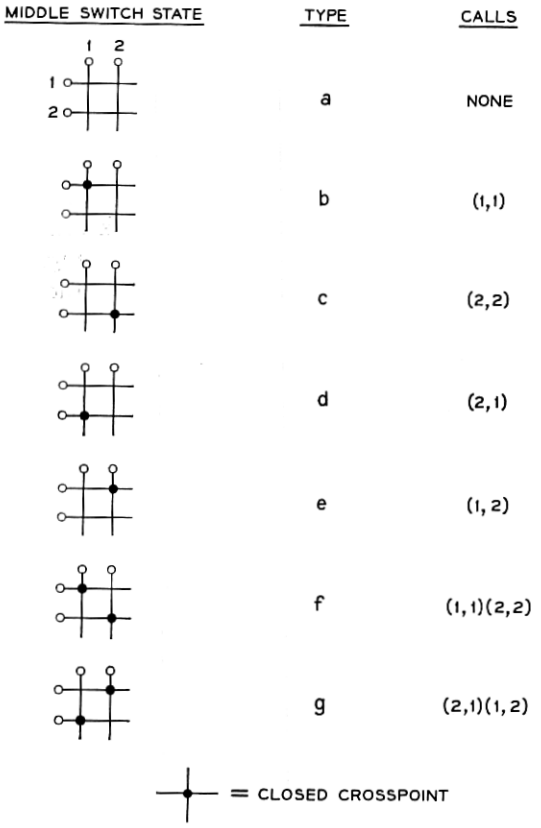

Fig. 13 — Seven possible states of a middle switch.

Theorem 1: *Let* ρ *denote the rule: Do not use an empty middle switch unless necessary. Let* x *be a state of the network of Fig. 10. Let* x *be reachable under* ρ. *Then for* $n \geqq 2$

$$S(x) \leqq [3n/2] \tag{1}$$

$$\left.\begin{aligned} b(x) + c(x) + f(x) &\leqq n \\ d(x) + e(x) + g(x) &\leqq n \end{aligned}\right\} \tag{2}$$

Proof: Each reachable state is reachable in a certain minimum number of steps. The theorem is true if x consists of one call and is reachable from the zero state in one step. As an hypothesis of induction, assume that the theorem is true for all states reachable in k steps or fewer. All changes in the state are either hangups, or new calls of the following kinds:

Type 1:

$$\begin{array}{lllll} & a(y) \to a(y) - 1 & & & \\ (1,1) & b(y) \to b(y) + 1 & \text{with} & c(y) = 0 \\ (2,2) & c(y) \to c(y) + 1 & \text{with} & b(y) = 0 \\ (2,1) & d(y) \to d(y) + 1 & \text{with} & e(y) = 0 \\ (1,2) & e(y) \to e(y) + 1 & \text{with} & d(y) = 0. \end{array}$$

Type 2: (preferred by ρ)

$a(y)$ remains fixed and

$$\begin{array}{llllll} (1,1) & f(y) \to f(y) + 1, & c(y) \to c(y) - 1 & \text{with} & c(y) > 0 \\ (2,2) & f(y) \to f(y) + 1, & b(y) \to b(y) - 1 & \text{with} & b(y) > 0 \\ (2,1) & g(y) \to g(y) + 1, & e(y) \to e(y) - 1 & \text{with} & e(y) > 0 \\ (1,2) & g(y) \to g(y) + 1, & d(y) \to d(y) - 1 & \text{with} & d(y) > 0. \end{array}$$

All states, reachable or not, satisfy the inequalities

$$\begin{aligned} b(y) + e(y) + f(y) + g(y) &\leqq n \\ c(y) + d(y) + f(y) + g(y) &\leqq n \\ b(y) + d(y) + f(y) + g(y) &\leqq n \\ c(y) + e(y) + f(y) + g(y) &\leqq n. \end{aligned}$$

The alternative preferred by ρ changes neither the value of $S(\cdot)$ nor the truth of (2) of the theorem. Consider a state x first reachable in

$k + 1$ steps. If x is first reachable by a hangup or by putting up a call of Type 2, then (1) and (2) are true of x. Suppose then that x is first reachable in $k + 1$ steps only by putting up a call of Type 1. Without loss of generality we can consider only the case where the new call is a (1, 1) call; the other three cases are symmetric. Let y be a state from which x is thus first reachable. Since the avoided alternative is used, we have

$$c(y) = 0.$$

Since a (1, 1) call is possible in state y, we must have

$$b(y) + d(y) + f(y) + g(y) \leqq n - 1$$
$$b(y) + e(y) + f(y) + g(y) \leqq n - 1$$

and from the induction hypothesis

$$d(y) + e(y) + g(y) \leqq n.$$

Hence,

$$2\{b(y) + d(y) + e(y) + f(y) + g(y)\} \leqq 3n - 2$$

or, since $c(y) = 0$

$$S(y) \leqq \frac{3n}{2} - 1.$$

However, $S(x) = S(y) + 1$, so $S(x) \leqq [3n/2]$. To show that (2) also holds of x consider that

$$b(y) + e(y) + f(y) + g(y) \leqq n - 1$$
$$c(y) = 0.$$

It follows that

$$b(y) + c(y) + f(y) \leqq n - 1.$$

However, since x is obtained from y by putting up a (1, 1) call of Type 1, we have

$$b(x) = b(y) + 1, \qquad e(x) = e(y)$$
$$c(x) = c(y) = 0, \qquad f(x) = f(y)$$
$$d(x) = d(y), \qquad g(x) = g(y).$$

Hence, (2) of Theorem 1 is true of x. This proves the result.

XV. A PROBLEM OF TRAFFIC CIRCULATION IN A TELEPHONE EXCHANGE

We shall describe and analyze a simple stochastic model for the operation of the control unit of a switching system. The connecting network is assumed to be nonblocking and is left out of account.

