Abstract

Traffic and trunking problems in telephone service were approached for nearly three-quarters of a century by probability methods. The approaches were made under idealized assumptions, such as: statistical equilibrium, Poisson input, independent events, exponential holding time distribution, and several others that were considered pertinent and expedient. In spite of the doubtful validity of these assumptions, the productivity of these approaches were truly remarkable. Further progress was impeded by practical limitations that restricted acceptable analytical methods almost exclusively to full availability arrangements of connecting devices. Other circumstances during recent years caused difficulties to be increasingly disturbing: unreliable predictions of traffic demands, accelerated and more massive traffic growth, progressive loss of random traffic character, increasing share of data transmissions, etc.

The changed conditions required appropriately changed approaches to traffic problems. Certain concurrent developments were helpful, such as: enlarged scope of theoretical studies, digital computers, traffic simulations, automated traffic measuring devices, mechanized network management, and adaptive routing. The most interesting of the latter starts out with an optimal steady-state routing pattern, but which is continually updated by monitoring the actual path taken from its origin by all traffic incoming to switching centers, and evaluating this information in computers for control purposes. The many aspects of such a feed-back system are very involved: such as optimization of route, link, and channel cost, number of switching points, stability, dead ending, vicious circles: they have been studied analytically and through simulations, but with inconclusive outcomes. Nevertheless, further patient research could be very worthwhile.

Traffic engineering in a service industry is the science of establishing the relationship between service demand, service implements and service results, as well as of measuring this relationship or any of its factors. Usually, the ultimate aim for establishing such a relationship is to optimize somehow the needed investment in implements for the potential demand relative to the received overall quality.

Pragmatism in Science

During the recent ten-twenty years, the approaches to telecommunication traffic engineering problems have basically changed. Similar changes in outlook can also be detected in many areas of present-day research, especially in physical and biological sciences. This approach is trying to avoid abstractions and is altogether very cautious about generalizations: they are permissible only as far as they are relevant to and interpret satisfactorily realities as experienced. Such principle does not intend to repudiate a theoretical versus an empirical outlook, but simply establishes a measure of relevancy for theoretical studies. There is nothing unusual or novel about such an outlook: in fact, it successfully dominated physical sciences all through their history. Periods with preeminence of abstractions and speculations were invariably supplanted by epochs of pragmatism achieving success through applications of empiricism with their appropriate interpretations. In other words, realities were found to be more serviceable than abstractions, physics more than metaphysics. Though theoretic physics of the nineteenth century accomplished the most brilliant feats in the history of natural science, through speculations and abstractions: still, the outlook on research had to be completely replaced during the past half century, due partly to having entirely new species of research tools available, and partly because the changed outlook forced the creation of such tools.

Inadequacy of Past Approaches

The preceding comments were brought up to show that the recent trends in telecommunication traffic engineering are just specific manifestations of a general trend in theoretical sciences. For nearly three-quarters of a century now, traffic and trunking problems in telephone service were approached by probability methods. The approaches were made on simple and somewhat idealized assumptions, such as: statistical equilibrium, Poisson input, independent events, exponential holding time distribution, and several others that were considered pertinent and expedient. Besides the stimulus bestowed upon probability research, the productivity of these approaches were truly remarkable, in spite of the doubtful validity of the assumptions.

In reality, telecommunications demand is not generated at random and its sources are not acting independently: there appears to be some inherent though elusive correlation. The inherent correlation between the coincident states of occupancies at various cross sections of the telecommunications plant is not at all elusive but quite obvious. Conversely, although the existence of some motivation between successive attempts from a traffic source, or between the receipt and origination of messages appear to be credible, to find a mathematical correlation by con-
lecture has turned out to be completely elusive. In the same way, telecommunications traffic is rarely, if ever, in statistical equilibrium: that invariably one has to consider the "busy hour" or "busy season" in traffic engineering evidence that there is no such thing as statistical equilibrium; and that, since the inception of telephony, there is, with rare exceptions, an incessant growth in telephone traffic suggest that traffic conditions will remain stationary from one day to the next, so to speak.

