MEMORANDUM
RM-3638-PR
AUGUST 1964

ON DISTRIBUTED COMMUNICATIONS:
IV. PRIORITY, PRECEDENCE, AND OVERLOAD

Paul Baran

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-3638-PR
AUGUST 1964

ON DISTRIBUTED COMMUNICATIONS:
IV. PRIORITY, PRECEDENCE, AND OVERLOAD

Paul Baran

This research is sponsored by the United States Air Force under Project RAND—Contract No. AF 49(638)-700 monitored by the Directorate of Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this Memorandum should not be interpreted as representing the official opinion or policy of the United States Air Force.

The RAND Corporation
1700 MAIN ST • SANTA MONICA • CALIFORNIA • 90406

PREFACE

This Memorandum is one in a series of eleven RAND Memoranda detailing the Distributed Adaptive Message Block Network, a proposed digital data communications system based on a distributed network concept, as presented in Vol. I in the series.* Various other items in the series deal with specific features of the concept, results of experimental modelings, engineering design considerations, and background and future implications.

The series, entitled On Distributed Communications, is a part of The RAND Corporation's continuing program of research under U.S. Air Force Project RAND, and is related to research in the field of command and control and in governmental and military planning and policy making.

The present Memorandum, the fourth in the series, is concerned with the establishment of traffic precedence doctrines designed to achieve optimum utilization of the communications resource, especially within a seriously degraded and overloaded network.

The proposed all-digital network has properties that differ in many respects from conventional communications networks. Network control features that can be performed only with great difficulty in conventional systems can be readily incorporated into a network of the type contemplated. While this Memorandum is oriented primarily

---

*A list of all items in the series is found on p. 61.

around the distributed digital network, the considerations are sufficiently general to be of interest to those concerned with the best utilization of an overloaded and impaired command and control communications network.

SUMMARY

In standard military communications systems, there is a natural upgrading of message precedence in times of stress. But, during the 99 per cent of the time of little or no stress conditions, the circuits are not conditioned for this high-stress traffic.

The Distributed Adaptive Message Block Network system delivers all traffic within the same time constraints, and that amount of time is within the limits of present emergency-level requirements. The circuits are prevented from overloading only by restriction of input--and no complete cutoff of input occurs to any user.

This Memorandum considers four separate techniques that can be used singly or in combination to achieve, through automation, "best" use of a seriously degraded and overloaded communications plant, within the framework of a rapidly changing organizational structure.

In the schemes considered, precedence is determined moment-by-moment, automatically for all traffic in the network. Precedence is computed as a composite function of:

1) the ability of the network to accept additional traffic;
2) the "importance" of each user and the "utility" of his traffic;
3) the data rate of each input transmission medium or the transducer used;
4) the tolerable delay time for delivery of the traffic.

During overload conditions, precedence status information is fed back through the network to limit the type and volume of data flow allocated to each user.

The dual goal is to prevent complete denial of communications to any network user while preventing network overloading. The surviving data rate, however meager under heavy network degradation, is always "equitably" rationed among the many network users. The definitions of "equitable" and "important" change from time to time, with control reserved to the commanding authority.

The examples used for describing the proposed notions are based upon the use of a high-data-rate, time-division, distributed system whose flexibility permits a rapid exchange of channel allocation between many users using few bits per second, and a few users using many bits per second. While the automated computational apparatus needed could, in all probability, be practicably implemented only in a future communications network, the concept is quite general and is easily visualized as being applicable to a pile of papers on a busy executive's desk, to a stack of computer programs awaiting processing, or to any other form of communications system.

ACKNOWLEDGMENTS

I am indebted to Jack Carne for many discussions that raised some of the notions included in this work. I would also like to acknowledge the continuing aid of Keith Uncapher and Wade Holland in the preparation of this series.

The parallel but independent work of Marvin Adelson at System Development Corporation is also to be noted. In an internal SDC working paper he has briefly suggested a somewhat related direction to the common problem. I would also like to acknowledge a conversation with Arthur Rosenberg of SDC on the subject, and to mention the excellent work of Col. A. J. Mandelbaum, appearing in reports of the Stanford Research Institute describing military communications traffic overload problems.*

I would also like to acknowledge many always stimulating discussions with R. H. Scherer, of the Office of the Director of Defense Research and Engineering, on communications, and in particular his statements on the differences between perishability and the importance of military traffic.

C. B. Laning of System Development Corporation, Marvin Adelson, now at the National Academy of Sciences, Jack Carne and James Farmer of RAND reviewed this manuscript and made many excellent suggestions which have been incorporated.

---

*Mandelbaum, A. J., Speed of Communications, Precedence and Traffic Control (U), Technical Report 2, Stanford Research Institute Project 2841, Menlo Park, California.

Inasmuch as the thoughts raised in this paper are highly speculative, and may be regarded by some as flights of fancy beyond the realm of justifiable experience, I alone must take the blame.

CONTENTS

I. INTRODUCTION

Simulation studies of switched communications network vulnerability, such as the RAND NATCOM model,* invariably include the assumption that any surviving link can carry all the traffic generated in the residual network. Since such vulnerability models have been so widely adopted, it is appropriate to consider some of the implications of the effects of overload on communications systems that are often ignored in the interest of simplification.

In this Memorandum we briefly touch upon the overload problem in present-day networks, reserving the bulk of our comments for improvements to a future system. Though this may appear to be ignoring the present in favor of a system that does not exist, the suggestions raised are intended primarily for a far-future time-frame. The general principles, however, might be used in such present-day problems as determining the order of program handling in a time-shared command-control computer.

In the process of describing the methods for automating the handling of mixed traffic of different levels of importance, data rate, and priority, simple mechanisms are mentioned which create an illusion of mechanized "judgment." Nothing magical is proposed; merely that a

---

*Eldrige, F. R., The Effectiveness of Command Control in Strategic Operations for the Mid-Sixties (U), The RAND Corporation, RM-3152-PR, October 1962 (Secret).

Reinertsen, R. W., The IBM 704 Computer Communications Vulnerability Model (FOUO), The RAND Corporation, S-135, August 15, 1960.

small set of executive policies can be automatically executed, with retention of the right to rapidly change the weights of these policies by human intervention. In essence, management's fundamental law of exception is mechanized so as to unburden humans processing information--nothing more.

We are dealing with two separate types of overload: terminal and network. If a called party is busy on the telephone, it may be said that the call is not completed because of an overload at the terminal. If the call is blocked by an overload at an intermediate switching center, or all circuits are busy, then the network is said to be overloaded.

There are two general categories of communications networks: store-and-forward, and line-switched. A torn-tape telegraph switching center which stores messages until a desired circuit is open is an example of the store-and-forward network. The conventional telephone network, which closes switches to provide a "real-time" path from user to user, is a line-switched network.

The overload phenomena have much in common, whether they be terminal or network, in store-and-forward or line-switching networks, and can be discussed on a general conceptual basis.

II. TRAFFIC OVERLOAD

A communications network can carry only a finite volume of traffic; if this volume is exceeded, a degradation in performance will result. If a local "step-by-step" civilian telephone office is overloaded and unable to accept a call, the busy signal is immediately heard. If the terminating central office is busy, the busy signal is returned shortly after the number is dialed. Usually, this delay is annoying but not critical. One can wait and try again.

Commercial communications networks are designed around an underlying assumption that each telephone is probably used less than two per cent of the time. A small amount of switching equipment can be safely shared by a large number of intermittent users.

It has been found, historically, that telephone subscribers are well satisfied if they are able to pick up the telephone and obtain service within ten seconds 99 per cent of the time. This level of service is called the probability time, P-T (0.01, 10); that is, the probability that service will be unavailable after ten seconds will occur only about one per cent of the time for any subscriber.

An additional doctrine of commercial common-carrier switched networks is that every subscriber shall receive the same grade of service.*

---

*R. I. Wilkerson's comments to M. Juncosa and R. Kalaba, in "Optimal Utilization of Trunking Facilities," Communications and Electronics, No. 40, January 1959, p. 1002.

Ninety-nine per cent service may sound like excellent communications service. But, to the military user such a network leaves much to be desired. The user has no choice as to exactly which one per cent of the time service is to be denied. The one per cent failure time must be expected to occur when there is an abnormally heavy demand upon the network. In civilian networks this usually happens during unexpected snowstorms, widespread fires, hurricanes, floods, etc. This can be tolerated since the goal of the commercial telephone utility is to provide service at the lowest cost most of the time; it is not basically intended for general emergencies.

Military crises, almost by definition, place abnormal loads on systems. When using a communications system designed under such civilian loading assumptions for military purposes, one can expect most service denials to occur precisely during those times when most needed. Though this may sound like a recitation of the obvious, the implications of such underlying ground rules have, on many occasions, been unappreciated by some planners of systems for the military.

This common, implicit, system-design assumption is particularly treacherous, as there is little opportunity to discover its existence under normal test operation. Communications networks are rarely exercised in real-time to simulate extensive communications network damage and overload.

STORE-AND-FORWARD OVERLOAD

Overloading is one of the causes of breakdowns in store-and-forward systems during military crises. During periods of high tension, not only does command-control traffic increase, but even logistics traffic is generated in greater volumes. In a crisis almost everyone feels obliged to communicate: the heavy increase of low-priority logistics traffic has been called the "underwear ordering" effect. The crisis will evoke a flood of backlogged requisitions into the system, all demanding immediate processing.*

A significant increase can usually be accepted by the local communications tributary station for processing; the bottleneck occurs farther downstream. Present-day "hard copy" written-text military communications networks are slow-speed store-and-forward systems. Long-time intermediate storage is used at the switching nodes to improve high-cost long-line-circuit usage. When the traffic volume arriving at the intermediate switching center from the many feed points is greater than the output circuit can handle, messages must be backlogged.

There are several different military precedence systems in use, which, theoretically, insure that the more urgent traffic is processed first. For example, in the BIX, Binary Information Exchange, system the precedence categories of "Flash," "Emergency," "Operational Immediate," "Priority," "Routine," and "Deferred" are used. These

---

*The ex post facto explanation heard from the requisitioner is, "If war is coming up, we better get all our requisitions in the mill so we will be ready to fight."

categories are divided into two subsets, "High Precedence" and "Low Precedence," each handled by a different set of rules within the store-and-forward switching center.

The earlier UNICOM system originally considered only three grades of precedence, "Right-of-Way," "Priority," and "Routine"; more recently this program has incorporated the use of four precedence grades. The Defense Communication Agency has standardized on the set shown in Appendix A.

Each user of the conventional military store-and-forward system is responsible for suitably marking his own messages fed into the network. The military communicator's goal is to process all highest-priority traffic so that it is delivered to the addressee within X minutes of transmission; all of the next grade traffic within Y minutes; etc.* Since these times cannot be met under heavy traffic conditions, a downgrading procedure, called "Minimize," is called into play (see Appendix B). "Minimize" attempts to reduce the volume of high-precedence traffic. Although helpful, the procedure has not been wholly effective in reducing network overload. The writer has heard it said that the following chain of events occurs: Messages inserted into the network are delayed enroute by a time factor unknown to the originating party. Not having any confirmation of receipt of his urgent message, the originator panics and sends another message--at an increased precedence level. Still not receiving an answer, the originator again panics and dumps still more high-precedence traffic into the network.

---

*The values of X, Y, etc., apparently are <u>not</u> standardized for <u>different</u> systems.

This mechanism may be viewed as being akin to a fast-acting servo controller connected in a loop with a long signal feedback lag-time; oscillation can be expected. A communication network that does not let the user know how long it will be before his message will be delivered to the end addressee may be theoretically oscillatory.