To set up a telephone call in a modern electromechanical automatic exchange usually involves a sequence of steps which are (traditionally and functionally) divided into two groups. The first group consists in collecting in a *register* the dialed digits of the called terminal. The second group, performed by a machine called a *marker*, consists in actually finding a path through the connecting network for the desired call, or otherwise disposing of the request for service. For even if a path to the called terminal be found, this terminal may already be busy.

In the exchange, enough registers and markers must be provided to give customers a prescribed grade of service. For engineering purposes, then, it is desirable to know the probability that r registers and m markers are busy. Let us assume that the exchange serves N customers, and that there are R registers and M markers. All calls are assumed to go to terminals *outside* the exchange.

We may think of each customer's line as being in one of a number of conditions, and moving from one condition to another. It makes no difference whether we ascribe these "conditions" to the line itself, or to a fictitious single customer if several people use the line. A given line may be *idle* (i.e., not in use); at some point in time it may request a connection, i.e., the customer picks up the receiver and starts *waiting for dial tone*; after obtaining a register he spends a certain amount of time *dialing*; he then *waits for a marker* to complete his call (freeing the register meanwhile); upon obtaining a marker, he must wait until the marker *completes* the connection; at this point he begins his *conversation*; at the end of his conversation his line becomes *idle* again.

One may now ask, what is the distribution of the N customers among these various conditions? Clearly, if not enough markers are provided there will be a tendency for the customers to collect in the "waiting for a marker" condition; a lack of registers will make the customers collect in the "waiting for dial tone" condition.

To obtain a simple probabilistic model for the "circulation" of customers, we assume that the probability that an idle customer starts a call in the next interval of time of length h is $\lambda h + o(h)$, the chance that a dialing customer completes his dialing in the next interval h is $\delta h + o(h)$, the chance that a busy marker finishes the call it is working on is $\mu h + o(h)$, and the probability that a conversation ends is $h + o(h)$,

all as $h \to 0$. The probability of more than one such event in h is $o(h)$ as $h \to 0$.

These assumptions are in turn consequences of assuming that the time a customer stays idle, the time a customer takes to dial, the time a marker takes to complete a call, and the holding time (conversation length) are all mutually independent random variables, each with a negative exponential distribution, and the respective means λ^{-1}, δ^{-1}, μ^{-1} and unity. The number λ is the calling rate per idle customer, δ and μ are the average rates of dialing and call completion by a marker (respectively), and time is measured in units of mean holding time, so that the hangup rate per call in progress is unity. The assumption that the marker operation times are exponentially distributed is not realistic, but we make it here in the interest of obtaining a global model whose statistical equilibrium equations can be solved in a simple way. This restrictive assumption could be avoided at the cost of complicating the mathematics. The important features of our model are depicted in Fig. 14; the labeled arrows indicate the rates of motion for various transitions.

The state of the system is adequately described by stating the number i of idle customers, the number r of customers that are dialing or waiting for dial tone, the number m that are being serviced by a marker or are waiting for a marker, and the number c of calls in progress. Actually, any three of these numbers suffice, since for physically meaningful states

$$i + r + m + c = N.$$

Let p_{irmc} be the equilibrium (or stationary) probability of the state

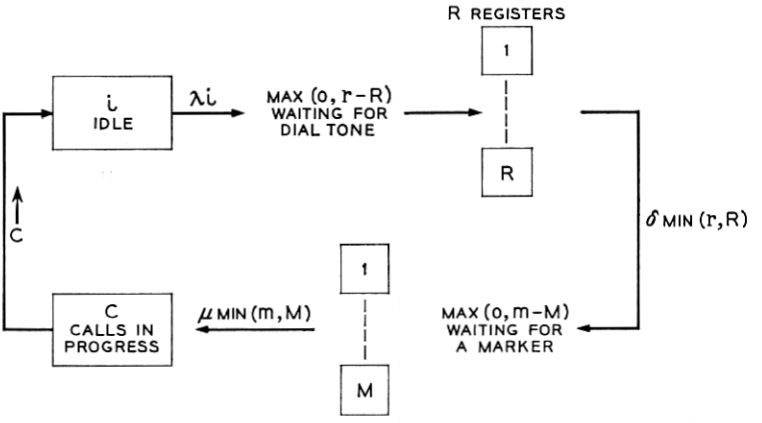

Fig. 14 — Diagram of a telephone system.

(i,r,m,c). The "statistical equilibrium" equations are, with suitable conventions at the boundaries.

$$(\lambda i + \delta \min(r,R) + \mu \min(m,M) + c)p_{irmc}$$
$$= (c + 1)p_{(i-1)rm(c+1)} + \lambda(i + 1)p_{(i+1)(r-1)mc}$$
$$+ \delta \min(r + 1,R)p_{i(r+1)(m-1)c} + \mu \min(m + 1,M)p_{ir(m+1)(c-1)}.$$

These equations state that the average rate at which a state is left equals the average rate at which it is reached from other states. We observe that the flow of calls in the exchange is in a sense *cyclic*; in making a call, each customer passes through four stages: idle, dialing, marker, conversation, then back to idle, in that order. This fact yields a way of solving the equations. Each side of the equilibrium equations has four terms, one for each of the four stages of a call. We shall find a way of assigning to each term on the left a corresponding *equal* term on the right which will cancel it.