Lastly, it is perfectly obvious that the population of traffic sources is not homogenous, and for many applications the assumption of an exponential holding time distribution is deceptive.

Further progress was impeded by such practical limitations that restricted acceptable analytical methods almost exclusively to full availability arrangements of connecting devices. Unfortunately, only a small portion of the engineering problems relate to full availability systems, and these happen to be the less interesting ones in most cases. The many different approaches that were proposed over the years to the more important and intricate arrangements, such as to grading schemes, or link systems, were extremely complex, virtually impossible to evaluate except in some trivial cases, and, as a result, were very controversial and were difficult to substantiate in practice. The application of one or the other became something like a profession of faith, the choice often made on subjective appeal, or the result of manipulation, or the reputation of the proponent.

The limitations were even more severe when switching systems were introduced where congestion conditions are producing delays. In addition to the complications mentioned in the preceding paragraph in relation to the trunking structure, there are also many incidental facets that resolve the circumstances on the orderly disposal of attempts as well as the traffic handling capability of the connecting means. Some of the most remarkable achievements in mathematics were accomplished in the attempt to find solutions in closed form to these problems; nevertheless they were only successful under extremely restrictive conditions which were but rarely satisfied in actual service. As a matter of fact, more often than not one may not even find a mathematical model that would correctly represent the service conditions. Even worse, in order to arrive at some estimate on the outcome, one must not only select judiciously a model that appears to be a good match to the actual conditions, but must also apply a formula which can yield quantitative answers that may, or most likely will have to improvise by interpolating between several of these that are conveniently available.

Trends in Telecommunications Traffic

If the preceding perspective appears to be defeatist, the experiences were in reality just the contrary. Abstract traffic research has yielded in surprisingly large number of applications surprisingly accurate estimates, while in those cases where it did not, it provided close enough guidelines without which telecommunications service could not have been as successful as it has proven to be.

That notwithstanding the analytical impediments which were just discussed, telecommunications engineering had such a successful history was to a large part due to the specified high service quality; that in effect the number of attempts that could not be completed due to insufficiency of number of connections, channels was negligible compared with the attempts which were unsuccessful for other reasons, and likewise that the number of attempts which were appreciably delayed due to such causes were few when compared with the delays brought about by external circumstances. Under such conditions, the propriety of the assumptions did not appear to be so pleasant. Conversely, in applications where for financial or for other reasons it was not practical to provide sufficient implements and long service delays became the rule, the appropriateness of the assumptions became irrelevant.

Nevertheless, circumstances that have developed during the past twenty years have brought forth practical difficulties and made the inadequacies of abstract research to be increasingly apparent.

Predicting traffic demands became less reliable due to the increased mobility of telephone users and the expanding range of their community of interest.

Increasing speed and volume in traffic growth. In former years, the rate of service facility additions were made to keep pace with the continuing modest rate of increase in the number of subscribers at a reasonably predictable and orderly manner. However, this recent increasing speed and volume in traffic growth, together with the increasing cost of labor and lesser availability of qualified manpower, have brought now the cost of delays as well as of re-arrangements higher than the cost of a temporary inventory of excess facilities.

The volume and scope of telecommunications service demands and, what is even worse, also of the facilities to cope with them have increased spectacularly. Coincidently, and with no doubt motivated by it, due to steady improvements in manufacturing and distributing methods more often than not it became uneconomical to provide facilities in single units, or even in small packages. Thus, it would be meaningless for such applications to be concerned with the examination of the number of circuits, or indeed with some odd number of circuits, if this would thwart the economical packaging of facilities. On the other hand, for occasional applications where such modular provisions would be clearly uneconomical, past traffic engineering approaches are still more than adequate, thus there seems to be little justification to add a great deal of new effort in refining these.

It seems as if traffic demands would be increasingly correlated, be less random, and the periods of statistical equilibrium be more confined. However, large scale growth in data communications and other applications of digital transmission may help toward equilibrium in the future.

With the increasing consistent use of mechanized alternate routing as a normal procedure, random distributions within various sectors became more and more disturbed. Anything that even resembles random input (if it has ever existed) is increasingly difficult to find, and it also seems that the structures of newer facilities have less of a tendency to randomize traffic — if anything they do the opposite.

