MEMORANDUM
RM-3420-PR
AUGUST 1964

ON DISTRIBUTED COMMUNICATIONS:
I. INTRODUCTION TO DISTRIBUTED COMMUNICATIONS NETWORKS

Paul Baran

PREPARED FOR:
UNITED STATES AIR FORCE PROJECT RAND

The RAND Corporation
SANTA MONICA • CALIFORNIA

MEMORANDUM
RM-3420-PR
AUGUST 1964

ON DISTRIBUTED COMMUNICATIONS:
I. INTRODUCTION TO DISTRIBUTED COMMUNICATIONS NETWORKS

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. Various items in the series deal with the concept in general and with its specific features, 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 first in the series, introduces the system concept and outlines the requirements for and design considerations of a digital data communications system based on the distributed concept, especially as regards implications for such systems in the 1970s. In particular, the Memorandum is directed toward examining the use of redundancy as one means of building communications systems to withstand heavy enemy attacks.

While highly survivable and reliable communications systems are of primary interest to those in the military concerned with automating command and control functions, the basic notions are also of interest to communications systems planners and designers having need to transmit digital data.

Various aspects of the concept as reported in this Memorandum were presented before selected Air Force audiences in the summer of 1961 in the form of a RAND briefing (B-265), and contained in RAND Paper P-2626, which this Memorandum supersedes.

---

*A list of all items in the series is found on p. 35.

SUMMARY

This Memorandum briefly reviews the distributed communications network concept and compares it to the hierarchical or more centralized systems. The payoff in terms of survivability for a distributed configuration in the cases of enemy attacks directed against nodes, links, or combinations of nodes and links is demonstrated.

The requirements for a future all-digital-data distributed network which provides common user service for a wide range of users having different requirements is considered. The use of a standard format message block permits building relatively simple switching mechanisms using an adaptive store-and-forward routing policy to handle all forms of digital data including "real-time" voice. This network rapidly responds to changes in the network status. Recent history of measured network traffic is used to modify path selection. Simulation results are shown to indicate that highly efficient routing can be performed by local control without the necessity for any central--and therefore vulnerable--control point.

A comparison is made between "diversity of assignment" and "perfect switching" in distributed networks. The high degree of connectivity afforded allows the use of low-cost links so unreliable as to be unusable in present type networks.

FOREWORD

The series that this Memorandum introduces describes work on distributed communications. Originally, it was thought that each of the eleven volumes would be able to stand by itself. But, somewhere downstream it became clear that this goal could not be fully met, as each part hinged upon others. Therefore, publication of the individual Memoranda of the series was delayed in order to release the set as a whole.

While the resulting mound of paper forms a frightening pile, it need not all be read in depth, nor will all readers be interested in all the volumes. It is suggested that the present volume be read first especially if the reader is not familiar with its antecedents, B-265 or P-2626. Then the reader should advance directly to the summary overview in Vol. XI. Once in context, it will be easier to selectively examine the other papers of the series in more detail.

Two types of papers will be found. The first set, Vols. I, IV, V, IX, and XI, describes in general terms the underlying system philosophy and what this system approach has to offer. The second set, Vols. II, III, VI, VII, VIII, and X, describes in nuts-and-bolts detail one possible way of implementing the proposed mechanisms. The purpose of this second set is to supply the technical details of the proposed system in sufficient detail, it is hoped, to permit the reader to focus his questions on the potential feasibility of the system in a meaningful manner.

It should be stated at the outset that we are dealing with an extremely complicated system and one that is even more complicated to describe. It would be treacherously easy for the casual reader to dismiss the entire concept as impractically complicated--especially if he is unfamiliar with the ease with which logical transformations can be performed in a time-shared digital apparatus. The temptation to throw up one's hands and decide that it is all "too complicated," or to say, "It will require a mountain of equipment which we all know is unreliable," should be deferred until the fine print has been read.

In the interim, let us agree on what we mean when we speak of "complexity." It can be defined in several ways; for example, by size, by flexibility, or by number of components. But these are not identical measures. Consider an ancient electro-mechanical computer composed of bays of clacking relays. The logical diagrams are simple--a few conceptually simple boxes perform almost trivial logical functions. But the physical dimensions of the package and the amount of maintenance effort required constitute a frightening aspect of complexity.

Conversely, consider a "shoe-box" of electronic equipment that performs all the functions the larger unit did, plus many new ones, and does them more quickly. It's smaller, more reliable, quieter, and requires less maintenance. But it may actually contain more components and its logical equations may be more difficult to comprehend. Is the shoebox more complex or less complex than its room-size electro-mechanical counterpart?

ACKNOWLEDGMENTS

In developing this work, I received a large number of excellent ideas and suggestions--so many, in fact, that it has become impossible to fully acknowledge each person who has contributed in some way without unduly lengthening these manuscripts.