LINE-SWITCHING OVERLOAD

A similar overload mechanism is sometimes observed in the civilian telephone system during unexpected civil crises. A caller dials a number and hears a busy signal. Upon receipt of the busy signal, he dials again. Again, hearing the busy signal, he becomes more impatient and once more ties up the shared central office equipment. In extreme cases, the telephone company may evoke a doctrine called line-load control. This consists of intermittently cutting off blocks of subscriber lines feeding the overloaded offices. Line-load control not only provides better service to the remaining users, but safeguards against a complete central office failure by overload of the time-shared common-control markers used in some dial systems. Such panic-induced failures, tying up an entire central office for several hours, have been noted. (The telephone utility in its efforts to maximize continuity of service in the face of overload, calls a series of separate techniques into play as the load builds up near the critical point; these are described in Appendix C.)

III. THE DISTRIBUTED ADAPTIVE MESSAGE BLOCK NETWORK

In order that a military communication network die gracefully, it must overload gracefully. The proposed broadband distributed network described in this series has several orders of magnitude greater communications capability than any existing military network. Although overloads would not normally be expected in such a network, we must still plan for such an eventuality.

In the Distributed Adaptive Message Block Network, all traffic from each originating circuit is chopped into small blocks of data called Message Blocks. Each Message Block is "rubber stamped" with the symbols of the end destination, the originating station, and some other housekeeping data. These Message Blocks are transmitted from Switching Node to Switching Node, eventually reaching the desired end station by a reasonably efficient path.

In order to handle real-time digital voice transmission, it was necessary to limit the amount of in-transit storage at each node to minimize differential-path delay times. Because of this restriction on the amount of storage capacity at each node, consideration was limited to only those routing doctrines that utilized little storage at the Switching Nodes. While such doctrines were expected to be less efficient in line utilization than the doctrines which allow backlog of many bits at each switching center, under simulation it has been found that little circuit-utilization performance is lost. Most of the high line-utilization of store-and-forward can be had with very little in-transit storage. By limiting the amount of

storage, the store-and-forward system can be made to exhibit one of the most useful features of a line-switching system; i.e., avoidance of store-and-forward oscillation by immediate confirmation of receipt of message. This is accomplished while retaining most of the store-and-forward system's advantage of greater alternate-routing capability and equitable sharing of a single channel by a plurality of users.

HOT-POTATO ROUTING

In the distributed network routing scheme under consideration, the policy is used that if the preferred path is busy, the in-transit Message Block is not stored, but rather sent out over a less efficient, but non-busy link. This rapid passing around of messages without delay, even if a secondary route must be chosen, is called a "hot-potato" routing doctrine. (Each node tries to get rid of its messages as if they were "hot potatoes" and the node is not wearing gloves.)

With such a doctrine, only enough storage at each node need be provided to permit retransmitting messages if an acknowledgment of correct receipt is not received from the adjacent station within a prescribed time interval. Message storage capacity is modest.*

---

*For example, if eight, separate, full-duplex, 1.5-megabit/sec, 150-mi links feed a single Switching Node, only about 32,000 bits of storage are required at each node. Present estimates are that it will take on the order of one millisecond for a 1024-bit Message Block to be read into a node, for a decision to be made as to the best direction in which to route the Message Block, and to start the Message

THE LINE-SWITCHING ILLUSION

Although the broadband distributed network is described as a store-and-forward system (which it is, as far as internal network operation is concerned), the fast switching time of the hot-potato doctrine presents the illusion of a virtual circuit, having a short delay, between two users. While it is anticipated that this mechanism will eliminate the "open servo loop" store-and-forward delay problem discussed above, this is only part of the problem.

Figure 1 portrays several Message Blocks arriving at a distributed network Switching Node. Each node contains a routing decision mechanism. Messages are shown to arrive simultaneously from N different lines. Three of the messages--those coming in on Lines 1, 3, and "N"--are all to be delivered to Station A. Line 2 carries a message directed to Station D, while Lines 4 and 5 carry messages which are to terminate at Stations E and F, respectively.

Block on its way to an adjacent node. Thus, if a Message Block has to traverse 30 such nodes in going across the U.S., only about 30 milliseconds will be required to pass through 30 Switching Nodes (plus another 20 milliseconds transit time through 4000 mi of transmission lines at 133,000 mi/sec. Thus, we anticipate that it may take only about 50 milliseconds to transmit data across the U.S. This is less than the delays encountered in forming digital voice into Message Blocks and unpacking the Message Blocks into audio voice.

This subject is discussed in detail in ODC-VII, -VIII. (ODC is an abbreviation of the series title, On Distributed Communications; the number following refers to the particular volume within the series. A list of all items in the series is found on p. 61.)

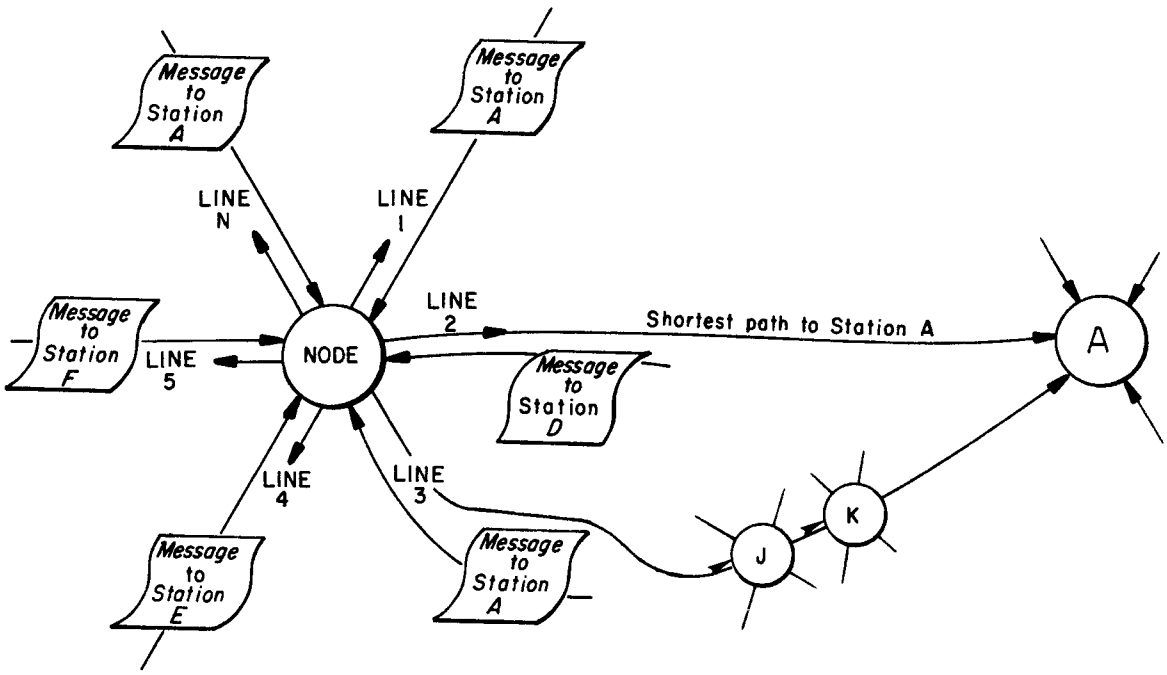

Fig. 1--Message Blocks Arriving at a Switching Node

A dynamically-updated routing table stored at each node indicates the best route for each Message Block to take to reach its directed end terminal station. In the example, three separate messages are received simultaneously at the node, each wishing to take the "best" path toward A. The shortest route is via Line 2, but all three messages cannot be sent out simultaneously, so two of the messages must travel over alternate paths. Although we could have built a system that stored the currently undeliverable messages until the line to A cleared, we have chosen to limit ourselves to a system which sends the "overflow" messages over less efficient paths, effectively preventing a common type of overload (see ODC-II, -III). When two messages seek the same preferred line, a random choice is used to select which message is sent out over the best path. In other words, if three signals are all to be directed towards A, it is not felt that it makes much difference whether the first-choice, second-choice, or third-choice path is taken, since only a few extra nodes must be traversed before the end station is reached. Simulation has shown that this use of secondary paths in lieu of storage is a surprisingly effective doctrine, enabling transmission of about 30-50 per cent of the peak theoretical capacity possible in a store-and-forward system having infinite storage and infinite delays.

AVOIDING NETWORK OVERLOADING--CHOKING

"Choking" is a policy automatically executed by the computer logic at each Switching Node to cut off new local traffic to a network approaching overload. A sample of

such a choking doctrine is that shown in Figs. 2-4 and described below.

Each node, as shown in Fig. 2, has N input lines feeding traffic into the node and N output lines, all carrying traffic away from the node. In the example, N = 8.

1) There is an input line and an output line for each remote station.
2) I_i = input data rate for the i'th line ($i = 1,2,\ldots,8$).
3) O_i = output data rate for the i'th line ($i = 1,2,\ldots,8$).
4) Compute

$$B = \sum_{i=I}^{7} I_i - \sum_{i=I}^{8} O_i \quad ,$$

which is the "backlog" traffic passing through the node. B should always be a negative number--otherwise, more traffic is piling up than is being carried away.

5) Compute L = f(B), where L is the volume of locally-generated traffic that may be allowed to enter the network.

If all or almost all of the output lines are busy, the local node avoids feeding new traffic into the network until there is open network output capacity. Implementation of this simple intermittent-connection choking doctrine in the illustration presents a varying network input capacity to the tributary communications center feeding Line 1, proportional to what the network can comfortably carry. Network simulation described in ODC-II and -III confirms that

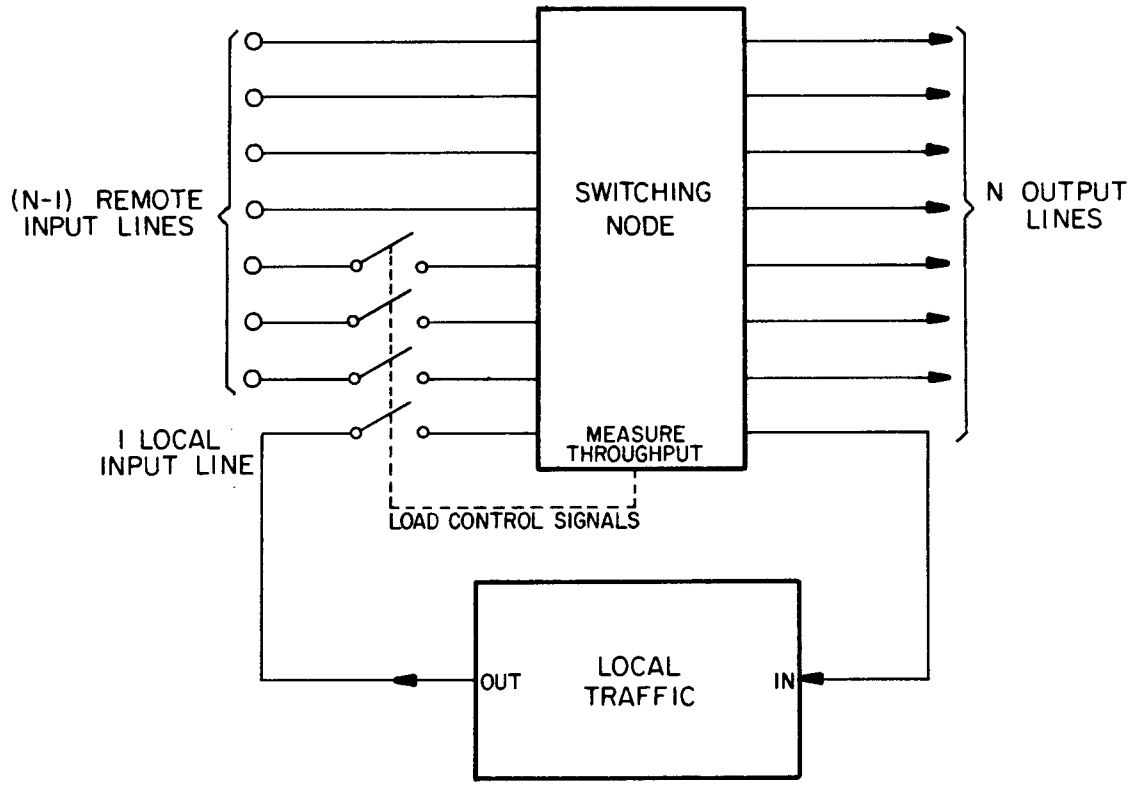

Fig. 2--Choking Input Traffic to Switching Node to Prevent Network Overloading

a simple policy that chokes off input signals in a binary manner is sufficient to protect any station or combination of stations from overloading the network while allowing each station to retain a high input rate.