The increasing volume of transmission of data further disturbs random traffic distribution, as well as the characteristics of the times. Therefore, the resulting changes in the traffic handling capabilities of trunking facilities are analogous to the changed probabilities of overloading frequency-division carrier repeaters. Further changes may be expected from increasing application of digital transmission techniques, although the nature of such changes is hard to predict at the present; and changes could be even more profound should techniques similar to Time Assignment Speech Interpolation (TASI) be more widely applied.

An increasing portion of trunking problems require delay analysis. While this is conceptionally more satisfactory than predictions based on theories, their mathematical analyses, as previously pointed out, is vastly more complex and require more precision in stating parametric conditions, rules, and assumptions. As it was also pointed out before, this is not always feasible: usually one has to take recourse to dubious approximations and interpolations demanding a great deal of subjective judgment and intuition.

The preceding difficulties are aggravated by the growing complexity of contemporary systems of complete assemblages of interconnected sub-systems.
Recent Advances

The evolution of telecommunications with the multiform drastic changes of traffic characteristics require appropriately changed approaches to traffic engineering problems. Fortunately, certain concurrent developments were helpful to advance a changed course and to place the emphasis in present-day traffic engineering on matters and methods other than that has been prevalent in times past. A few of these developments are listed below:

* A large amount of extremely ingenious research is being devoted successfully on the theories of delays. They became refined and steadily expanded to include increasingly sophisticated situations, as posed by present-day network structures and newer service demands, such as for data communications, including updated telegraphy, and remote operation of computers.

* Digital computers that are capable to evaluate formerly irreducible analytic expressions. Thus, the final limitation to the application of a theory is not any more the degree of difficulties in reducing it in a closed form for numerical computations, but instead in the capacities and capabilities of computers, and whether or not the problem justifies the cost of their employment.

* Versatile traffic simulation methods with digital computers brought about a better understanding of complex network problems, as well as an ability to confirm or refute proposed mathematical patterns. Simulations are productive as long as they are not made merely to verify that there were no errors made in developing the formal analytic procedure; but, instead, if they produce explicit numerical outputs generated by specified inputs through the simulated specific network structure. The limitations are again in the capacities and capabilities of computers, as well as in the substantial expense in exploring the whole range of parameters. Moreover, one should not lose sight of the fact that source traffic at the present is being simulated by pseudo-random inputs, while it has been pointed out repeatedly in the foregoing that telecommunications traffic is but rarely at random.

* The development of efficient and progressively automated observing, recording, and evaluating devices of telecommunication traffic. Next to what the researcher possesses knowledge in this twain, this is probably the most productive tool to perform traffic research. The telecommunication network, and particularly the switching equipment in its own analogue computer, capable of producing an exact record of its response to realistic instead of artificial inputs. With complete automation and advance planning the cost is most reasonable relative to the economics obtainable from the knowledge, and by appropriate programming the amount of information being produced can be limited to what is significant and usable.

* Further advances for making studies more realistic are feasible by combining some aspects of the preceding procedures. For example, instead of feeding pseudo-random inputs to a simulation process, one could select from an appropriate segment of the network suitable records of real traffic for the problem on hand, compressed in time and redistributed as required by the simulation program. The author is not aware of if this has been done or even proposed before for telecommunication research or operation, though its application appears to be feasible and reasonably simple. It would be a significant advance in producing realistic predictions and understanding on the behavior of projected systems and arrangements.

New and efficient methods for mechanization of network management. The remaining discussions will be devoted to this facet.

Network Management

The concepts of network management and adaptive routings are to some degree interrelated, though by no means synonymous. Anyone who is planning or operating a telecommunication network, for instance a traffic engineer, is of course a network manager. In a more restrictive sense, network management is a routine, almost day-by-day or minute-by-minute adjustment of the rules in routing traffic within a network, according to or anticipating its state of occupancy. In colloquial usage it implies human surveillance, decision and action. However, the recent concept of automated network management is among the most challenging of the newer approaches.