I wish to take this opportunity to thank the following contributors, each of whom reviewed one or more of the Memoranda in the series and who offered highly appreciated and accepted suggestions. The process of review of a manuscript does not necessarily imply full agreement with all that is said, so I alone must accept responsibility for any mistakes in the work.

Reviewers included:* Marvin Adelson (National Academy of Sciences), C. L. Baker, Edward Bedrosian, Sharla Boehm, J. L. Bower, J. B. Carne, L. J. Craig, J. I. Derr, F. E. Eldridge (Office of the Assistant Secretary of Defense, Comptroller), T. O. Ellis, James Farmer, N. E. Feldman, H. Hambrock (North Electric Company), W. B. Holland, J. L. Hult, C. B. Laning (System Development Corporation), C. R. Lindholm, I. S. Reed, E. E. Reinhart, R. H. Scherer (Office of the Director of Defense Research and Engineering), J. W. Smith, Harold Steingold, C. G. Svala (North Electric Company), Rein Turn, K. W. Uncapher, T. G. Williams (Philco Corporation).

*Unless otherwise noted, those listed are with The RAND Corporation.

CONTENTS

PREFACE iii

SUMMARY v

FOREWORD vii

ACKNOWLEDGMENTS ix

Section
I. INTRODUCTION 1

II. EXAMINATION OF A DISTRIBUTED NETWORK 3
Node Destruction 6
Link Destruction 9
Combination Link and Node Destruction .. 9

III. DIVERSITY OF ASSIGNMENT 13
Simulation 13
Comparison with Present Systems 15

IV. ON A FUTURE SYSTEM DEVELOPMENT 16
Future Low-Cost All-Digital Communications Links 17
Variable Data Rate Links 19
Variable Data Rate Users 19
Common User 20
Standard Message Block 20
Switching 23
Forgetting and Imperfect Learning 30
Lowest-Cost Path 33

V. WHERE WE STAND TODAY 34

LIST OF PUBLICATIONS IN THE SERIES 35

I. INTRODUCTION

Let us consider the synthesis of a communication network which will allow several hundred major communications stations to talk with one another after an enemy attack. As a criterion of survivability we elect to use the percentage of stations both surviving the physical attack and remaining in electrical connection with the largest single group of surviving stations. This criterion is chosen as a conservative measure of the ability of the surviving stations to operate together as a coherent entity after the attack. This means that small groups of stations isolated from the single largest group are considered to be ineffective.

Although one can draw a wide variety of networks, they all factor into two components: centralized (or star) and distributed (or grid or mesh) (see Fig. 1).

The centralized network is obviously vulnerable as destruction of a single central node destroys communication between the end stations. In practice, a mixture of star and mesh components is used to form communications networks. For example, type (b) in Fig. 1 shows the hierarchical structure of a set of stars connected in the form of a larger star with an additional link forming a loop. Such a network is sometimes called a "decentralized" network, because complete reliance upon a single point is not always required.