Figure 3 represents a variable volume of traffic passing through a sample node. Each of eight input lines is assumed to be able to carry a 1,500,000-bit/sec peak capacity, with the following sample choking doctrine imposed: If the total traffic in the network approaches a volume greater than a preset value, new input will be prevented from entering the network. Traffic already in the network is given preference over newly-entering traffic. Once traffic has entered into the network, the user is almost guaranteed delivery with the certainty that his own traffic will not cause a local bottleneck at another switching center and get bogged down in transit.

In Figs. 2 and 3 the number of lines in simultaneous use is measured to provide a rapid estimate of those intermittent periods when new traffic can be safely entered into the network. The measuring detector in this case might be a simple majority weighting element responding, in this case, to a threshold of five out of eight. If the traffic volume is less than threshold, the full input traffic capacity at the link is allowable. But, when the nodal threshold is exceeded, the input is cut off in a binary manner, as shown in Fig. 4. The local user sees these very short on-off periods smoothed out, with the network input capacity appearing as a slowly varying allowable data rate; i.e., the maximum input rate that can be fed into the network without causing an overload.

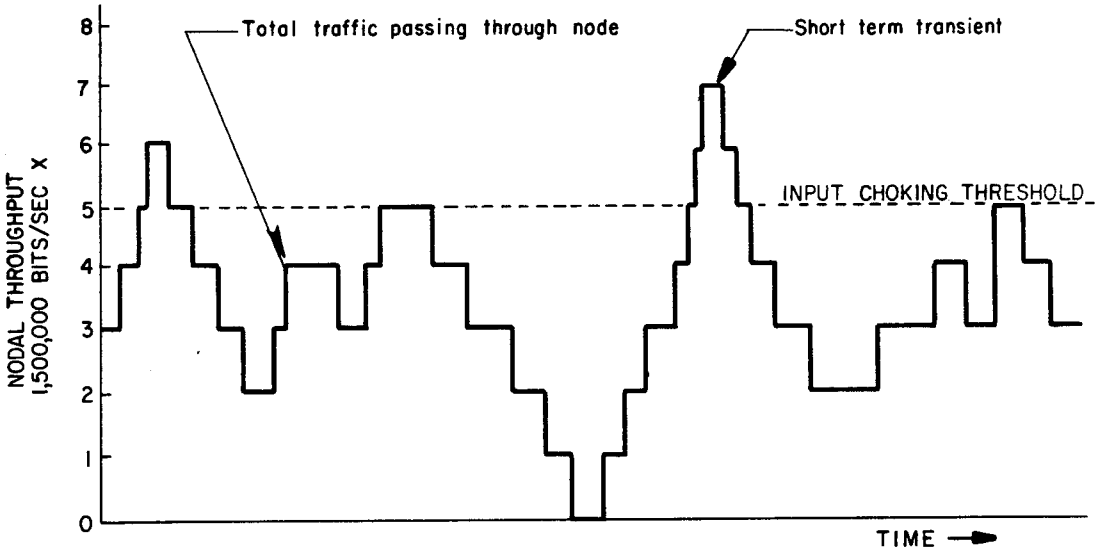

Fig. 3--Traffic Load Through Sample Node

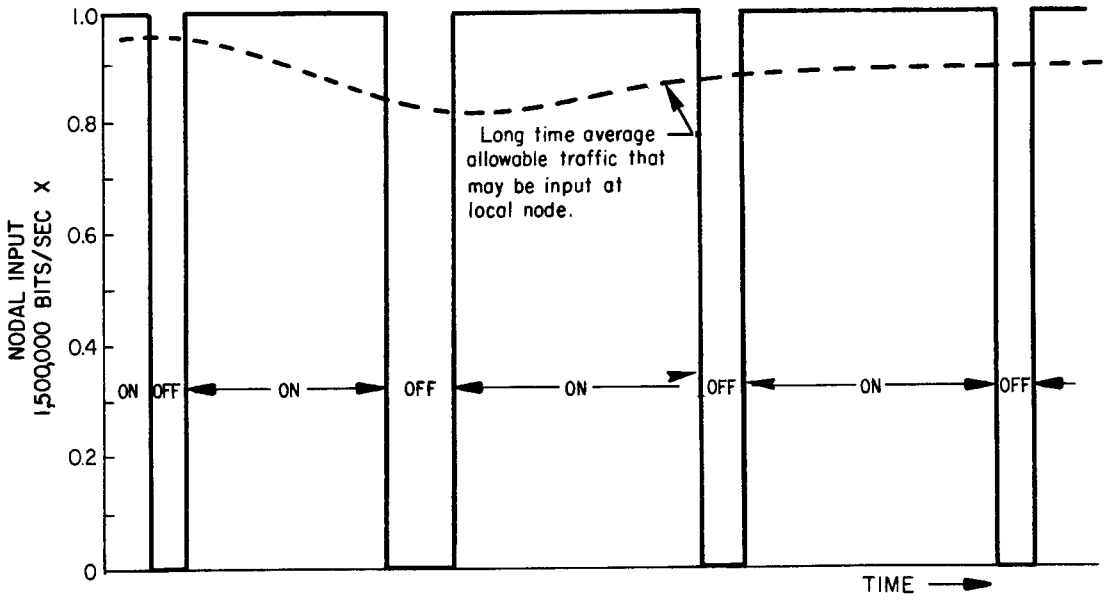

Fig. 4--Intermittently-Connected Local Input
to Limit Average Local Data Rate

The next question we shall consider is allocating this total locally-allowable data rate among a large set of local users or "subscribers."

THE MONETARY ANALOGY

Each local communications center in a communications network can only feed up to a certain number of bits per second into the network without overload. This total variable volume may be thought of as being a communications resource, or a "currency," measured in bits rather than dollars. This currency can theoretically be spent in any manner desired. At present, in military networks we generally allow each user to demand as much of the limited resource of communications capability as he alone desires, limited only by the precedence of his traffic. During times of crisis, military commanders will try to spend more communications currency than exists, and there will be an effect identical to inflation. Messages once labeled "Deferred" will be stamped "Operational Immediate," and jammed back into the input hopper. It is analogous to the inflationary competition of competing buyers for scarce goods. We temporarily delude ourselves into thinking that we are buying more capability by inflating the precedence indicator. It is only when we are forced to face up to reality by a currency devaluation that we appreciate what has happened. The "Minimize" doctrine, described in Appendix B, can be seen to be analogous to a "currency devaluation." In designing a priority handling system, we should never permit ourselves to believe that

we have more (or less) usable communications capability than we really have. This implies network status control feedback loops.

IV. USER CONTROL

A tributary communications station is a traffic concentrator. Messages or lines from many individual users are combined at the communications office, thus making most economic use of an expensive and limited communications facility through time-sharing.

Although communications centers of the future can be smaller and more completely automated, we would like to retain human judgment to allow redistribution of the communications resource during crises.

THE COMMUNICATIONS CONTROL CONSOLE

The job and the equipment we describe does not exist today. At each local communications center we propose there be a "responsible" military officer who is charged with local traffic control and who is responsible for changing the gross allocation of the communications resource when necessary.

The Communications Controller is to be an arbiter with the task of allocating the available data capacity or communications resource among the many local users of the communication network.

To aid the Communications Controller to allocate surviving capability equitably among individual demands, a Communications Control Console is proposed. This console provides a picture of both the availability and the demand for communication service, together with devices for its sub-allocation. The commander views his communications capacity (resource), observing present allocation

(of currency) and the size of individual demands (spending) for service.

Figure 5 is an illustration of a hypothetical Communications Control Console.* A portion of this console is shown in more detail in Fig. 6. The meter in Fig. 6 indicates the summation of total demand by all local subscribers of channel capacity in the common measure of bits per second. The right-hand meter indicates the network's ability to absorb new traffic in bits per second after choking.

Since traffic already in the network is given precedence over new traffic, overload is prevented by the diminution of acceptance of new traffic. Thus, the total input data rate from all local users into the network must be limited until an equilibrium point is reached between the traffic that is admitted into the network and the traffic that the network can rapidly deliver. Each communications center need not use its full allocation of capability for feeding traffic into the network. Conversely, a commander at a particular remote site at a crisis point may handle a heavier-than-normal communications load--if necessary, by requesting nearby stations to reduce their traffic input in order to allow heavier local use of the network capability. Flexibility is thus reserved to allow for the many different centers of various importance which utilize the network.

Further, each individual network user ("subscriber") can ask the Communications Controller for more than his

---

*This is the Priority Control Console described in detail in ODC-VIII.

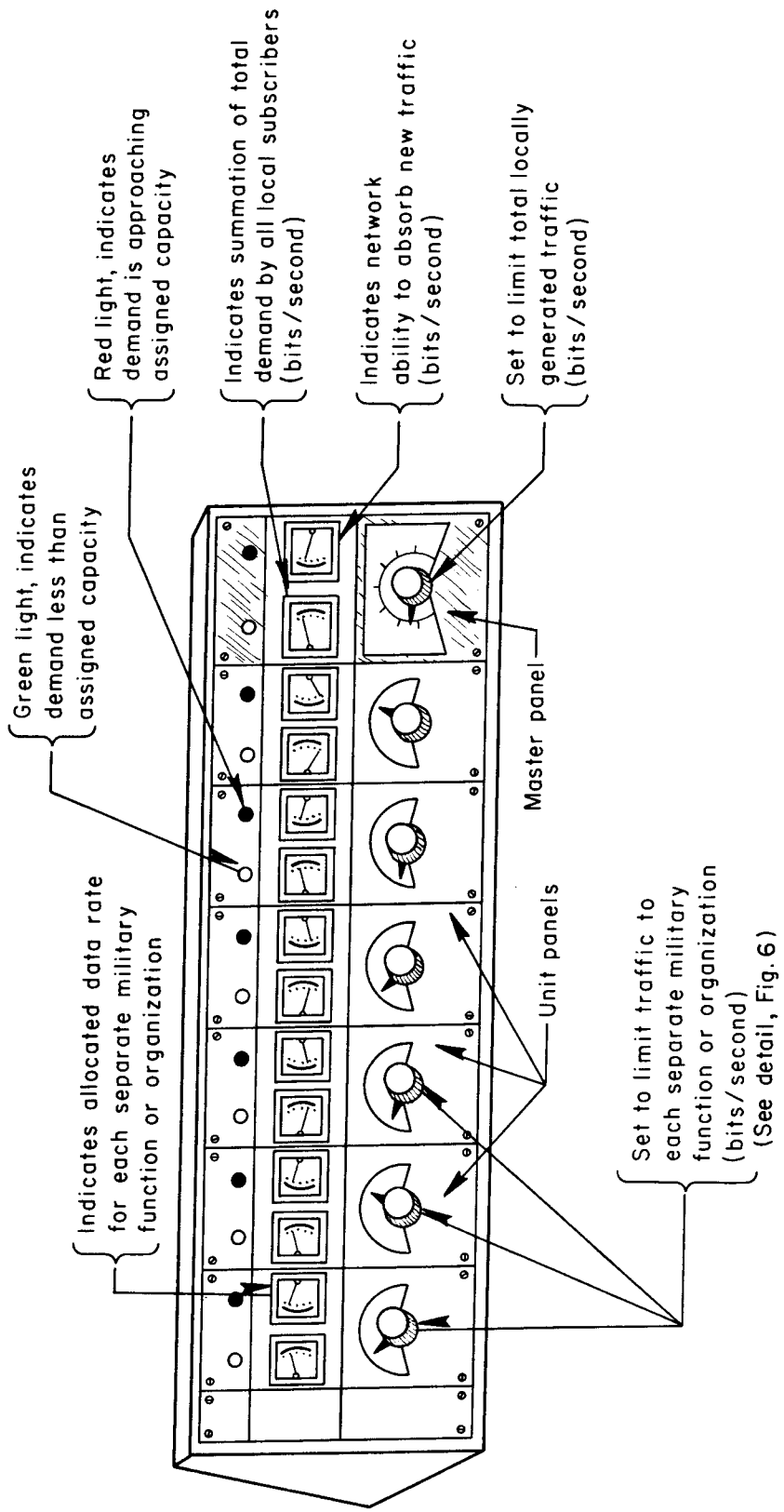

Fig. 5--The Communications Control Console (Priority Control Console)