Since early days, temporary network control through human administration has been practiced particularly during periods of heavy traffic obstruction or in emergency situations. Traffic controlled by automatic routing arrangements, with a rigid hierarchy or with some degree of flexibility have also been widely used now for quite some time, and their fields of application and scope are broadening all the time. With displays of network occupancies and facilities for the remote control of routings, full-time continuous centralized administrative guidance became an effective method for the management of complex communication networks, supplementing or temporarily supplanting the normal rules of routings.

The concept of adaptive routing implies that the network modifies or establishes its own patterns of traffic routing, by observing and analyzing the itinerary and experience of current network traffic as well as of the immediate past. The ground rules, of course, are established at the outset within the control mechanics. In a way, one may argue that adaptive routing has always been practiced. A switch finding a trunk occupied passes over it. Or, when a trunkgroup is fully occupied in hierarchical routing, a new attempt will look for the next available one higher up on the ladder; with a possible change in pattern according to the time of the day or by some other external conditions. Or, after being set up over a series of several segments, a connection finding a busy condition somewhere along the line may revert to one of the earlier stages, and may try a prescribed alternate route hoping to be successful there, though with no assurance whatsoever.

None of these procedures, however, are really self-adaptive but simply follow rules prescribed for the situation an attempt is encountering at the moment. In true adaptive routing the network has accumulated a certain amount of experience for some time and governs its actions by taking into account the chances of success, together with other factors such as cost, quality, transmission delay, and the like.

Preliminaries

Systems using adaptive routing in the sense just stated have been proposed, abstractly analyzed, simulated on digital computers in several variances, and performance characteristics predicted; but as far as it is known, no such system has been put in actual service. These self-adaptive systems were thought to be most attractive for military communications, mainly because they were considered to be relatively insensitive to destruction, which may come through hostile
action, or equipment failure, or noise, fading, or other propagation difficulties: by reacting spontaneously to such circumstances and substituting the most opportunistic, and for the semblance of normality in traffic flow can be maintained. Collateral attractions are seen in the capabilities of immediately adapting new stations and routes and discontinuing those that are abandoned, some inherent security and adaptability to cryptography, various levels of priorities, and message switching of telegrams and data, which appear to be the most indispensable requirements in military communications. With the present very limited knowledge about adaptive routings, it is not clear how they compare with more traditional methods in respect of traffic carrying capacities and overload capabilities, as well as in overall cost; however, for military communications this latter factor is evaluated from different points of view than for commercial use.

A very challenging complete communication system, which among other things employs adaptive routing, was proposed by Mr. Paul Barran of the RAND Corporation of Santa Monica, California, and was studied in considerable detail by him and several associates. Their report was released by RAND in 13 volumes on "Distributed Communications" between 1964 and 1966. Summary and conclusion from this report were published by Messrs. B. W. Boehm and R. L. Mohley in the June, 1969, issue of the IEEE Transactions on Communication Technology, pp. 340 through 349. The "piece de resistance" of this study is in the adaptive routings: but, besides, it also offers several novel suggestions in respect to digital transmission, microwave, security provisions, multiplexing, switching, and simulation techniques. As to its economics, the proponents claim that because of the self-adapting capability overall system costs can be reduced by using less reliable components.

The Communication Net

The proposed network is essentially an enmeshed perfect grid structure of switching points, including terminal stations, and trunk lines interconnecting these. One of the characteristics of the described network is that each switching point is connected only to its immediate neighbors, although this does not appear to be requisite, and perhaps one may conversely define a neighbors those switching points that are directly interconnected any two of the network may be connected through an arbitrary number of switching points and connecting trunks, usually over a diversity of routes, of the approximate environment to which it belongs. Another characteristic of the described arrangement is that a trunk link seems to be made up of only a single channel, which must be capable of transmitting very fast a very large number of data bits. Again, this does not appear to be an innate characteristic of the method; instead a link could conceivably be made up of an arbitrarily large number of trunks according to traffic requirements. In the sequel, therefore, trunk and trunk group are interchangeably used.