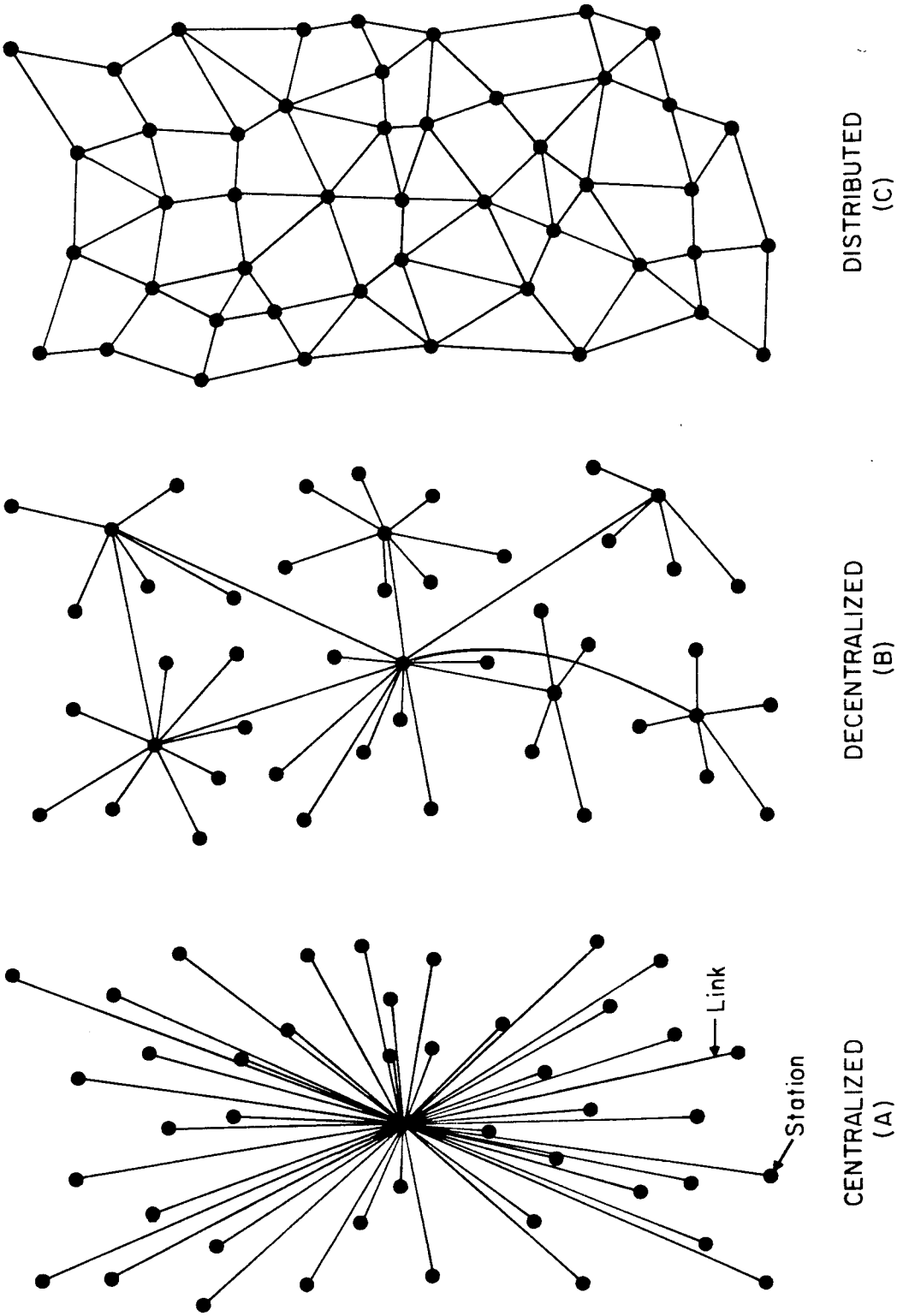

FIG. 1 — Centralized, Decentralized and Distributed Networks

II. EXAMINATION OF A DISTRIBUTED NETWORK

Since destruction of a small number of nodes in a decentralized network can destroy communications, the properties, problems, and hopes of building "distributed" communications networks are of paramount interest.

The term "redundancy level" is used as a measure of connectivity, as defined in Fig. 2. A minimum span network, one formed with the smallest number of links possible, is chosen as a reference point, and is called "a network of redundancy level one." If two times as many links are used in a gridded network than in a minimum span network, the network is said to have a redundancy level of two. Figure 2 defines connectivity of levels 1, 1½, 2, 3, 4, 6, and 8. Redundancy level is equivalent to link-to-node ratio in an infinite size array of stations. Obviously, at levels above three there are alternate methods of constructing the network. However, it was found that there is little difference regardless of which method is used. Such an alternate method is shown for levels three and four, labelled R'. This specific alternate mode is also used for levels six and eight.*

Each node and link in the array of Fig. 2 has the capacity and the switching flexibility to allow transmission between any ith station and any jth station, provided a path can be drawn from the ith to the jth station.

Starting with a network composed of an array of stations connected as in Fig. 3, an assigned percentage of nodes and links is destroyed. If, after this operation, it is still possible to draw a line to connect the ith station to the jth station, the ith and jth stations are said to be connected.

---

*See Craig, L. J., and I. S. Reed, "Overlapping Tessellated Communications Networks," IRE Trans. Comm. Sys., CS-10 (1962) 125-129.