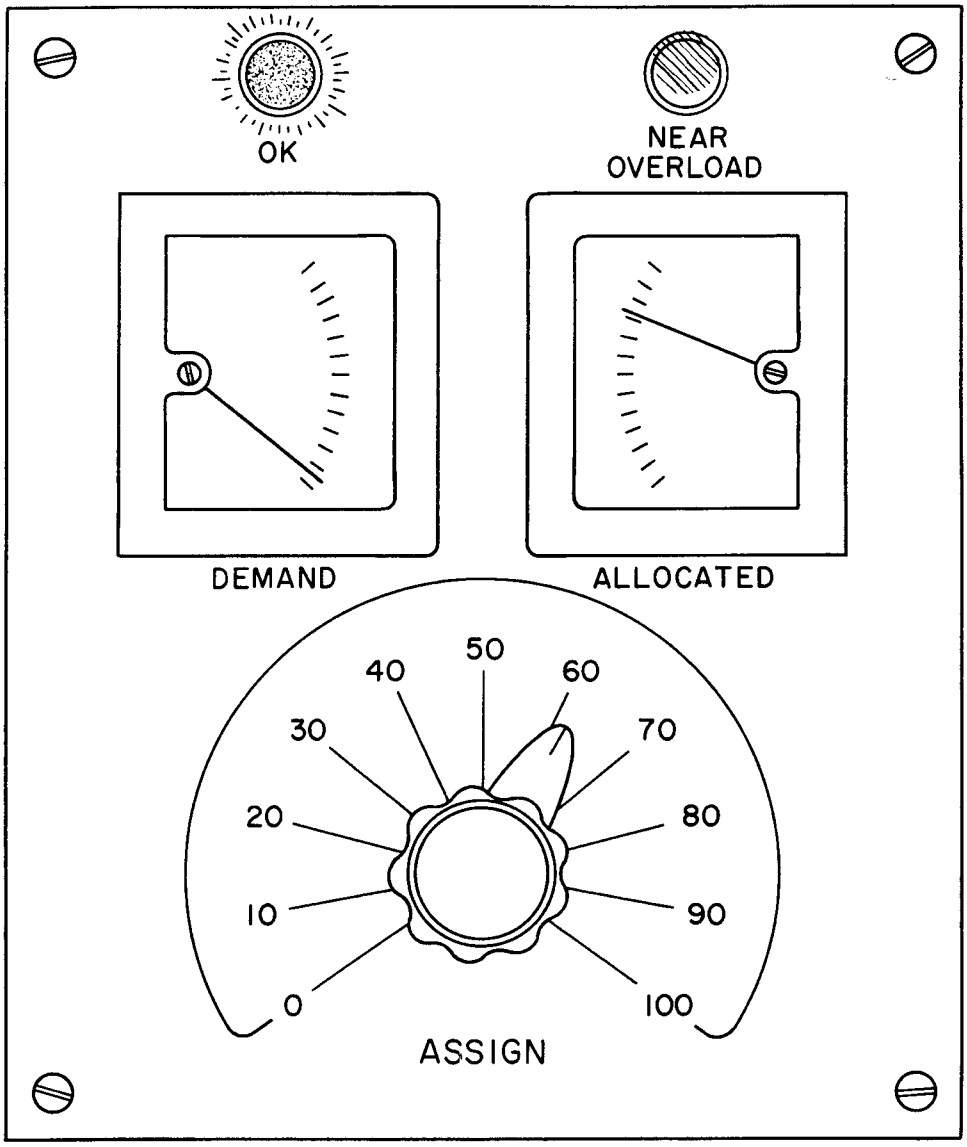

Fig. 6--Detail of Communications Control Console:
Traffic Monitoring and Control Unit
(See Fig. 5)

currently allocated share of the communications resource. This change of assignment is effected by the Communications Controller turning a dial, labeled "Allocated" in Fig. 6. This control sets the limit to the total volume of locally-generated traffic (in bits/sec). If, for example, a station was requested to cut its loading to 50 per cent of its allocated capacity, the dial would be set accordingly.

In Fig. 5 two lights are shown above each panel. The green light (left side) indicates situation normal: demand safely under the allocated capacity. The red light (right side) warns that traffic volume is approaching the overload point. Each local Communications Controller in turn can ration his assigned total communications resource among his many competing local users. For the sake of illustration, assume that each user performs a vastly different function; for example, intelligence, early warning, logistics supply, etc. To allow control of these separate, further-subdivided, functional categories, individual panel units similar to the master panel are provided. The first such unit might control the gross volume of intelligence traffic; the second, early warning traffic; etc. Thus the Communications Controller would also have at his fingertips the ability to vary the allocation of data rate to each separate function.

The Communications Controller will not play the console as an organ, since gross changes in loading are slowly-occurring phenomena; rather, he will normally leave the controls set to fixed positions, except when a crisis or overload approaches as indicated by the red warning light.

He then decides which users with growing demands should deprive others with less important duties, and to what degree. (Every user who desires more communications capability than he is momentarily allocated is always allowed to talk to the Communications Controller to present his case for an increase in allotment of bits.)

The console does not seek to supplant human judgment. It simply provides an automated facility to instantaneously implement the human executive decision. It administers the will of the commander swiftly and without an arbitrary cutoff. No individual function need ever be totally deprived of instantaneous communication, as is necessary with the binary decision rules necessary to enforce today's precedence doctrines.

AN ANALOG MODEL OF THE CONSOLE

Figure 7 is a sketch of a schematic analog-signal model of the Communications Control Console. This hypothetical implementation is included to suggest that the circuitry required to display the control traffic loading need not necessarily be complex.* In Fig. 7, an AC input voltage proportional to local available data rate is inserted at the terminals in the upper-left-hand corner of the illustration. A rotary autotransformer, T-1, is connected to the control knob of the unit. A second voltmeter, V-2, is connected to the output of T-1, to show the reduced amount of data rate that may be utilized locally. Each

*The all-digital version of this console is described in ODC-VIII, where all signals necessary for the console are included.

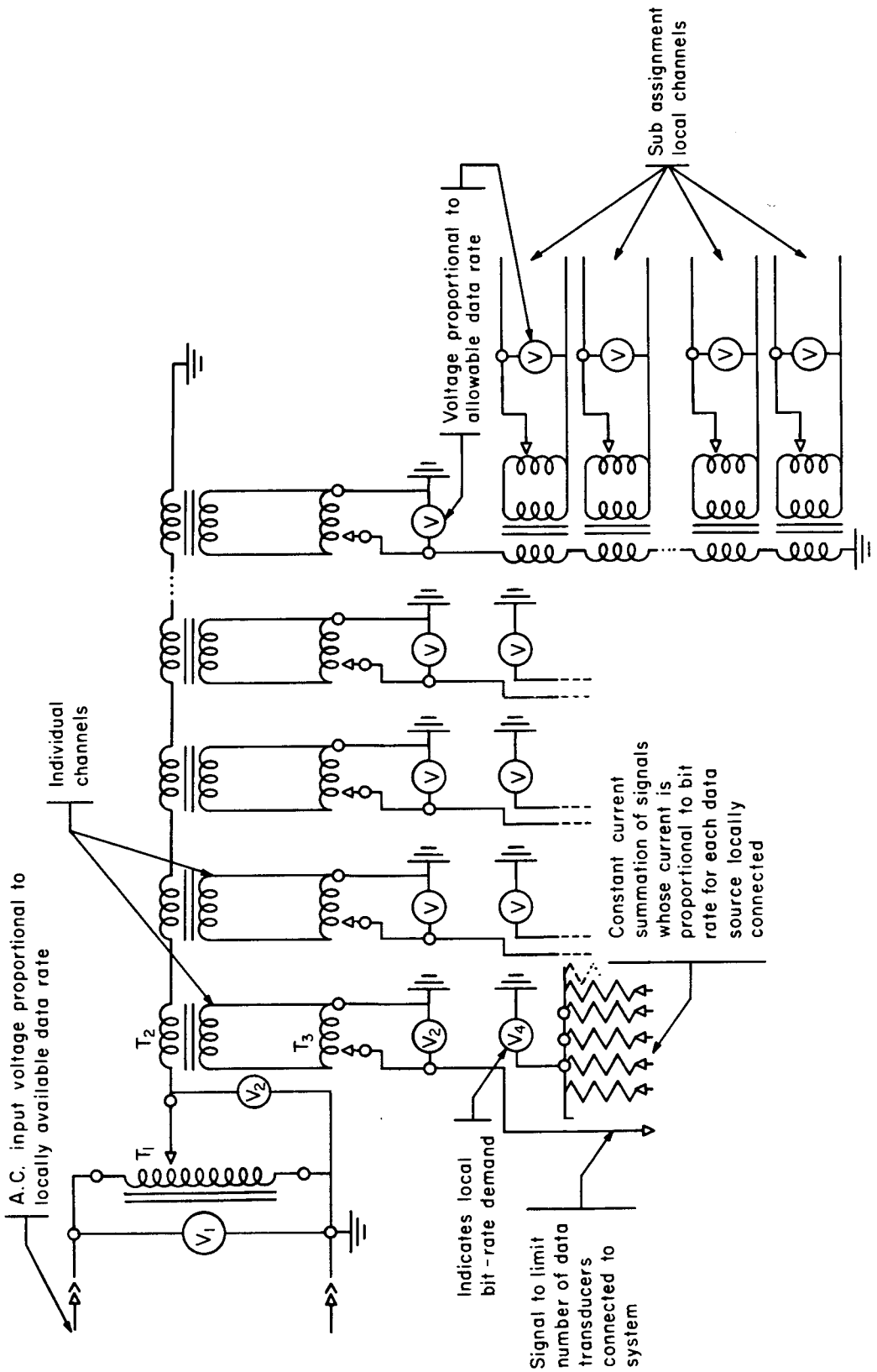

Fig. 7--Traffic Monitoring and Control Unit Circuitry

unit panel is connected to a separate autotransformer. Each autotransformer, in turn, can be connected to lower-level sub-assignment autotransformers. These autotransformers create voltages proportional to the assigned data rate of each of the individual unit panels. Voltmeter V-4 indicates the local bit-rate demand. This signal can be approximated by a simple constant-current summation of binary control signals whose currents are proportional to the bit-rate demand for each data source energized.*

For example, if four teletype machines are connected to the system, four separate resistors, each passing a current proportional to the data rate required by each, are used. Each time a teletype machine is energized, an associated relay closes attaching an AC voltage to the resistor associated with the connected teletype. If all four teletypes are in use, the currents passing through all four resistors would be summed, producing a voltage whose value would be approximately equal to the total bit-rate demand of all four individual teletypewriter devices.