The solution of the equations for $(i,r,m,c) \neq (N,0,0,0)$ is proportional to

$$f_{i,r,m,c} = \frac{N!}{i!r!m!c!} \cdot \frac{\prod_{j=0}^{r} \max(1,j/R) \prod_{j=0}^{m} \max(1,j/M)}{\lambda^i \delta^r \mu^m}.$$

The constant of proportionality is the probability of the "zero" state

$$p_{N000} = \left(1 + \sum_{\substack{i+r+m+c=N \\ i,r,m,c \geqq 0 \\ i<N}} f_{i,r,m,c}\right)^{-1}$$

obtained from the normalization condition for probabilities. The algebraic character of the solution is closely analogous to the actual pattern of circulating traffic in Fig. 14, for the easiest way of showing that f_{irmc} is actually a solution of the statistical equilibrium equations is to make the following correspondence between terms on opposite sides of the equations:

$$\lambda i p_{irmc} \sim (c + 1)p_{(i-1)rm(c+1)}$$
$$\delta \min(r,R)p_{irmc} \sim \lambda(i + 1)p_{(i+1)(r-1)mc}$$
$$\mu \min(m,M)p_{irmc} \sim \delta \min(r + 1,R)p_{i(r+1)(m-1)c}$$
$$cp_{irmc} \sim \mu \min(m + 1,M)p_{ir(m+1)(c-1)}.$$

It can be seen that each term on the left cancels the corresponding

one on the right when f_{irmc} is substituted. Each term represents the (total) rate of occurrence of one of the four kinds of possible event: request for service, completion of dialing, completion of a call, and hangup. In the life history of a given call, these events occur in the natural cyclic order given. Events associated with corresponding (i.e., canceling) terms are next to each other in this cyclic order.

XVI. AN OPTIMAL ROUTING PROBLEM

Our final example is a variational problem involving both combinatoric and probability. We shall exhibit some particular answers to the following question: If requested connections can be put up in a connecting network by several different routes, leading to different states, which routes should be chosen so as to minimize the probability of blocking? This question poses a variational problem in which many possible methods of operating a connecting network of given structure are compared, rather than one in which different network structures are compared.

We shall consider this question for a connecting network that is of little practical significance because it is obviously wasteful of crosspoints. Its virtues, however, are that it is perhaps the simplest network for which our question can be asked, and that it clearly exhibits the principles and arguments involved, so that these can be understood. The network is shown in Fig. 15, the squares standing for square 2 by 2 switches.

The possible states of this network are determined by all the ways

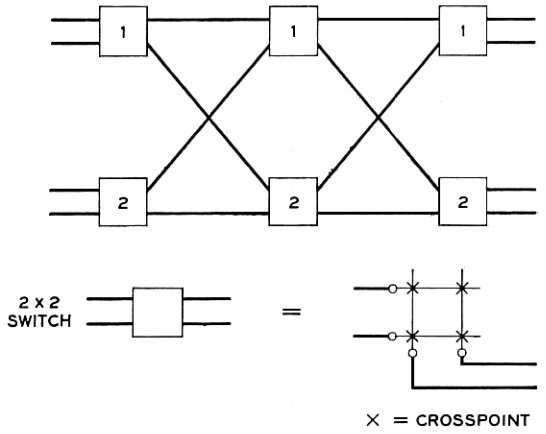

Fig. 15 — A simple network in which optimal routing is studied.

in which four or fewer inlets on the left can be connected pairwise to as many outlets on the right, no inlet being connected to more than one outlet, and vice-versa. These possible states are depicted in a natural arrangement in Fig. 16; states which differ only by permutations of customers or switches have been identified in order to simplify the diagram. That is, there is essentially only one way to put up a single call, there are four ways of having two calls up, two ways each of having three and four calls up. These "ways" have been arranged in rows according to the number of calls in progress, and lines have been drawn between states that differ from each other by only the removal or addition of exactly one call.

For ease of reference, let us number the states in the (partly arbitrary) way indicated in Fig. 16; insofar as possible, we have used small num-

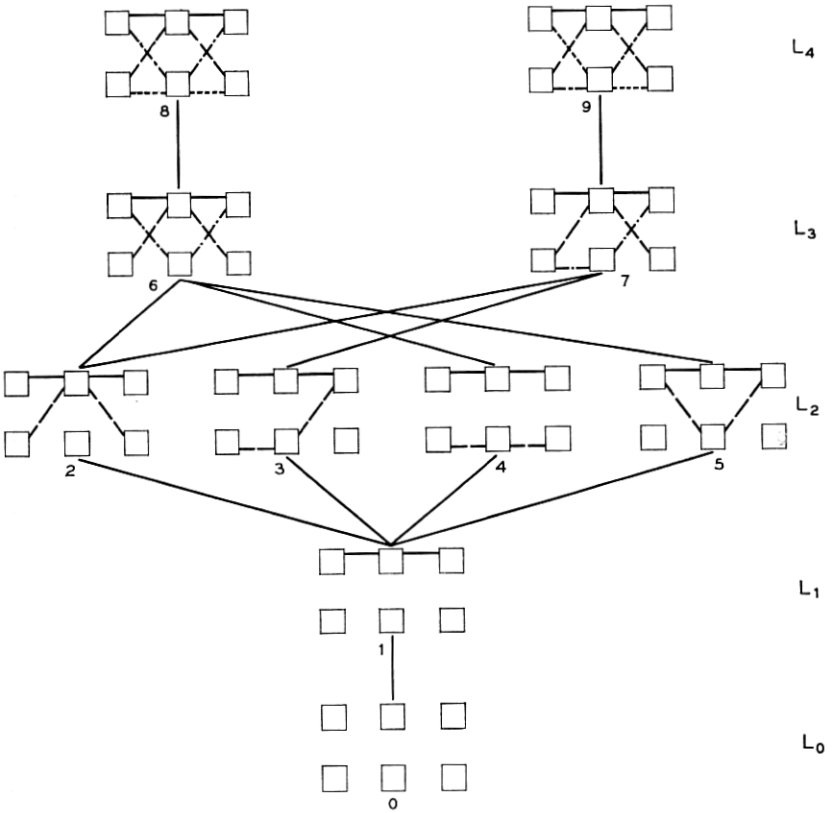

Fig. 16 — (Reduced) state diagram for the network shown in Fig. 15.

bers for states with small numbers of calls. The set of possible states of our example then consists of (essentially) ten different configurations of calls in the basic network of Fig. 15. The state diagram, with each state identified now only by its number within a small circle, is schematized in Fig. 17. Also indicated in this schema are two important sets of quantities associated with the states. To the left of each state is the number of idle inlet-outlet pairs, and to the right of each state is the number of idle inlet-outlet pairs that can actually be connected, i.e., that are not blocked.