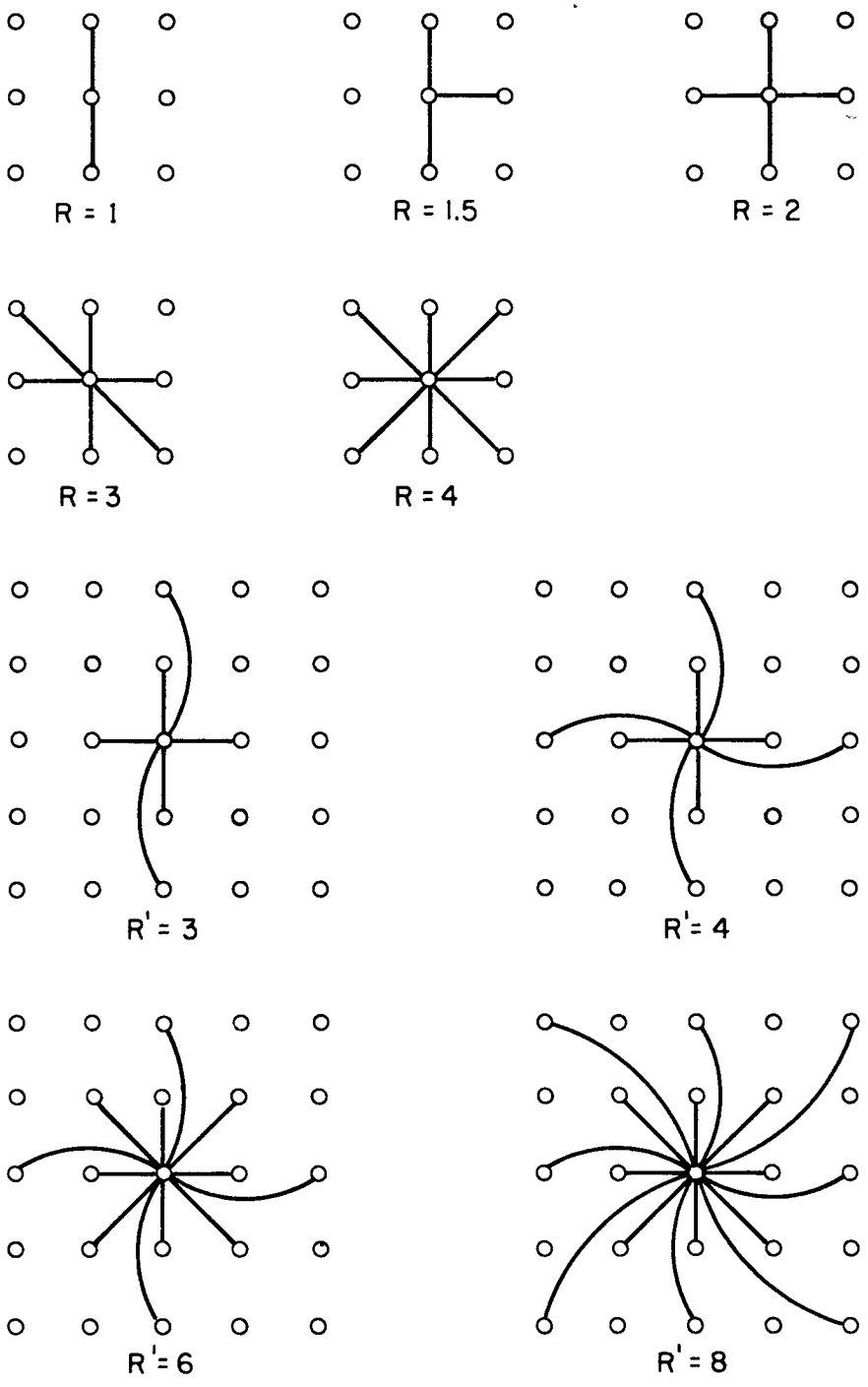

FIG. 2 - Definition of Redundancy Level

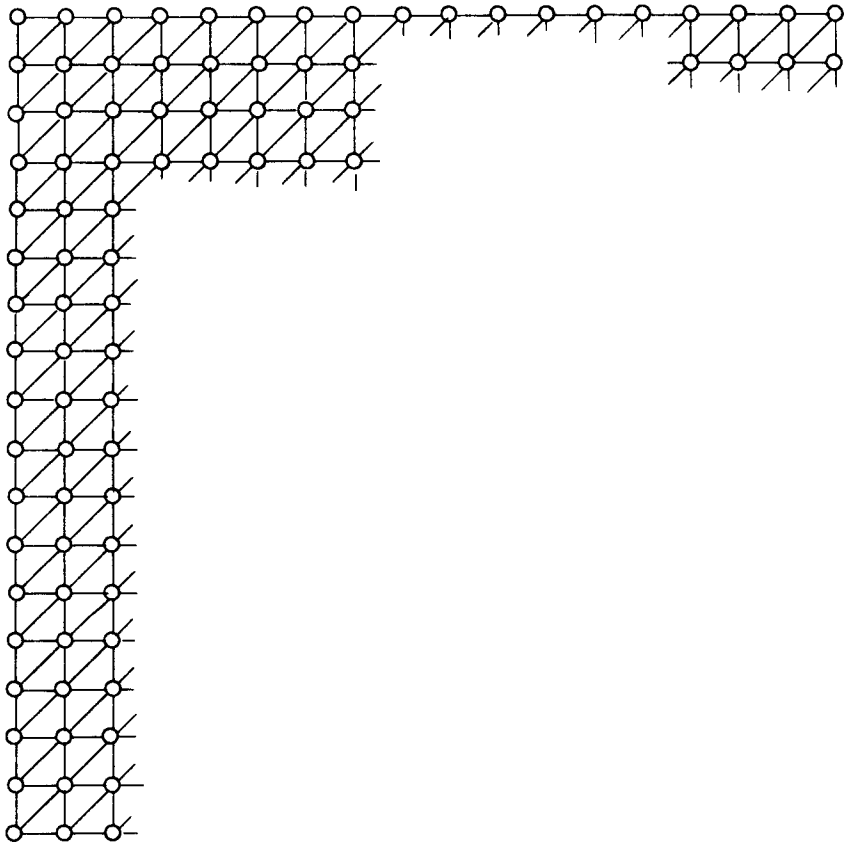

FIG. 3 – An Array of Stations

NODE DESTRUCTION

Figure 4 indicates network performance as a function of the probability of destruction for each separate node. If the expected "noise" was destruction caused by conventional hardware failure, the failures would be randomly distributed through the network. But, if the disturbance were caused by enemy attack, the possible "worst cases" must be considered.