On the extreme right-hand side of Fig. 7 it can be seen that each single-unit channel can be subdivided into other smaller sub-assignment panels. These panels can be connected to match the same hierarchical organizational structure of the network feeding the local communications transducers to the communications tributary station. (Nothing, of course, restricts any user from being connected to more than one communications tributary station.)

---

*In practice, this would be done at the Multiplexing Station which already has equipment to measure activity.

V. MEDIA CONTROL

To this point, we have described how available data rate, a volatile commodity whose measure is in bits per second, may be allocated among the many users of a communication network. We shall next consider a method to encourage each individual user to make most efficient use of his allocated communications resource.

The common-user digital network of the future encounters a wide variety of data-generating devices. These devices can load the network at rates differing as much as 100,000 to 1. Some of these devices are briefly described in Appendix D, and include telephone, teletype, telegraph, and facsimile. Each of these media places a different bit-rate load upon the communications system.

In the civilian world, the decision of choice of medium tends to be primarily economic. We ordinarily use the cheapest medium commensurate with the value and perishability of the information. In the non-dollar economy of the military world we need a mechanism analogous to the concept of money to encourage each individual user to select the medium which makes most efficient use of the total communications resource.

The reader is cautioned that the following development is not the result of rigorous analysis, inasmuch as the analytical tools are simply not available. But, in their absence, we have chosen the following approach; the numbers used are not important since their purpose is illustrative only.

Let us start with a tentative assumption that each alternative communications medium conveys roughly about the

same amount of "information theory" type information per unit time under real-time, human-to-human, or human-to-machine information interchange. This is not too difficult to visualize: For example, humans speak and type at the same rate, about 60 words per minute. But, in the information theory sense, such language is highly redundant. Tests have been run to determine the non-redundant information input rate of humans, and an extensive body of literature has developed examining the input data rate for man's input senses. The results of these tests indicate that the maximum possible human input data rate is probably less than about 50 bits/sec, based upon tachistoscopic picture element recognition (visual input), language text (also visual input), voice (auditory input), and vibratory Morse code (tactile input).

As "information" can be sent in a variety of ways, we wish to encourage each user always to use the most "efficient" medium. During peacetime, it might be most "efficient" to send a letter by facsimile if transmission bandwidth is cheap and keyboard operator's time expensive. However, during overload the available supply of the communications commodity decreases, and, assuming a free economy, the "price" should change.

Under these conditions, one would like to encourage the use of more-efficient, narrower-band transmission devices, such as, say, teletype. We would like to have a "cost" table of these different data sources in order to provide a basis for determining the inflation-free price of service when the communications resource is to be rapidly rationed (see Table 1).

Table I

DATA RATES FOR DIFFERENT DIGITAL GENERATING DEVICES

		No. Bits/Sec (Bauds)
TELETYPE	5-Unit	60
	Fieldata	80
VOICE	19.2 kilobits	19,200
	2.4 kilobits	2,400
FACSIMILE	9.6 kilobits	9,600
DATA	Keyboard	40
	300 c/m Card Reader	5,000
	High-Speed Card Reader	20,000
	Partial Core Dump (10%)	100,000
	Entire Core Dump	1,000,000
	Magnetic Tape Transfer	200,000

Data rate alone, however, does not provide a complete measure of network loading; some devices have a short duty cycle, such as one computer sending the contents of its core to a remote computer. While such devices place a heavy peak demand for service, they are highly intermittent. On the other hand, a pulse-coded telephone call places a lower peak demand load, but ties up network capacity for a longer period and results in heavier average loading. Therefore, we should include an expected message-duration or holding-time factor in the network-load weighting table.

Two separate factors are at work here. Many separate low-data-rate devices time-shared or concentrated into a single high-data-rate link permit better averaging, as compared to a few correspondingly-higher-data-rate users. But, as many of the high-data-rate users "get in" and "get out" fast, they have a short holding time. This helps the averaging process. To be precise in this computation, a better understanding of the number of users, their use statistics, and the network characteristics appears mandatory, and shall be deferred until such information is available. It is sufficient for purposes of explaining the concept to use the following tentative rationale in preparing the sample loading chart of Table II.

Start with the data rate in bits per second of the transducers which we have briefly considered. Next, give "credit" to those devices expected to be in use for only a comparatively small portion of the time. Column 1 lists the peak data rate in bits/sec for each of the input devices. Column 2 lists the percentage of network users using any single type of data input device. (If, for

Table II

APPROXIMATION OF NETWORK LOADINGS FOR
DIFFERENT DIGITAL DATA GENERATING DEVICES

		(1) Data Rate, Bits/Sec (Bauds)	(2) Per-centage of Users	(3) % Time in Use	(4) Data Rate x Users, %	(5) Long Time Average Loading	(6) Rank Order Loading
TELETYPE	5-Unit	60	3	100	180	60	2
	Fieldata	80	5	100	400	80	3
VOICE	19.2 kilobits	19,200	60	100	1,152,000	19,200	10
	2.4 kilobits	2,400	10	100	24,000	2,400	7
FACSIMILE	9.6 kilobits	9,600	4	100	38,400	9,600	9
DATA	Keyboard	40	10	20	80	8	1
	300 c/m Card Reader	5,000	3	10	1,500	500	5
	High-Speed Card Reader	20,000	2	20	8,000	4,000	8
	Partial Core Dump (10%)	100,000	1	0.1	100	100	4
	Entire Core Dump	1,000,000	1	0.1	1,000	1,000	6
	Magnetic Tape Transfer	200,000	1	30	60,000	60,000	11

example, ten per cent of the links served 40-bit/sec keyboards, the value of ten per cent would be shown in Col. 2.)

Column 3 shows the expected maximum duty cycle of each input device averaged over a long period. (A telephone could conceivably be used 100 per cent of the time, while it is doubtful that a computer core dump would occur more often than 0.1 per cent of the time.)

Column 4 is the product of data rate and users, and provides an indication of the loading demand by each type of input source. Column 5 lists the rank order of entries of Column 4 to indicate the type devices that make heaviest average demands upon the data resource. Thus, for example, in our network we will expect many digital telephone users but few computer core dumps.

This determination of allocation of loading demand can be further refined by including the following two factors.

First, we could use the investment cost of each input device as a factor indicating its "importance." Thus, for example, another multiplicative factor could be included to allow for the higher investment cost of computers, relative to telephones.

Secondly, we could also use the reciprocal of the total number of each data input device as the metric. Using this rationale in a distributed command-control system, the true vulnerability is probably some function of the reciprocal of the number of times a specialized input device is replicated in the network. For example, if three special-purpose computers are tied together with a communications

network, destruction of a single computer installation (or its communications) is correspondingly more serious than if the complex contains fifty computers sharing the work. In the first case, there would be loss of one-third of the complex's capacity; in the second, only 1/50 capacity. Thus, we might wish to tend to favor the more "critical" elements as compared to the less unique.

The metric being developed as a measure of the communications resource is a vector having many components. To this point, we have described a few of the components. We shall now consider yet another component--the conventional military precedence indicator and how it might be blended into the vector. The present-day precedence indicator system concept is based primarily upon the speed of delivery of a message. Historically, it has probably grown out of the old commercial telegraph tariffs; i.e., telegram, deferred, day letter, and night letter.

Column 1 of Table III lists present Defense Communications System Precedence Categories, together with target processing times. Column 3 lists the approximate ratio of these time categories. An underlying consideration in the following development is the mixed requirement that, while we wish to give priority treatment to the higher-precedence traffic of equal network loading, we must also satisfy the goal that we preserve a minimum transmission capability for the lower-precedence traffic. Thus, instead of a blanket rule that all traffic of a given precedence grade will be transmitted before handling the next lower precedence grade, we choose to use the time ratios of these precedence categories to act as a preference weighting factor.

Table III

USE OF SPEED OF SERVICE CRITERIA TO ESTABLISH
A NUMERICAL WEIGHTING PRECEDENCE FACTOR

Precedence Categories	DCS Speed of Service Criteria[a]	Time Ratio	Precedence Factor
Flash	~ 0	1[b]	1
Emergency	30 min	10	10
Ops. Immed.	>30 min < 1 hour	10-20	20
Priority	> 1 hour < 5 hours	20-100	100
Routine	> 3 hours < 8 hours	60-160	160
Deferred	> 8 hours < Start of next day's business	160-300	300

[a] See Appendix A.

[b] Based upon message length of 125-150 groups @ 60 wpm ≃ 3 min or 1/20 hr.

Table IV combines the network loading factors for different digital services, shown in Table I, together with the numerical values of preference weighting factor derived from Table II.

Table V is included to illustrate certain combinations of "lower-precedence" traffic which can automatically force preferential treatment over those forms of "high-precedence" traffic making extremely inefficient use of network data rate. The entries of Table V are listed in order of preference. This illustrates how we can encourage each network user to make an efficient choice of the form of data transmission, regardless of his chosen precedence grade, without depriving any low-precedence user from transmitting a small volume of traffic which adds negligible loading to the network.

Table IV

LONG-TIME NETWORK LOADING AS A FUNCTION OF DIGITAL SERVICES AND NUMERICAL PREFERENCE WEIGHTING FACTOR

		Precedence Grade					
		x1 Flash	x10 Emergency	x20 Opn. Immed.	x100 Priority	x160 Routine	x300 Deferred
1	Keyboard	8	80	160	800	1.28K	2.4K
2	5-Unit 704	60	600	1.2K	6K	9.6K	18K
3	Fieldata	80	800	1.6K	8K	12.8K	24K
4	Partial Core Dump	100	1K	2K	10K	16K	30K
5	300 c/m Card Reader	500	5K	10K	50K	80K	150K
6	Entire Core Dump	1K	10K	20K	100K	160K	300K
7	2.4-Kb Voice	2.4K	24K	48K	240K	384K	720K
8	1200 c/m Card Reader	4K	40K	80K	400K	640K	1.2M
9	FAX	9.6K	96K	192K	960K	1.536M	2.88M
10	19.2-Kb Voice	19.2K	192K	384K	1.92M	2.892M	5.76M
11	Mag. Tape Transfer	60K	600K	1.2M	6M	9.6M	18M

K = 1,000

M = 1,000,000

Table V

RANK OF ORDER OF SERVICE PREFERENCE

		Preference Grade					
		x1 Flash	x10 Emergency	x20 Ops. Immed.	x100 Priority	x160 Routine	x300 Deferred
1	Keyboard	1	4	6	10	14	18
2	5-Unit 704	2	8	13	21	24	30
3	Fieldata	3	9	15	22	28	34
4	Partial Core Dump	5	12	16	27	29	35
5	300 c/m Card Reader	7	20	26	38	41	44
6	Entire Core Dump	11	25	32	43	45	49
7	2.4-Kb Voice	17	33	37	48	51	55
8	1200 c/m Card Reader	19	36	40	52	54	58
9	FAX	23	42	47	56	59	61
10	19.2-Kb Voice	31	46	50	60	62	63
11	Mag. Tape Transfer	39	53	57	64	65	66

VI. PERISHABILITY CONTROL

A communications network can deliver more traffic than a recipient can answer. We wish to prevent oscillation by feeding back processing status as a function of perishability.

First, we shall consider the number of parameters needed to specify the perishability of "messages" of various utilities transmitted in a communications network. Perishability and importance are not synonymous. Inasmuch as we pay for data transmission we assume that we must receive some utility when the message is correctly received by the recipient. The earlier the message arrives, the more useful it will be, as represented in Fig. 8. Such a message might be a telegraph request for a hotel reservation, transmitted while one is rushing out the door to catch an airplane. If the message arrives after the time requested for the reservation, it has a utility of zero. If the message arrives early enough to guarantee a room, then it has a relatively high utility. While we lack a good metric for communications utility, it is sufficient in this discussion to let it be equal to the quantity, $1-P_{pb}$, where P_{pb} is the probability of sleeping on a park bench. (Perhaps the economist's term "disutility" would be more fitting--the price for which one would be willing to sleep on the park bench.) Only two parameters of specification were needed to provide a reasonable measure of the value of the communication as a function of time:

a) Peak value of utility at t = 0 (Point A in Fig. 8);

b) Last time the message had any value (Point B).