Only in the state numbered 4 are there any blocked calls. It is to be noticed that state 4 realizes essentially the same assignment of inlets to outlets as state 2, which has no blocked calls. The difference between the two is that in state 2 all the traffic passes through one middle switch, leaving the other entirely free for any call that may arise. Clearly, then, this difference illustrates the "packing rule" that one should always put through a call using the most heavily loaded part of the network that will still accept the call.

The question naturally arises, therefore, whether this packing rule is

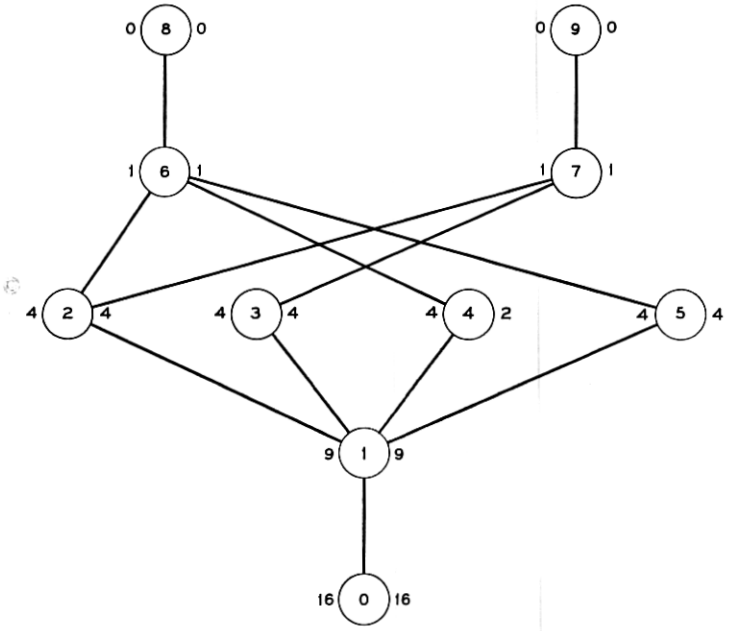

Fig. 17 — Schema of state diagram.

in any sense optimal for our particular example. We shall prove that it is, in two senses. It is clear from an inspection of the state diagram that only in state 1 is there ever a choice of route, and that this choice is always between states 2 and 4. From the fact that state 4 is the only state with any blocked calls, it is intuitively reasonable to expect that the probability of blocking is the least if the "bad" state 4 is avoided as much as possible, i.e., if from state 1 we always pass to either 2, 3, or 5, and visit 4 only when we have to, via a hangup from state 6.

The next task is to choose a probabilistic model for the operating network; this will be done in the simplest possible way. We postulate that in any state of the system, the probability that a given idle inlet-outlet pair request connection in the next interval of time h is $\lambda h + o(h)$, the chance that an existing connection cease is $h + o(h)$, and the chance that more than one event (new call or hangup) occur in h is $o(h)$, as $h \to 0$. The number λ is the calling rate per idle pair, and time is measured in units of mean holding time, so the "hangup" rate is unity. New calls that are not blocked are instantly connected, with some specific choice of route, while blocked calls are lost and do not affect the state of the system, their terminals remaining in the idle condition.

To complete the probabilistic description of the behavior of the system, it remains to specify how routes are chosen. In our example, this amounts to specifying whether, for certain calls arising in state 1, the route leading to state 2 or that leading to 4 is chosen. At first we shall only consider methods of choice that are independent of time, i.e., the choice is made in the same way each time.

The methods of choice over which we shall take an optimum may be parametrized as follows: each time a choice is to be made between going to state 4 and state 2, a coin is tossed with a probability α of coming up heads. If a head comes up we choose state 4; if a tail, we choose state 2; the toss of the coin is independent of previous tosses and of the history of the system. The parameter α may take on any value in the interval $0 \leqq \alpha \leqq 1$; the value $\alpha = 0$ corresponds to choosing state 2 every time; the value $\alpha = 1$ corresponds to choosing state 4 every time; a value of α intermediate between 0 and 1 means that 4 is chosen over 2 a fraction α of the time.

Introducing a natural terminology (from the theory of games), we may say that a choice of α represents a *policy* or *strategy* for making routing decisions; a value 0 or 1 of α represents a *pure strategy*, in which the route is specified by a rigid rule, and there is no randomization; an intermediate value of α represents a *mixed strategy*.