To bisect a 32-link network requires direction of 288 weapons each with a probability of kill, $p_k = 0.5$, or 160 with a $p_k = 0.7$, to produce over an 0.9 probability of successfully bisecting the network. If hidden alternative command is allowed, then the largest single group would still have an expected value of almost 50 per cent of the initial stations surviving intact. If this raid misjudges complete availability of weapons, or complete knowledge of all links in the cross section, or the effects of the weapons against each and every link, the raid fails. The high risk of such raids against highly parallel structures causes examination of alternative attack policies. Consider the following uniform raid example. Assume that 2,000 weapons are deployed against a 1000-station network. The stations are so spaced that destruction of two stations with a single weapon is unlikely. Divide the 2,000 weapons into two equal 1000-weapon salvos. Assume any probability of destruction of a single node from a single weapon less than 1.0; for example, 0.5. Each weapon on the first salvo has a 0.5 probability of destroying its target. But, each weapon of the second salvo has only a 0.25 probability, since one-half the targets have already been destroyed. Thus, the uniform attack is felt to represent a known worst-case configuration in the following analysis.

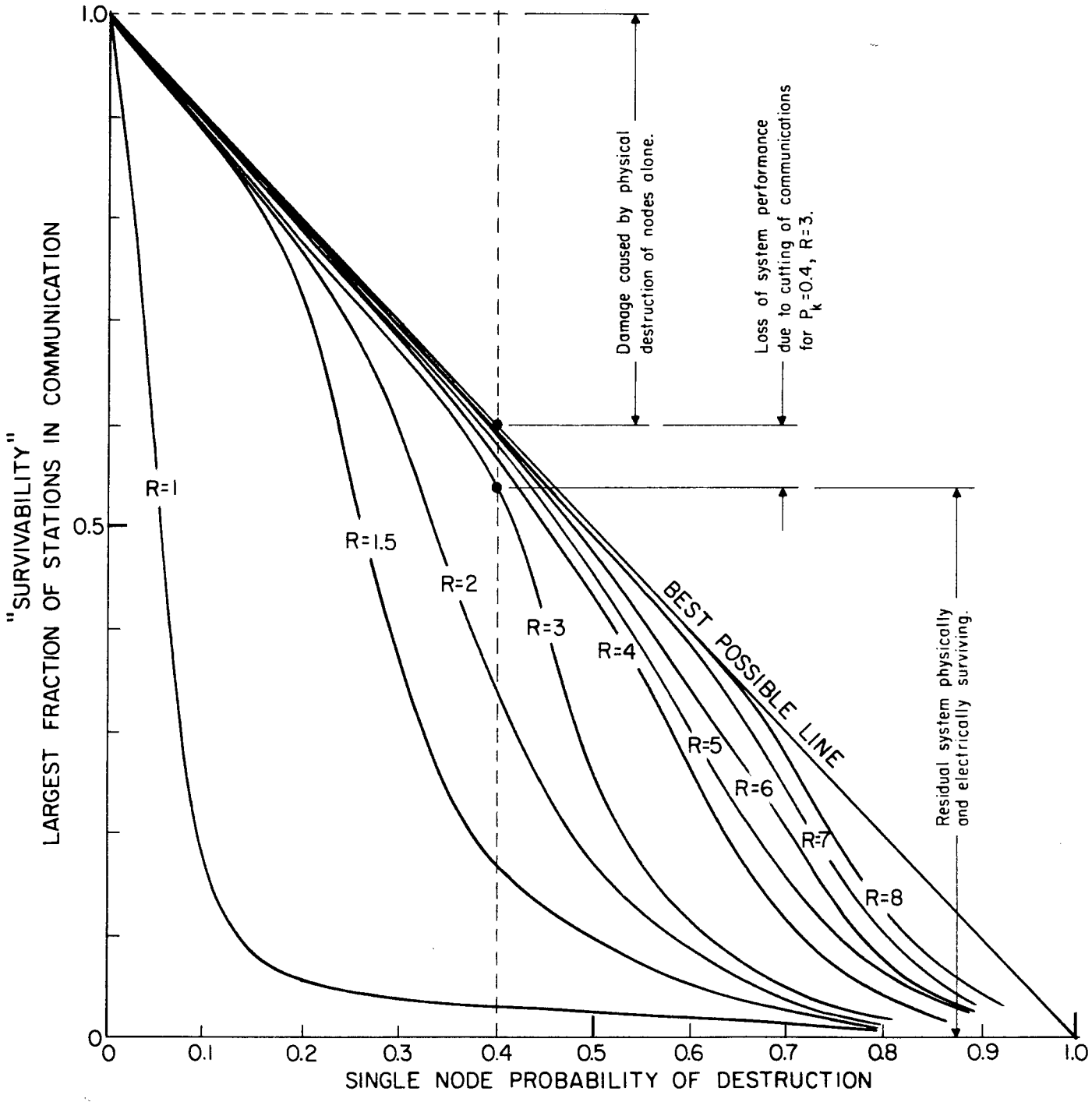

FIG. 4 — Perfect Switching in a Distributed Network — Sensitivity to Node Destruction, 100% of Links Operative.