A second message example, illustrated in Fig. 9, is "I am bringing a guest home for dinner tonight." Here, three parameters of specification would be useful:

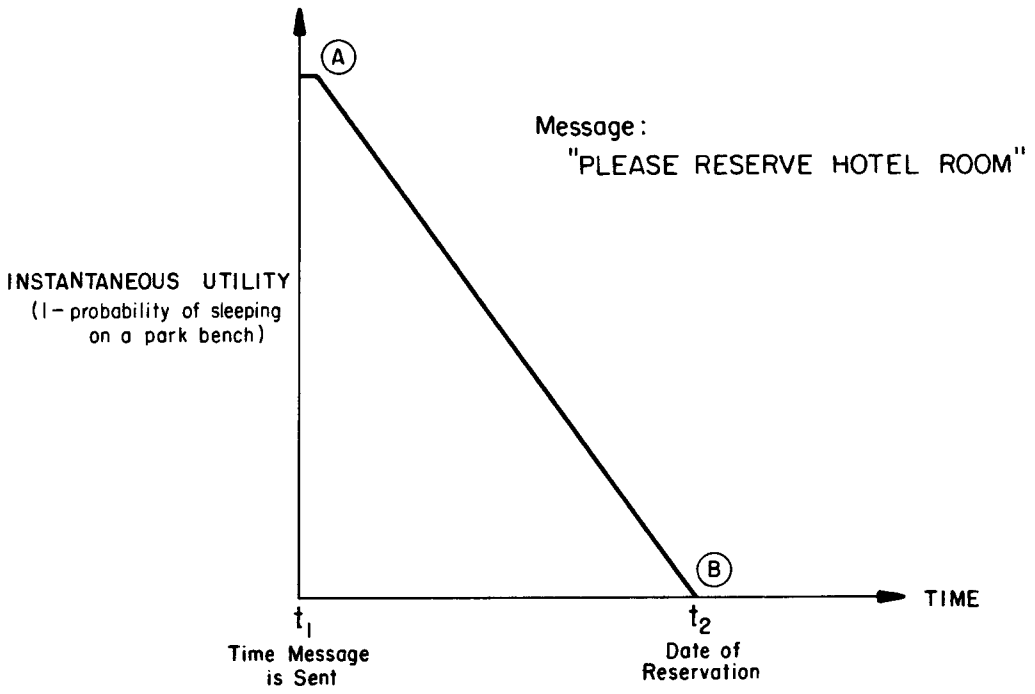

Fig. 8--An Example of Utility of Message as a Function of Elapsed Time

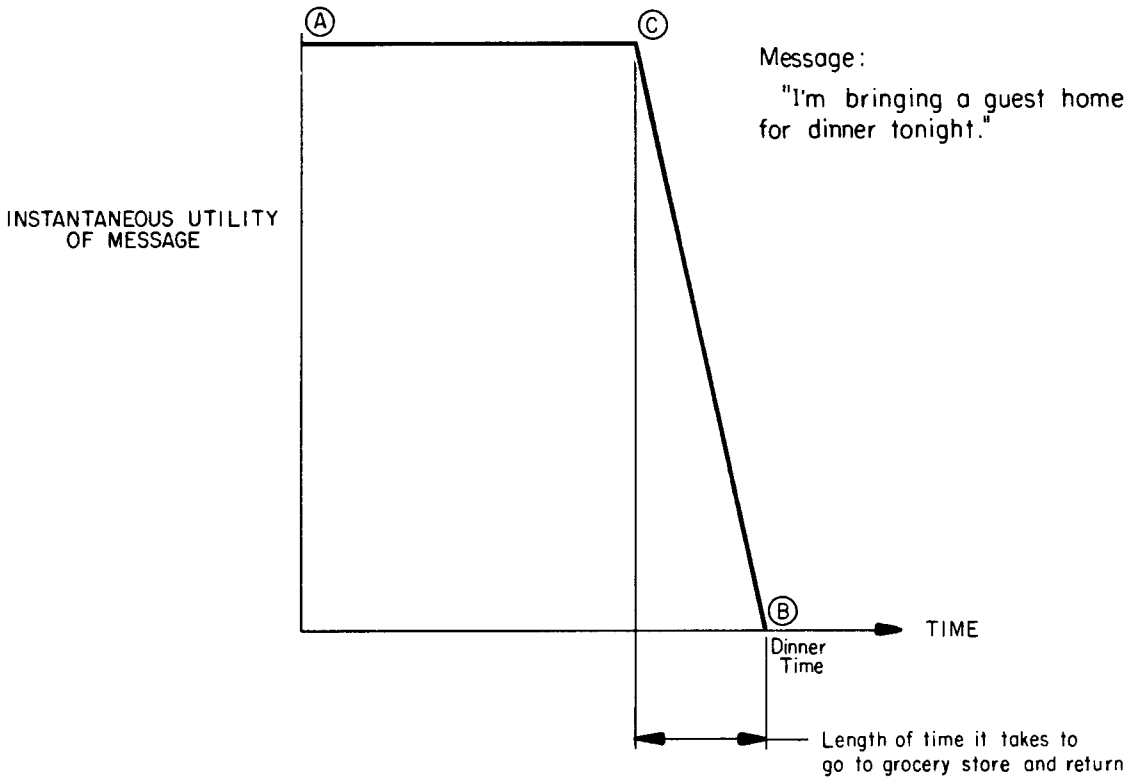

Fig. 9--An Example of Utility of Message
As a Function of Delivery Time

a) Peak value of utility at t = 0 (Point A);
b) Last time the message had any value (Point B);
c) Last time the message had a high value of utility (Point C).

(Parameter C might correspond to the last time one's wife could drive over to the grocery store and pick up more beer for thirsty guests.)

Figure 10 shows the value of instantaneous utility for three different messages: a message to initiate a large long-term project; a message to reserve a hotel room; and a message saying, "Merry Christmas." Only three points are needed to approximate the shape of each of these curves. We may, for example, ask the three questions:

a) How "important" is the message?
b) When would we like the message delivered?
c) When is the latest time the message would be of any value?

While it is possible to think of examples that require more than three specification parameters to delineate the instantaneous utility curve, three will suffice for our purpose. If the communications network knows these few parameters for all traffic in the network, it will be able to perform a rather sophisticated control function. Consider the narrow-funnel-neck problem, where many communications links terminate at a single individual. A communications system may have a capability of delivering more messages which need to be processed than a single end individual can handle. In our future communications system we seek to automatically inform the sender not only that

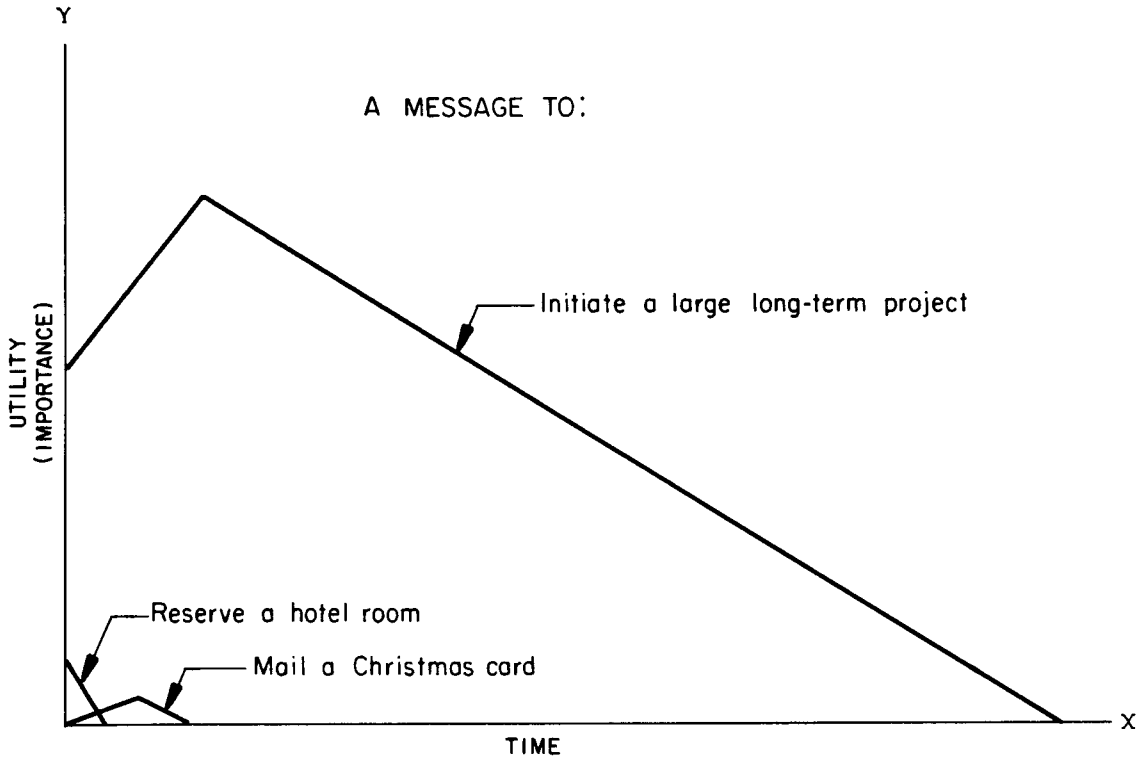

Fig. 10--Instantaneous Utility Curves for Three Different Messages

the end addressee is busy, but that he has a processing backlog of an expected value of K minutes.

If this predicted time is acceptable to the sender, he will do nothing. Otherwise, he could increase his precedence indicator and try again.

In Fig. 11, Messages 101-105 arrive sequentially at an end receiver who requires more time to process messages than the time interval between receipt; i.e., he is overloaded. The question then becomes one of determining the best order for processing these messages. For example, we could use the conventional first-in, first-out processing method. Let us consider the alternative of computing the integral of the peak utility of all messages awaiting attention and ordering the messages in descending sequence of the integrals of peak utility. Thus, the end addressee receiver sequentially picks out what we hope will be the most urgent messages of the set requiring attention. If equipment at the receiver observes the length of time it takes to process each message, it then can heuristically estimate the time when it will be able to accommodate each stacked message. Thus, the sender can be immediately informed of the expected time of action. Even more important, however, is the receipt of an immediate response that the end addressee will not be able to take action within the zone of utility. For example, Fig. 12 shows many messages with different values of utility, all directed to a single human bottleneck. Each message arrives with a zero urgency level, but the level rises in proportion to the urgency, until the item is either processed or the point is reached at which the