The ratio of the number of trunk groups to the number of switching points in an infinitely extended array is an indication of the system's capability for route diversification. This ratio depends on the average number of neighbors, as defined above, directly connected to a switching point, and for the networks that were studied was found to be a maximum of four, beyond which one gets rapidly into the region of diminishing returns. Conceivably, this number may also have some significance on the number of alternates that one can efficiently use in a conventional hierarchical scheme.

In the proposed system, a message is chopped up into fixed block lengths of, say, 1024 data bits, transmitted serially one block at a time independently of another, interspersed according to the momentary demand by blocks belonging to other messages. The bits of each block contain the transfer value, which is the most significant feature of the process, and which will be discussed in detail little later.

Obviously, the process is not very efficient in message transmission, being tailored to communication methods other than voice, and in a way to message switching rather than to circuit switching modes. With these presumptions, the time taken up by the switching delays through an arbitrary number of switching points, plus the transmission delays through an arbitrary number of trunks, plus the reassembling of the several blocks of a message, cannot be objectionable, particularly since with the assumed high bit rate each of these delays is very short, perhaps just a fraction of a second. Furthermore, when message blocks must wait, the delay conditions are favorable, the individual block lengths being short and equal.

Considering its adoption to voice communication with circuit switching, it has been found (to mention an analogous example) that the transmission delays in telephone communications over stationary satellite links are not as objectionable as was previously assumed; nevertheless, they do pose definite limitations (two or more satellite links in tandem are not permissible), which undoubtedly would also be the case with the proposed adaptive routing system. But, as several times in the foregoing, fractionalizing of the messages into blocks may not be an essential condition for an adaptive routing arrangement, and therefore one should maintain an open mind while examining the potentialities of the proposed system, considering the use for which it was originally intended. The present discussion may not be an exact rendition of what the proponents had in mind, but includes also some "apropos" that have occurred to the reviewing Kibitzer.

At the empty virgin state of the network each switching point starts out with complete ignorance as to the state, or even of the existence of its neighbors; or of any other point or trunk group in the network, for that matter. Soon as traffic is being generated, each switching point learns enough information to be capable of routing its traffic efficiently. To do this, whenever a traffic parcel is to be dispatched at that point, it will choose the best outgoing trunk that is occupied. But instead of waiting for the best trunk to become available, if it just happens to be busy or otherwise inoperative, it chooses the second, third, etc. best that can be used.

Transfer Value

It remains now to define what is meant to be the best. The best route may be the physically shortest path, or the one with the greatest traffic handling capability, or the one that handles a unit of traffic at the lowest cost, or the best conditions are favorable; yardstick, or a combination of some of these which would bring about the overall most efficient and effective transportation value.

The before-mentioned "transfer value" expresses quantitatively the ranking of each route by the chosen yardstick, as an integer carried within each message block. It is blank at the outset, but may be incremented each time the message block passes through a switching point. The value of the increment could be equal to a value expressed by whatever yardstick is appropriate for the trunk over which the block is just
being transmitted, or modified in some way as may be prescribed by the specific rules of the network.

Each switching point includes a modest sized single-purpose computer whose principal output is a routing table of the complete network. It provides the current estimate of the transit time, or whatever other quantity may be selected to express the transfer value, to each destination in the network over each trunk radiating out of this switching point, and ranking these trunks accordingly in references to the destination.

When a message block arrives at a switching point, its transfer value, together with the original point of dispatch and the particular incoming trunk, are noted and processed in the computer for an eventual update of the routing table.

The transfer value of a recent arrival at a switching point from some originating station should be a current reasonable estimate for the transfer value of message blocks that would be heading in the near future for the same station. It indicates which neighbor of the switching point happens to be at this moment most directly or efficiently connected with the originating station. If the computer accepts this information, which it may or may not do, it will be used by the switching point after inserting it in the routing table, to select the preferential outgoing trunk over which a message to this destination should be directed.

Obviously, as traffic loads vary and trunks become inaccessible (or accessible) for one reason or another, the transfer values and ranking of trunk links are continuously revised in the routing tables, reflecting the state of momentary traffic flow in the network.