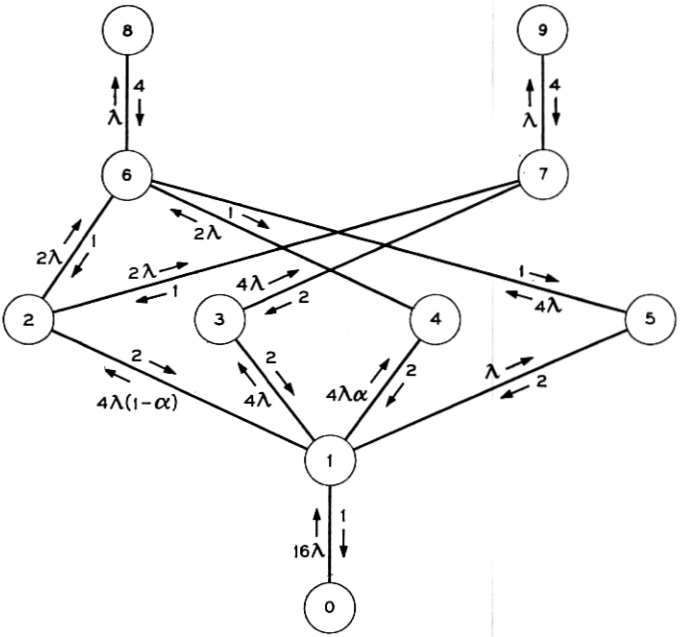

Fig. 18 — Schema of state diagram showing transition rates.

A choice of α determines a matrix $Q = Q(\alpha,\lambda)$ of transition rates (Fig. 18) among states of the system, and so a Markov stochastic process taking values on those states. As a measure of performance we shall use the fraction b *of requests for connection that encounter blocking*, defined as follows: let $b(t)$ be the number of blocked calls occurring in the interval $(0,t]$; and let $r(t)$ be the number of requests for service occurring in $(0,t]$; then

$$b = \lim_{t\to\infty} \frac{b(t)}{r(t)}.$$

It can be shown that this limit exists and is constant with probability one, so b is well defined.

The number $b = b(\alpha,\lambda)$ can be calculated from the matrix Q as follows: if $(i,i = 0, \cdots, 9)$ is a state, let $\beta(i)$ be the number of blocked idle pairs in state i, and let $\gamma(i)$ be the number of calls in existence in state i. The stationary state probabilities $\{p_i, i = 0, \cdots, 9\}$ exist and are the unique solution of the matrix-vector equation $Qp = 0$. Then b is given by

$$b = \frac{\sum_{i=0}^{9} p_i \beta(i)}{\sum_{i=0}^{9} p_i [4 - \gamma(i)]^2}$$

$$= \frac{2p_4}{\sum_{i=0}^{9} p_i \delta(i)}, \quad \text{with} \quad \delta(i) = [4 - \gamma(i)]^2$$

$$= \frac{(p,\beta)}{(p,\delta)}$$

where the inner product (p,x) is $\sum_{i=0}^{9} p_i x_i$.

We may therefore formally state our variational problem for this example as follows: to find that α in the interval $0 \leqq \alpha \leqq 1$ for which the ratio

$$b = \frac{(p,\beta)}{(p,\delta)} = \text{minimum}$$

subject to the conditions $Qp = 0$, $\sum_{i=0}^{9} p_i = 1$.

It is natural to expect that in choosing an optimum routing method in the example above there is no point in randomizing, i.e., using a mixed strategy with α unequal to either 0 or 1. That this is so is not obvious from our mathematical statement of the problem, and requires proof. We shall demonstrate a more general result:

Theorem 2: *Let x and y be vectors of* 10 *dimensions, with y nonnegative and not identically zero.*

$$\begin{matrix}\min \\ \text{or} \\ \max\end{matrix} \left\{ \frac{(p,x)}{(p,y)} \;\middle|\; Qp = 0, \quad \sum_{i=0}^{9} p_i = 1, \quad 0 \leqq \alpha \leqq 1 \right\}$$

is always achieved for $\alpha = 0$ or $\alpha = 1$.

Proof: The equation $Qp = 0$ may be written out in the detailed form

$$(i) \qquad 16\lambda p_0 = p_1$$

$$(ii) \qquad (9\lambda + 1)p_1 = 16\lambda p_0 + 2\sum_{i=2}^{5} p_i$$

$$(iii) \qquad (4\lambda + 2)p_2 = 4\lambda(1 - \alpha)p_1 + p_6 + p_7$$

$$(iv) \qquad (4\lambda + 2)p_3 = 4\lambda p_1 + p_7$$

$$(v) \qquad (2\lambda + 2)p_4 = 4\lambda\alpha p_1 + p_6$$

$$(vi) \qquad (4\lambda + 2)p_5 = \lambda p_1 + p_6$$

$$(vii) \qquad (\lambda + 3)p_6 = 2\lambda p_2 + 2\lambda p_4 + 4\lambda p_5 + 4p_9$$

$$(viii) \qquad (\lambda + 3)p_7 = 2\lambda p_2 + 4\lambda p_3 + 4p_8$$

$$(ix) \qquad 4p_8 = \lambda p_6$$

$$(x) \qquad 4p_9 = \lambda p_7 .$$

These are the standard "statistical equilibrium" equations for the probabilistic model we have assumed. They can be solved by successively eliminating every p_i except p_0 and obtaining a solution of the form

$$p_i = f_i p_0 , \qquad i \neq 0.$$

The value of p_0 is then determined by the normalization condition $\sum_{i=0}^{9} p_i = 1$ as

$$p_0 = \frac{1}{1 + \sum_{i=1}^{9} f_i} .$$

The f_i are of course functions of λ and α. We shall prove that they are *linear* functions of the parameter α.

We first eliminate p_1 and note that $f_1 = 16\lambda$. Since the relations (iii)–(iv) contain the variables $\{p_i , i = 2,3,4,5\}$ only on the left, these variables may be eliminated entirely from (ii), and from (vii)–(x). But substitution for these variables in (vii) and $(viii)$ in terms of (iii)–(vi) introduces α and p_0 only in inhomogeneous terms. Hence, f_6 and f_7 are linear in α, and so all $\{f_i , i = 1, \cdots , 9\}$ are linear in α.