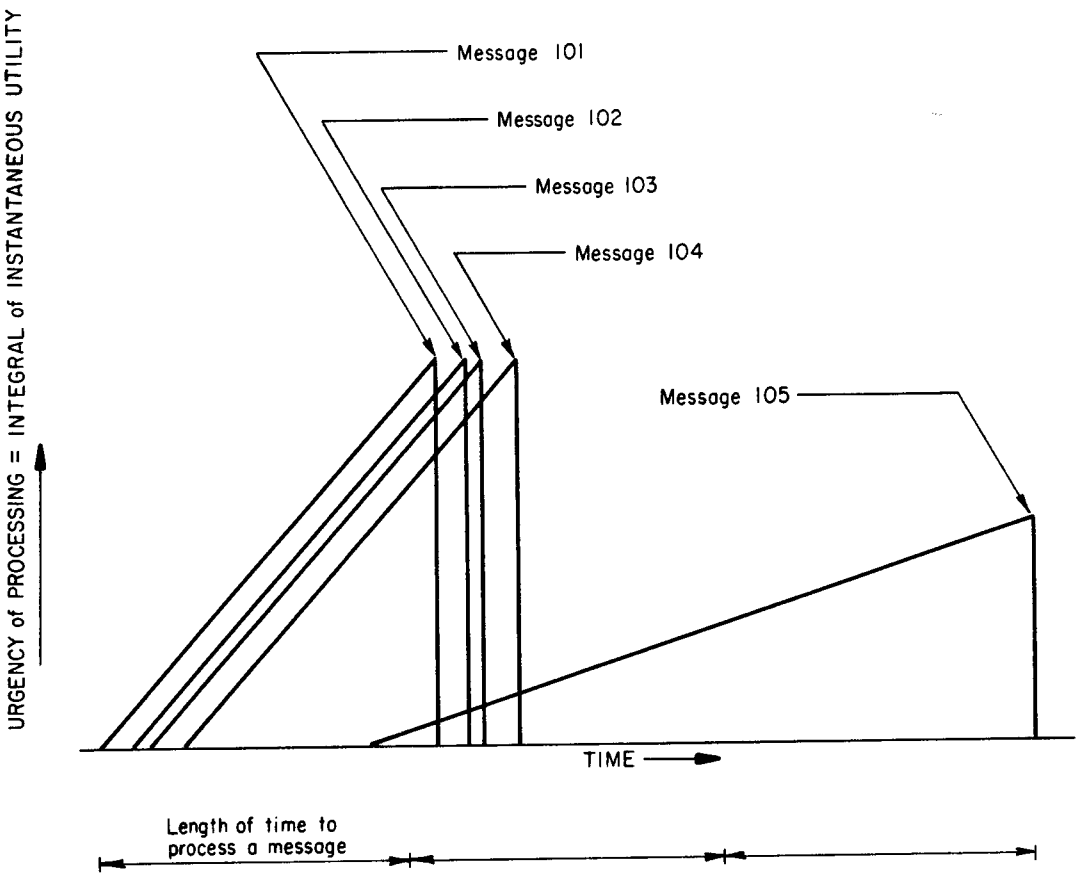

Fig. 11--Several Almost-Simultaneous Messages Requiring Servicing by a Single Processor

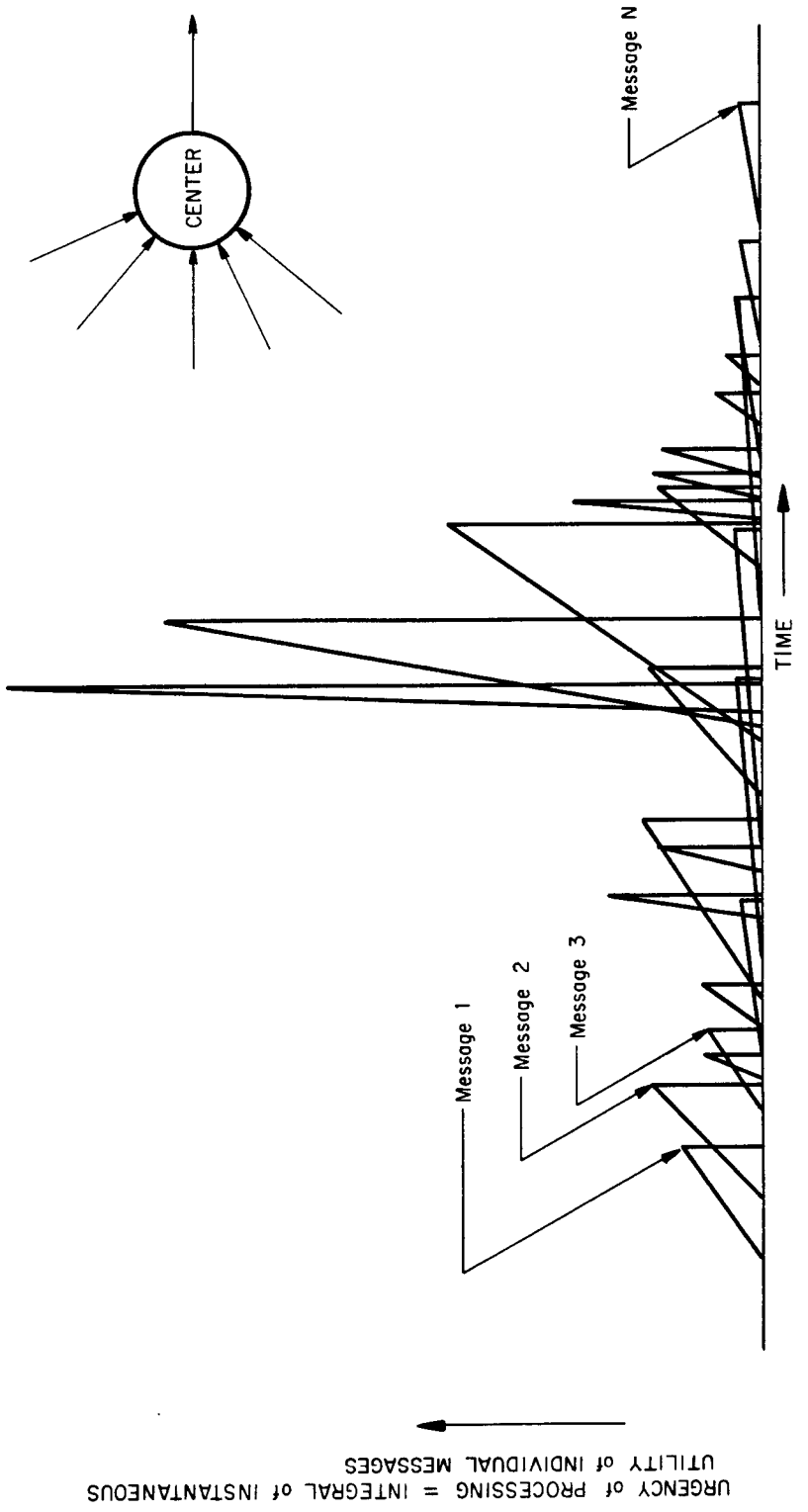

Fig. 12--Many Messages with Different Utilities Passing Through a Single Processor

message is too old to be of any value. Thus, the sender can be immediately informed of whether and when he can communicate to the addressee, allowing him an opportunity to take alternative action if necessary. While only the end addressee ever really knows what is really important, this feature would allow a major pre-filtering of messages to be processed by the receiver so that the receiver will be able to spend most time on the most important of the important tasks demanding his attention.

VII. CONCLUSIONS

A precedence system which is time-dynamic in operation has been proposed. The assignment of capacity and capability is designed to be rapidly changed by the military communications controller. The goal sought is to automate those procedures which help a communications network overload gracefully under attack. The priority order is generally consistent with present military precedence concepts except that means are included to encourage the use of lower-data-rate generating devices in preference to the high-data-rate generating devices in times of overload.

While we have discussed only hard-copy messages, the same mechanisms appear suitable for telephone calls. The feedback information that the called party will not be able to answer within a prescribed period can be announced by an easily recognizable tone pattern. Such signaling tones at the beginning, at the mid-point, and near the end of such rationed calls should also serve to shorten their length. Similar tones could also indicate that the calling party is working against an incoming backlog, and could indicate the urgency of that backlog. All the information needed to implement such a seemingly complex precedence system is available at the Multiplexing Station (see ODC-VIII) in the Distributed Adaptive Message Block Network.

While the techniques discussed best lend themselves to an all-digital distributed network, the general notions apply in any situation where volume-limited communications

is found. We feel that a usable priority structure might one day be evolved that can be implemented on a semi-automatic basis without extreme, continuing demands on executive judgment.

The communications control system suggested seeks to implement the law of management by exception. Each person has a certain job that he must do; as long as he is able to operate within organizational constraints, it is not necessary to request help from a higher hierarchical level. If the demand exceeds capability, only exceptional cases need be transmitted to a higher level for action.

In the proposed control system, changes in network loading are spotted and displayed to predict that a re-construction of allocation of the communication resource may soon be required. If the commander concurs, he can instantly re-allocate his available resource among the many demanders for the resource without complete deprivation to any.

We seek to reduce the inflationary struggle for priority treatment of messages by the generators of communications far down the hierarchical chain. These sources are the sensors which perceive some of the most important information in need of transmission in a crisis. At present, these individuals do not always have the authority to obtain immediate precedence treatment for their messages without a serial series of approvals before a person is reached who has enough authority to place urgent traffic in the network. We think we can do better.

Appendix A

DCS SPEED OF SERVICE CRITERIA*

Precedence	Precedence Prosign	Communication Handling Time	Remarks
FLASH	Z	As fast as possible.	Interrupts lower-precedence.
EMERGENCY	Y	30 min.	Ahead of lower-precedence; interrupts when necessary.
OPERATIONAL IMMEDIATE	O	30 min to 1 hr.	Ahead of lower-precedence; interrupts when appropriate.
PRIORITY	P	1-5 hr.	Does not interrupt lower-precedence in progress.
ROUTINE	R	3-8 hr.	In order received after higher-precedence.
DEFERRED	M	8 hr to start of business following day.	In order received after higher-precedence.

NOTES:

1) Criteria are based on average message length of 125-150 groups.
2) Communication handling times indicate elapsed time from receipt at originating communication center to receipt at addressee center.

*Maloney, E. S., Col., USMC, Defense Communications System Speed of Service Criteria, DCA Circular 70-4, 312, 11 December 1961, p. 2.

Appendix B

MINIMIZE--SELECTED EXCERPTS FROM AIR FORCE REGULATIONS*

2. What the Program Does

The imposition of MINIMIZE conditions is a warning to the users of the facilities and services ... that it is vital to drastically reduce normal message and long distance telephone traffic so communications directly connected with the emergency or exercise will not be delayed.

4. Determining Need for Imposing MINIMIZE

In the event of an emergency or anticipated emergency, the Air Force Communications Message Traffic Control Unit (CMTCU) will furnish telephone reports to the Director of Communications-Electronics, Headquarters USAF, on circuit and traffic conditions throughout the Air Force communications system (AFR 100-49).

5. Authority to Impose MINIMIZE

Except as noted below, only the Chief of Staff, USAF, or, in his absence, his deputy, has the authority to impose MINIMIZE either Air Force-wide or for a specific area or areas. However, the commander of an overseas major air command may institute MINIMIZE if he has obtained prior coordination of the theater commander concerned and approval of the Chief of Staff, USAF.

---

*Air Force Regulation 100-11, Department of the Air Force, Washington, 15 January 1957, pp. 1-4.

6. Procedure for Imposing MINIMIZE

When MINIMIZE is implemented Air Force-wide, it will be imposed by the words "MINIMIZE AIR FORCE-WIDE." When it is applied to a particular area, it will be indicated by designating the area concerned, for example, "MINIMIZE FEAF." Implementation of MINIMIZE by Headquarters USAF will be accomplished by both ALMAJCOM and AFCOMMSTA general messages to insure that all major air commands and Air Force communications centers receive this notification by the most expeditious means. When a commander of an overseas major air command institutes MINIMIZE under the provisions of paragraph 5, it will be accomplished by means of a specifically addressed "BOOK" message to all concerned.

9. Administrative Control of Messages

Strict control must be established to prevent unnecessary use of electrical communications once MINIMIZE is effected. This will be accomplished as follows:

a. Before drafting a message addressed to the affected area or areas, the message drafter will satisfy himself that the immediate situation will be adversely affected if a message is not sent.

b. Each message-releasing authority will perform an adequate check to insure that all message traffic released for electrical transmission falls within the categories of traffic authorized in paragraph 10. Except as noted below, messages not acceptable for release by electrical transmission will be

referred to the drafter for cancellation or consent to dispatch by other means.

c. Message-releasing authorities may release for electrical transmission any message traffic which has been delayed to the extent that early receipt of the content by the addressee(s) is considered vital.