Modifications

Little consideration will show, and can also be demonstrated by simulation, that under certain conditions the entries in the routing tables will have a tendency to perpetuate themselves and become stale. This will particularly be true if for some reason the momentary "best" route happens to be a circuitous one and, due to prevailing somewhat light traffic conditions and lack of messages arriving from the particular point, was not reduced to reflect a more recent situation. For this reason, transfer values are not always accepted at par by the computer for updating the routing table, but will proceed, as was described, with a "skeptical learning and forgetting." In this process certain modifying factors are applied to the received transfer values before being used to update current values in the routing table. To provide the most satisfactory operations, the factors would have to be predominantly empirical, will differ according to whether they increase or decrease the increment, and will of course occasionally miss the target.

It was not intended at the outset to employ message storage. However, during simulation studies it was found that at times of somewhat light usage, storing a modest number of message blocks improves delivery, but that it is not very effective when trunk groups are heavily loaded. In the arrangement that was tried, the message block whose best choice trunk is unavailable is stored and dispatched during the next time frame. If the storage position is full, the block is steered to the next best trunk. If both the storage is full and there are no trunks, the message is dropped. A limited length storage can be helpful to span over short transients of noise and propagation fades.

With the sectionizing of messages into blocks, prolonged storage and roundabout routes should be avoided lest consecutive blocks of the same message could get out of step. Therefore, the transfer value entries in the routing tables must have a top limit, also established from experience carrying message blocks. Transfer values in excess of the maxima are considered undeliverable and dropped. Within the limited range of the computer simulations, this maximum was found to be 5 to 10 times the transfer value of the trunk route in the network. Shorter limits unnecessarily discarded usable messages, while long maxima defeated the intended purpose.

Perfect Routing

In a different approach to adaptive routing, the time required to reach a destination under ideal conditions is calculated by the various switching node computers from the structure of the network by using a so-called "dynamic programming" interactive method. In this system the computer must also store a map of the network besides the routing table. Changes in the availability of the network are detected by nearby switching points, from where the changes are propagated to other points of the network. The computers at the various locations then update their maps and the ideal routing tables.

This technique is said to be very suitable for small networks, it provides the perfect routing at almost all times, but as its size increases the demands on computer storage may soon get out of hand. Furthermore, it reduces the utility of the network as the size of the large volume of status messages flooding the system.

In addition to these disadvantages, the method, and particularly its routing table, is insensitive to prevailing traffic loads, especially while traffic flow prevails in one direction.

Simulations

With the above background, it may now be of interest to review briefly the conclusions gained from computer simulation studies. In reflecting on these conclusions, one must keep in mind their limitations. The simulations were reasonably thorough but by no means exhaustive when one considers the great variety of circumstances in potential applications; that the scale of the network was quite small, switching node grids being on the order of 8 x 7 or 5 x 10; that only single-channel connecting trunks were studied, applied in a message switching node transmitting short blocks of information with holding times of fraction of a second in real time; and, most importantly, that the main criteria for its worth were the capability to maintain traffic while portions of the network were destroyed, and the percentage of messages that were deliverable within a certain length of time, or delivered at all, during normal and during impaired conditions, instead of its general economy and efficiency during normally prevailing traffic and serve conditions as is usually desired in commercial operations.

As might be expected, the second, ideal routing method, where the routing tables are updated by dynamic programming, performed nearly perfectly as to the criteria given in the preceding paragraph; but as it was also mentioned previously it adapts itself poorly to variable traffic loads and its cost increases steeply with increasing size of the network. On the other hand, surprisingly, the first method in which transfer values of arriving message blocks were used to update existing routing table entries performed poorly, and got increasingly worse as the condition of the network deteriorated. However, when the updating was regulated by various empirical modifiers, the routing was more than 95% correct after processing 2,000 message blocks, even though half of the trunks were inoperative. Its message delivery performance was even better: after an initial processing of 500 message blocks nearly 100% of the messages were delivered within the maximum prescribed transfer time. The propitious views on this and similar performance results should be tempered by the fact that they were taken at a low 25% occupancy of the overall

526/5