Clearly, we have

$$\frac{(p,x)}{(p,y)} = \frac{(f,x)}{(f,y)}$$

because the normalization terms $1 + \sum_{i=1}^{9} f_i$ cancel out, and so it follows that $(p,x)/(p,y)$ is a *bilinear* function of α, i.e., it has the form

$$g(\alpha) = \frac{A_1 + B_1\alpha}{A_2 + B_2\alpha}$$

where A_1 , A_2 , B_1 , and B_2 are constants. Now

$$\frac{d}{d\alpha} g(\alpha) = \frac{B_1(A_2 + B_2\alpha) - B_2(A_1 + B_1\alpha)}{(A_2 + B_2\alpha)^2}$$

$$= \frac{B_1A_2 - B_2A_1}{(A_2 + B_2\alpha)^2}$$

which is of the same sign as its numerator. Thus $g'(\alpha)$ is either always nonpositive or nonnegative, and so any extremum of $g(\alpha)$ in $0 \leqq \alpha \leqq 1$ is assumed at the boundary, either for $\alpha = 0$ or $\alpha = 1$. Since the solution p of $Qp = 0$ is known to have all strictly positive components for all α in the unit interval, we have $A_2 + B_2\alpha = (p,y) > 0$.

It follows in particular that the minimum of blocking probability b is achieved for $\alpha = 0$ or $\alpha = 1$. It is unthinkable that visiting a blocking state (state 4) more frequently should decrease b, so we conjecture (and shall shortly prove that) α should be zero rather than one.

Before doing this though, let us observe that there is only one blocking state (viz., 4), and that the blocking probability b can be written as

$$b = \frac{2p_4}{16p_0 + 9p_1 + 4\sum_{i=2}^{5} p_i + p_6 + p_7}.$$

These facts and our intuition suggest that b should be a monotone increasing function of

$$f_4 = \frac{p_4}{p_0}.$$

This conjecture is correct, and provides an easy way of showing that $\alpha = 0$ gives the least blocking probability. Let us prove it.

From (*i*) and (*ii*) we find that

$$\sum_{i=2}^{5} p_i = 8\lambda(9\lambda + 1)p_0 - 2\lambda p_0 = 72\lambda^2 p_0$$

whence

$$b = \frac{2f_4}{16 + 144\lambda + 288\lambda^2 + f_6 + f_7}.$$

From (*vii*)–(*x*) we find that

$$p_6 + p_7 = \frac{1}{\lambda + 3}\left(\lambda(p_6 + p_7) + 4\lambda\sum_{i=2}^{5} p_i - 2\lambda p_4\right)$$

$$= \tfrac{4}{3}\lambda\sum_{i=2}^{5} p_i - \tfrac{2}{3}\lambda p_4$$

$$= 96\lambda^3 p_o - \tfrac{2}{3}\lambda p_4.$$

Therefore

$$b = \frac{2f_4}{16 + 144\lambda + 556\lambda^2 + 192\lambda^3 - \frac{2}{3}\lambda f_4}.$$

This is of the form

$$\frac{2x}{a - cx}$$

where a and c are strictly positive constants. Now

$$\frac{d}{dx}\frac{2x}{a - cx} = \frac{2}{a - cx} + \frac{2cx}{(a - cx)^2}$$

$$= \frac{2a}{(a - cx)^2} \geqq 0.$$

Hence, b is a monotone increasing function of f_4. It follows that b is a minimum if f_4 is a minimum.

To prove that the blocking probability b is a minimum for $\alpha = 0$, it remains to calculate p_4 from the equilibrium equations. By eliminating all the equilibrium probabilities except p_6 and p_7, we find

$$p_6 = \frac{1}{\lambda + 3}\left(\frac{8\lambda^2(1 - \alpha)16\lambda p_0}{4\lambda + 2} + \frac{8\lambda^2\alpha 16\lambda p_0 + 2\lambda p_6}{2\lambda + 2} + \frac{4\lambda^2 16\lambda p_0 + 4\lambda p_6}{4\lambda + 2} + \lambda p_7\right)$$

$$p_7 = \frac{1}{\lambda + 3}\left(\frac{8\lambda^2(1 - \alpha)16\lambda p_0}{4\lambda + 2} + \frac{(16\lambda)^2\lambda p_0 + 4\lambda p_7}{4\lambda + 2} + \lambda p_6 + \frac{2\lambda}{4\lambda + 2}(p_6 + P_7)\right).$$

We have purposely not simplified the terms so that their origin can be verified. From these two equations we find that

$$f_6 = \frac{p_6}{p_0}$$

$$= X^{-1}128\lambda^3\left(\frac{1 - \alpha}{4\lambda + 2} + \frac{\alpha}{2\lambda + 2} + \frac{1}{8\lambda + 4} + \frac{\lambda\left(\frac{1 - \alpha}{4\lambda + 2} + \frac{1}{2\lambda + 1}\right)}{\lambda + 3 - \frac{3\lambda}{2\lambda + 1}}\right)$$

where

$$X = \lambda + 3 - \frac{\lambda}{\lambda + 1} - \frac{2\lambda}{2\lambda + 1} - \frac{2\lambda^3 + 2\lambda^2}{2\lambda^3 + 4\lambda + 3}$$

$$= \frac{2\lambda^3 + 5\lambda^2 + 7\lambda + 3}{2\lambda^2 + 3\lambda + 1} - \frac{2\lambda^3 + 2\lambda^2}{2\lambda^3 + 4\lambda + 3}$$

$$> 0.$$

The coefficient of α in f_6 is

$$\frac{128\lambda^3}{2\lambda + 2}\left(1 + \frac{2\lambda + 2}{4\lambda + 2}\left(1 + \frac{\lambda}{\lambda + 3 - \dfrac{3\lambda}{2\lambda + 1}}\right)\right).$$