Such worst-case attacks have been directed against an 18x18-array network model of 324 nodes with varying probability of kill and redundancy level, with results shown in Fig. 4. The probability of kill was varied from zero to unity along the abscissa while the ordinate marks survivability. The criterion of survivability used is the percentage of stations not physically destroyed and remaining in communications with the largest single group of surviving stations. The curves of Fig. 4 demonstrate survivability as a function of attack level for networks of varying degrees of redundancy. The line labeled "best possible line" marks the upper bound of loss due to the physical failure component alone. For example, if a network underwent an attack of 0.5 probability destruction of each of its nodes, then only 50 per cent of its nodes would be expected to survive--regardless of how perfect its communications. We are primarily interested in the additional system degradation caused by failure of communications. Two key points are to be noticed in the curves of Fig. 4. First, extremely survivable networks can be built using a moderately low redundancy of connectivity level. Redundancy levels on the order of only three permit withstanding extremely heavy level attacks with negligible additional loss to communications. Secondly, the survivability curves have sharp break-points. A network of this type will withstand an increasing attack level until a certain point is reached, beyond which the network rapidly deteriorates. Thus, the optimum degree of redundancy can be chosen as a function of the expected level of attack. Further redundancy buys little. The redundancy level required to survive even very heavy attacks is not great--on the order of only three or four times that of the minimum span network.

LINK DESTRUCTION

In the previous example we have examined network performance as a function of the destruction of the nodes (which are better targets than links). We shall now re-examine the same network, but using unreliable links. In particular, we want to know how unreliable the links may be without further degrading the performance of the network.

Figure 5 shows the results for the case of perfect nodes; only the links fail. There is little system degradation caused even using extremely unreliable links--on the order of 50 per cent down-time--assuming all nodes are working.

COMBINATION LINK AND NODE DESTRUCTION

The worst case is the composite effect of failures of both the links and the nodes. Figure 6 shows the effect of link failure upon a network having 40 per cent of its nodes destroyed. It appears that what would today be regarded as an unreliable link can be used in a distributed network almost as effectively as perfectly reliable links. Figure 7 examines the result of 100 trial cases in order to estimate the probability density distribution of system performance for a mixture of node and link failures. This is the distribution of cases for 20 per cent nodal damage and 35 per cent link damage.