10. Messages Authorized for Electrical Transmission

Message-releasing authorities will approve for electrical transmission only those messages meeting one or more of the criteria listed below. Messages which do not meet these criteria may be dispatched by courier, air mail or ordinary mail in accordance with pertinent Air Force directives:

a. FLASH or EMERGENCY precedence messages (AFM 11-4).

b. REDLINE messages (AFR 100-3).

c. INDICATIONS messages (AFR 100-4).

c. OPERATIONAL IMMEDIATE precedence messages pertaining to:

(1) USAF wartime capability plans.

(2) Aircraft movements.

(3) MANOPED weather messages.

(4) Casualty messages.

3. Messages pertaining to:

(1) The existing emergency or exercise.

(2) Joint War Room Annex (JWRA) operations.

(3) Emergency Air Staff Actions (EASA).

11. Handling of Messages by Communications Centers and Relay Stations

Under MINIMIZE, messages will be handled as follows:

a. Disposition of Original Messages on Hand. Upon receipt of MINIMIZE, Air Force tributary stations will transmit all messages which bear a FLASH, EMERGENCY, or OPERATIONAL IMMEDIATE precedence. All other messages will be returned to message releasing authorities for appropriate action under paragraph 10.

b. Disposition of Relay Tapes on Hand. Upon receipt of MINIMIZE, Air Force tape relay stations will transmit all messages which bear a FLASH, EMERGENCY, or OPERATIONAL IMMEDIATE precedence. All other messages will be disposed of as prescribed by USAF Supplement 1 to ACP 127 (B).

Appendix C

COMMERCIAL TELEPHONE TRAFFIC OVERLOAD PROTECTION TECHNIQUES

The following techniques are either in use by or are being considered by the Bell Telephone Company to minimize system degradation under overload.

1. Reduced Inter-sender Timing

The length of time spent by a sender waiting for dial pulses from the subscriber's telephone is reduced under heavy traffic conditions. For example, upon lifting the telephone from its cradle a dial tone is heard. If you do not start to dial immediately, you needlessly tie up shared common-control equipment. Under overload, the period that the "sender unit" waits before assuming that the subscriber is balking is reduced. This forces the subscriber to dial his number more quickly.

2. Recorded Message

When there is an overload in the Direct Distance Dialing system, calls can be passed to an automatic playback device. A canned recording announces an overload condition, and requests that the caller wait and place his call later. A variation on this recording is used in emergency. Such recorded messages have proved to be highly effective in the past. When people are specifically requested not to use the telephone unless it is an emergency, they will generally refrain from doing so.

3. Operator Spacing

During dial overload, a manual operator can cut in to ask for the number being called. The operator will tell the calling party that she will call back later when the line is free. This effects a delayed spreading of calls during peak periods.

4. Variable Divisions of Circuits

Whenever only two circuits are left in a group of trunks between two cities in the Direct Distance Dialing switching hierarchy, these two lines are reserved for calls approaching from the higher level of the switching hierarchy. This procedure provides precedence for calls that have proceeded a greater distance into the system than incoming calls.

5. Network Management

On certain peak days, such as Christmas, network management policies are enforced which prohibit alternate routing. A traffic supervisory console is used to prevent round-about alternate routes during periods of heavy network use, and to restrict traffic to short efficient routes. The decision to invoke such network management is limited to those in the Regional Offices of the DDD system. (All Regional Control Centers intercommunicate on a single "hot-wire" loop circuit, which permits each center to inform the others of its actions.)

6. Switching Operator Position

During conditions of high overload in which the pattern of demand changes in a predetermined manner (on Mothers' Day, for example), there is a shortage of operator positions in the suburbs and a surplus in the cities. It is possible to remotely switch operator positions and transfer trunks by key switches to process local suburban traffic. An operator position in downtown New York can be wired to handle calls originating in a particular suburb (alleviating overload peaks).

7. Line Load Control

Line load control, as presently practiced, divides all subscribers into three separate groups. First there are the subscribers who are deemed to merit communications in emergencies. Such users include police, fire, mayors, newspapers editors, pay telephone booths, etc. These users are in turn divided into two groups. Each group comprises about 40-45 per cent of all subscribers. During emergencies, a switch at the Central Office can be closed to deny access to new calls to one group of 45 per cent of the users. After ten minutes, service is then denied the second 45-per-cent group. No one's connection is interrupted during line load control; it is simply that no new calls will be accepted from the 45 per cent to whom access is denied.

8. Priority

It is said to be possible to call the operator in an emergency, to identify yourself, tell her it is an

emergency, and expect that she will handle the call in the appropriate manner (this is sometimes called Category "0" Priority within the Bell System).

Appendix D

DIGITAL COMMUNICATIONS MEDIA

TELETYPE

Although a wide variety of teletype codes is in use, we will probably most often encounter conventional five-unit and eight-unit start-stop teletype signals. These signals are convertible to binary-stream transmission with buffering storage or by sampling at a rate much higher than the length of a single teletype bit.

VOICE

Digital voice transmission can be performed at a wide variety of data rates. Seven-sample pulse-code modulation (PCM) using 8000 samples per second gives excellent voice quality, but requires 56,000 bits/sec. Differential PCM can be performed with 38,400 bits/sec. Recently, this writer heard tests of High Information Delta Modulation (HIDM) using a data rate of only 19,200 bits/sec.* The quality was good and the intelligibility excellent. HIDM uses relatively simple analog-digital-analog conversion equipment and appears to meet the requirements for military voice transmission, including good dynamic range. The use of this type of modulation has been assumed in the discussion in the text of this Memorandum.

---

*Winkler, M. R., "High Information Delta Modulation," IEEE International Convention Record, Paper 47.3, New York, New York, March 28, 1963.

Another class of digital voice equipment is vocoder equipment. In the vocoder, voice energy is frequency-division separated by a bank of band-pass filters; the output signal strength of each filter is measured and transmitted as a digital signal. Vocoder equipment is generally expensive and the reconstructed voice is of low quality. Its chief virtue is that it makes extremely efficient use of bandwidth, requiring less than 2400 bits/sec. There are indications that very-high-quality, high-intelligibility vocoders might be built in the future having a data rate of 5-10 kilobits/sec.

(One development that occurred after the preparation of this Memorandum will change the loading figures for voice by a large factor, and should at least be mentioned here. In the description of the Multiplexing Station (ODC-VIII) it is pointed out that it is easy to suppress blank spots in a voice stream without losing quality or breaking synchronization. This would probably reduce the average voice-conversation data rate, when using HIDM for example, from 19,200 bits/sec to about 5000 bits/sec without any drop in quality.)

FACSIMILE

Facsimile transmission, normally an analog signal, can be converted by conventional analog-to-digital means to operate at about 9.6 kilobits/sec.

COMPUTER DATA

A wide variety of manually and semi-manually operated data generating devices are used with computers. These

include typewriter keyboards, coded insertion card interrogators, Hollerith card readers, punched paper tape readers, and high-speed teletypewriters. These devices are generally characterized by low bit-rate requirements.

In the future, we may also find cases where computers might "talk" to one another. Here, the amount of data exchanged can be small, if only processed data is exchanged or if humans are involved. But, there are also applications where it is desirable to exchange raw data between machines. Here, the assumption of equal information per unit time is not applicable, because we are not limited by the human factor in such a loop. One application might be a dump or transfer of the major parts of the high-speed core memory of a computer into a remote computer at a high bit-rate--on the order of one million bits/sec. Magnetic tape can also be read at one station and rewritten into a remote tape unit for later processing, perhaps at rates up to 250,000 bits/sec.

ON DISTRIBUTED COMMUNICATIONS:

List of Publications in the Series

I. Introduction to Distributed Communications Networks, Paul Baran, RM-3420-PR.

Introduces the system concept and outlines the requirements for and design considerations of the distributed digital data communications network. Considers especially the use of redundancy as a means of withstanding heavy enemy attacks. A general understanding of the proposal may be obtained by reading this volume and Vol. XI.

II. Digital Simulation of Hot-Potato Routing in a Broadband Distributed Communications Network, Sharla P. Boehm and Paul Baran, RM-3103-PR.

Describes a computer simulation of the message routing scheme proposed. The basic routing doctrine permitted a network to suffer a large number of breaks, then reconstitute itself by rapidly relearning to make best use of the surviving links.

III. Determination of Path-Lengths in a Distributed Network, J. W. Smith, RM-3578-PR.

Continues model simulation reported in Vol. II. The program was rewritten in a more powerful computer language allowing examination of larger networks. Modification of the routing doctrine by intermittently reducing the input data rate of local traffic reduced to a low level the number of message blocks taking excessively long paths. The level was so low that a deterministic equation was required in lieu of Monte Carlo to examine the now rare event of a long message block path. The results of both the simulation and the equation agreed in the area of overlapping validity.

IV. Priority, Precedence, and Overload, Paul Baran, RM-3638-PR.

The creation of dynamic or flexible priority and precedence structures within a communication system handling a mixture of traffic with different data rate, urgency, and importance levels is discussed. The goal chosen is optimum utilization of the communications resource within a seriously degraded and overloaded network.

V. History, Alternative Approaches, and Comparisons, Paul Baran, RM-3097-PR.

A background paper acknowledging the efforts of people in many fields working toward the development of large communications systems where system reliability and survivability are mandatory. A consideration of terminology is designed to acquaint the reader with the diverse, sometimes conflicting, definitions used. The evolution of the distributed network is traced, and a number of earlier hardware proposals are outlined.

VI. Mini-Cost Microwave, Paul Baran, RM-3762-PR.

The technical feasibility of constructing an extremely low-cost, all-digital, X- or K_u-band microwave relay system, operating at a multi-megabit per second data rate, is examined. The use of newly developed varactor multipliers permits the design of a miniature, all-solid-state microwave repeater powered by a thermoelectric converter burning L-P fuel.

VII. Tentative Engineering Specifications and Preliminary Design for a High-Data-Rate Distributed Network Switching Node, Paul Baran, RM-3763-PR.

High-speed, or "hot-potato," store-and-forward message block relaying forms the heart of the proposed information transmission system. The Switching Nodes are the units in which the complex processing takes place. The node is described in sufficient engineering detail to estimate the components required. Timing calculations, together with a projected implementation

scheme, provide a strong foundation for the belief that the construction and use of the node is practical.

VIII. The Multiplexing Station, Paul Baran, RM-3764-PR.

A description of the Multiplexing Stations which connect subscribers to the Switching Nodes. The presentation is in engineering detail, demonstrating how the network will simultaneously process traffic from up to 1024 separate users sending a mixture of start-stop teletypewriter, digital voice, and other synchronous signals at various rates.

IX. Security, Secrecy, and Tamper-Free Considerations, Paul Baran, RM-3765-PR.

Considers the security aspects of a system of the type proposed, in which secrecy is of paramount importance. Describes the safeguards to be built into the network, and evaluates the premise that the existence of "spies" within the supposedly secure system must be anticipated. Security provisions are based on the belief that protection is best obtained by raising the "price" of espied information to a level which becomes excessive. The treatment of the subject is itself unclassified.

X. Cost Estimate, Paul Baran, RM-3766-PR.

A detailed cost estimate for the entire proposed system, based on an arbitrary network configuration of 400 Switching Nodes, servicing 100,000 simultaneous users via 200 Multiplexing Stations. Assuming a usable life of ten years, all costs, including operating costs, are estimated at about $60,000,000 per year.

XI. Summary Overview, Paul Baran, RM-3767-PR.

Summarizes the system proposal, highlighting the more important features. Considers the particular advantages of the distributed network, and comments on disadvantages. An outline is given of the manner in which future research aimed at an actual implementation of the network might be conducted. Together with the introductory volume, it provides a general description of the entire system concept.