This is positive, because

$$1 - \frac{2\lambda + 2}{4\lambda + 2}\left(1 + \frac{\lambda}{\lambda + 3 - \dfrac{3\lambda}{2\lambda + 1}}\right) = 1 - \frac{\lambda + 1}{2\lambda + 1}\left(\frac{4\lambda^2 + 5\lambda + 3}{2\lambda^2 + 4\lambda + 3}\right)$$

$$= 1 - \frac{4\lambda^3 + 9\lambda^2 + 8\lambda + 3}{4\lambda^3 + 10\lambda^2 + 10\lambda + 3}.$$

However,
$$\frac{p_4}{p_0} = \frac{32\lambda^2\alpha}{\lambda + 1} + \frac{p_6/p_0}{2\lambda + 2}.$$

Hence,
$$\frac{df_4}{d\alpha} > 0.$$

We shall now consider the problem of optimal routing in our (trivial) network from a different point of view. Instead of minimizing the *ratio* of unsuccessful attempts to attempts, let us simply minimize the average number of unsuccessful attempts in any finite number of events, counting changes of state and unsuccessful attempts as events.

In our example, the only choice is between states 2 and 4, when a particular call requests connection in state 1. By a *policy*, let us mean a function $p(\cdot)$ on the nonnegative integers taking the values 0 and 1. Let x_n be the state of the network after n events, $n \geqq 0$. We say that the system is operated according to policy $p(\cdot)$ if, for each $n \geqq 0$, given that $x_n = 1$ and a choice occurs, the system moves to

state 2 if and only if $p(n) = 1$

state 4 if and only if $p(n) = 0$.

Now our intuitive feeling is that going to state 2 is preferable over

going to state 4 under all circumstances. At the cost of anticipating results to be proven, let us partially order all the possible policies by the definition: If $p(\cdot)$ and $q(\cdot)$ are policies, then

$$p \geqq q \quad \text{if and only if} \quad p(n) \geqq q(n) \quad \text{for all} \quad n \geqq 0.*$$

The shift transformation T of policies $p(\cdot)$ is defined by the condition

$$Tp(n) = p(n + 1) > n \geqq 0.$$

It is evident that $p \geqq q$ implies $Tp \geqq Tq$. Let $E_{0,p}(x) \equiv 0$, and define

$$E_{n,p}(x) = E\left\{\begin{array}{l}\text{number of unsuccessful attempts after } n \text{ events} \\ \text{starting from state } x \text{ if the system is operated ac-} \\ \text{cording to policy } p(\cdot)\end{array}\right\}$$

Let S be the set of states $\{0, 1, \cdots, 9\}$.
We shall prove

Theorem 3: *If* $p \geqq q$, *then for all* $n \geqq 1$ *and* $x \in S$

$$E_{n,p}(x) \leqq E_{n,q}(x).$$

As a preliminary result (not without its own interest) we shall need the

Lemma: *For* $n \geqq 1$ *and any policy* $p(\cdot)$

$$E_{n,p}(4) = \max_{x \in S} E_{n,p}(x).$$

This says that starting in the (sole blocking) state 4 is always the worst way to start, no matter how long we run the system.

Proof: For $n = 1$ and $x \neq 4$, $E_{n,p}(x) = 0$ since no unsuccessful attempts can occur in any state except 4. However,

$$E_{1,p}(4) = \frac{2\lambda}{2 + 2\lambda}$$

so the lemma is true for $n \leqq 1$. Assume as an hypothesis of induction that it is true for $n \leqq k$. Now for $x \neq 4$, $E_{k+1,p}(x)$ is a convex combination of values of $E_{k,Tp}(\cdot)$, so clearly for $x \neq 4$

$$E_{k+1,p}(x) \leqq \max E_{k,p}(y) = E_{k,p}(4).$$

However, elementary probability arguments establish that

$$E_{k+1,p}(4) = E_{k,p}(4) + Pr\{x_k = 4 \mid x_0 = 4\}E_{1,T^k p}(4)$$

so the lemma is proven.

* Read "$p \geqq q$" as "p is better than q"!

Proof of Theorem 3: For any policy $s(\cdot)$

$$E_{1,s}(x) = 0 \quad \text{if} \quad x \neq 4$$

$$E_{1,s}(4) = \frac{2\lambda}{2 + 4\lambda}.$$

Hence,

$$E_{1,p}(x) = E_{1,q}(x) \quad \text{for all} \quad x \in S.$$

Assume as an hypothesis of induction that $p \geqq q$ implies

$$E_{n,p}(x) \leqq E_{n,q}(x)$$

for all x and all $n \leqq k$. Now for $x \neq 4$ or 1 and any policy $s(\cdot)$, $E_{k+1,s}(x)$ is a convex combination of values of

$$E_{k,Ts}(\cdot).$$

For $x = 4$, we have for any policy $s(\cdot)$

$$E_{k+1,s}(4) = \frac{2\lambda}{2 + 4\lambda} + \text{convex combination of } E_{k,Ts}(\cdot)$$

where the coefficients of the convex combination are transition probabilities independent of the policy $s(\cdot)$, and

$$\frac{2\lambda}{2 + 4\lambda} = \Pr\left\{\begin{matrix}\text{first event is a} \\ \text{blocked attempt}\end{matrix} \text{ start in state 4}\right\}.$$

Hence, $p \geqq q$ and $x \neq 1$ implies

$$E_{k+1,p}(x) \leqq E_{k+1,q}(x)$$

For $x = 1$ and any policy $s(\cdot)$ we have

$$E_{k+1,s}(1) = \frac{4\lambda}{1 + 9\lambda}\{s(1)E_{k,Ts}(2) + [1 - s(1)]E_{k,Ts}(4)\} + \frac{1 + 5\lambda}{1 + 9\lambda} \text{ convex combination of } E_{k,Ts}(\cdot)$$

where the coefficients of the convex combination are independent of $s(\cdot)$, and

$$\frac{4\lambda}{1 + 9\lambda} = \Pr\left\{\begin{matrix}\text{first event requires} \\ \text{routing decision}\end{matrix} \text{ start in state 1}\right\}.$$