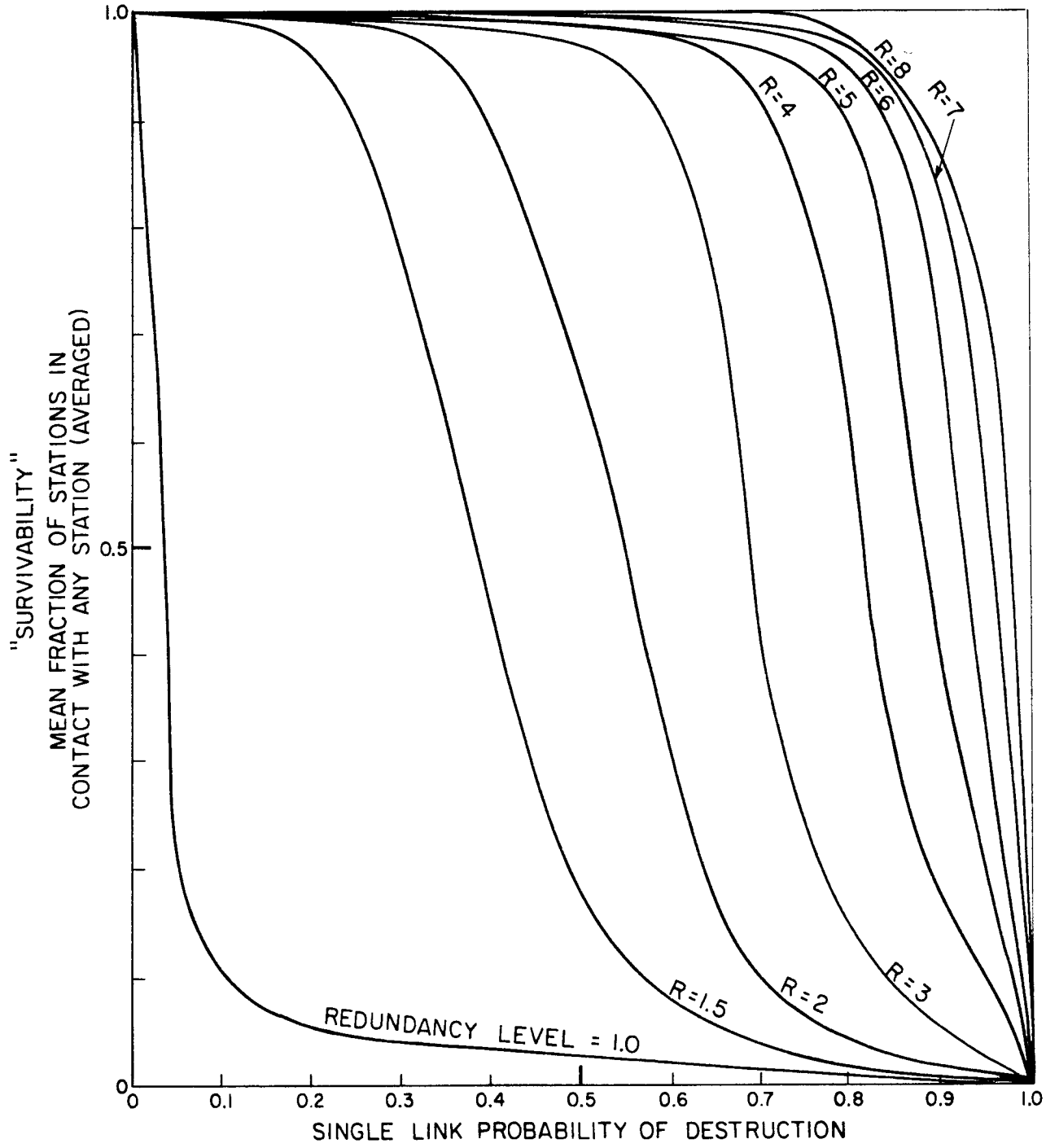

FIG. 5 — Perfect Switching in a Distributed Network — Sensitivity to Link Destruction, 100% of Nodes Operative.