Suppose now that $p \geqq q$. It is sufficient to show that

$$p(1)E_{k,\tau p}(2) + [1 - p(1)]E_{k,\tau p}(4)$$
$$\leqq q(1)E_{k,\tau q}(2) + [1 - q(1)]E_{k,\tau q}(4).$$

If $p(1) = q(1)$, this follows from the hypothesis of induction. The only other possibility is that $p(1) = 1$ and $q(1) = 0$. By the lemma and the hypothesis of induction we find

$$E_{k,\tau p}(2) \leqq E_{k,\tau p}(4)$$
$$\leqq E_{k,\tau q}(4).$$

This proves Theorem 3. The result at once shows that the policy $p \equiv 1$ is optimal in the sense that it minimizes

$$\limsup n^{-1}E_{n,p}(x).$$

XVII. ACKNOWLEDGMENT

The author takes pleasure in expressing his gratitude to J. Riordan, R. I. Wilkinson, E. Wolman, A. Descloux, and W. Helly for reading the preliminary draft. Their encouragement and suggestions have led to substantial improvements.

REFERENCES

1. Beneš, V. E., Algebraic and Topological Properties of Connecting Networks, this issue, pp. 1249–1274.
2. Beneš, V. E., On Rearrangeable Three-Stage Connecting Networks, to appear.
3. Beneš, V. E., Markov Processes Representing Traffic in Connecting Networks, to appear.
4. Beneš, V. E., A Thermodynamic Theory of Traffic in Connecting Networks, to appear.
5. Syski, R., *Introduction to Congestion Theory in Telephone Systems*, Oliver and Boyd, London, 1960.
6. Kosten, L., The Historical Development of the Theory of Probability in Telephone Traffic Engineering in Europe, Teleteknik, **1,** 1957, pp. 32–40.
7. Wilkinson, R. I., The Beginnings of Switching Theory in the United States, Teleteknik (English Edition), **1,** 1957, pp. 14–31.
8. Molina, E. C., Computation Formula for the Probability of an Event Happening at Least c Times in N Trials, Amer. Math. Monthly, **20,** 1913, pp. 190–193.
9. Jensen, A., An Elucidation of A. K. Erlang's Statistical Works Through the Theory of Stochastic Processes, in the Life and Works of A. K. Erlang, Trans. Danish Acad. Sciences, 1948, pp. 23–100.
10. Engset, T., Die Wahrscheinlichkeitsrechnung zur Bestimmung der Wähleranzahl in Automatischen Fernsprechämtern, E.T.Z., **31,** 1918, pp. 304–306.
11. O'Dell, G. F., An Outline of the Trunking Aspect of Automatic Telephony, J. Inst. Elec. Engrs., **65,** 1927, pp. 185–222.
12. Crommelin, C. D., Delay Probability Formulae, P. O. Elec. Engrs. J., **26,** 1933–1934, pp. 266–274.
13. Molina, E. C., Application of the Theory of Probability to Telephone Trunking Problems, B.S.T.J., **6,** 1927, pp. 461–494.
14. Pollaczek, F., Über eine Aufgabe der Wahrscheinlichkeitstheorie, Math. Zeit., **32,** 1930, pp. 64–100 and 729–750.

15. Khinchin, A. I., Matematicheskaya Teoriya Statsionarnoi Ocheredi, Matematicheskii Sbornik, **39,** 1932, pp. 73–84.
16. Fry, T. C., *Probability and Its Engineering Uses*, D. Van Nostrand, New York, 1928.
17. Kolmogorov, A. N., *The Foundations of Probability*, Second Edition, Chelsea, New York, 1956.
18. Palm, C., Intensitätsschwankungen im Fernsprechverkehr, Ericsson Technics, **44,** 1943, pp. 1–189.
19. Feller, W., On the Theory of Stochastic Processes, with Particular Reference to Applications, *Proc. [1st] Berkeley Symp. Math. Stat. and Prob.*, 1949, pp. 403–432.
20. Kosten, L., On the Influence of Repeated Calls in the Theory of Probabilities of Blocking, De Ingenieur, **59,** 1947, pp. 1–25.
21. Kosten, L., Manning, J. R., and Garwood, F., On the Accuracy of Measurements of Probabilities of Loss in Telephone Systems, J. Royal Statistical Soc., B, **11,** 1949, pp. 54–67.
22. Scudder, F. J., and Reynolds, J. N., Crossbar Dial Telephone Switching System, B.S.T.J., **18,** 1939, pp. 76–118.
23. Jacobaeus, C., A Study on Congestion in Link Systems, Ericsson Technics, **51,** 1950, pp. 1–68.
24. Fortet, R., and Canceill, B., Probabilité de Perte en Selection Conjuguée, Teleteknik, **1,** 1957, pp. 41–55.
25. Lee, C. Y., Analysis of Switching Networks, B.S.T.J., **34,** 1955, pp. 1287–1315.
26. Le Gall, P., Methode de Calcul de L'encombrement dans les Systèmes Téléphoniques Automatiques à Marquage, Ann. des Telecom., **12,** 1957, pp. 374–386.
27. Lundkvist, K., Method of Computing the Grade of Service in a Selection Stage Composed of Primary and Secondary Switches, Ericsson Review, No. 1, 1948, pp. 11–17.
28. Elldin, A., Applications of Equations of State in the Theory of Telephone Traffic, Thesis, Stockholm, 1957.
29. Clos, C., A Study of Non-Blocking Switching Networks, B.S.T.J., **32,** 1953, pp. 406–424.