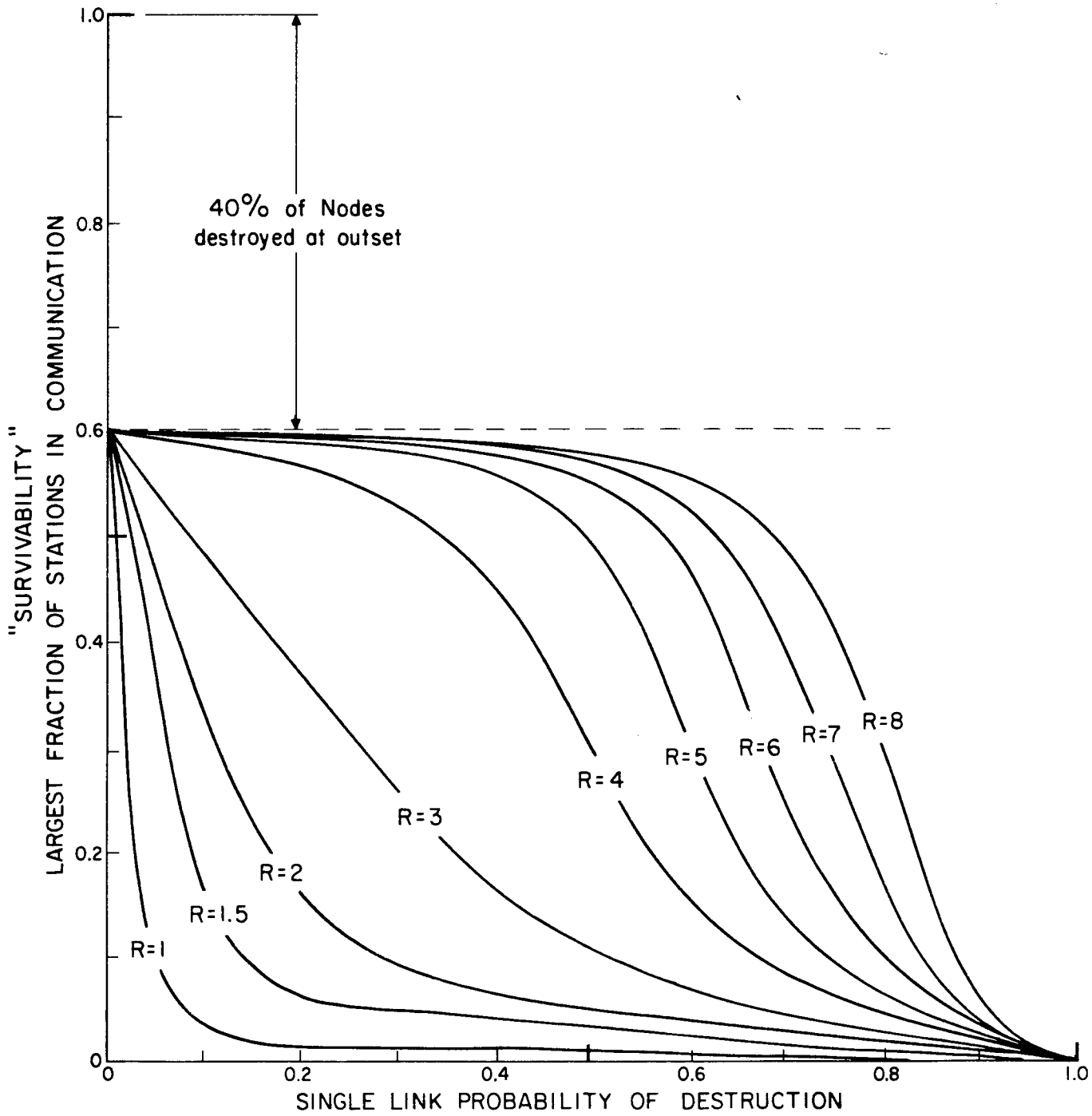

FIG. 6 – Perfect Switching in a Distributed Network — Sensitivity to Link Destruction After 40% Nodes Are Destroyed.

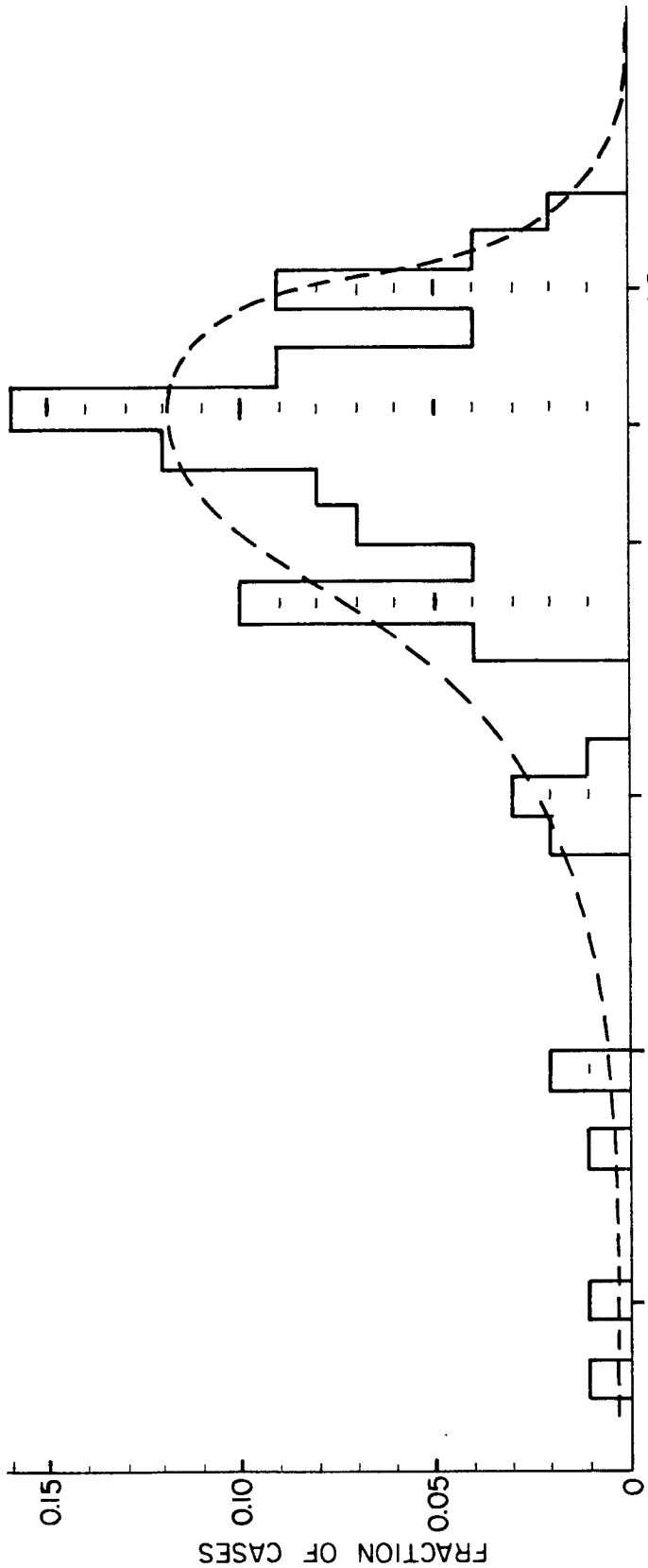

FIG. 7 — Probability Density Distribution of Largest Fraction of Stations in Communication Perfect Switching, R = 3, 100 Cases, 80% Node Survival, 65% Link Survival.

III. DIVERSITY OF ASSIGNMENT

There is another and more common technique for using redundancy than in the method described above in which each station is assumed to have perfect switching ability. This alternative approach is called "diversity of assignment." In diversity of assignment, switching is not required. Instead, a number of independent paths are selected between each pair of stations in a network which requires reliable communications. But, there are marked differences in performance between distributed switching and redundancy of assignment as revealed by the following Monte Carlo simulation.

SIMULATION

In the matrix of N separate stations, each ith station is connected to every jth station by three shortest but totally separate independent paths ($i=1,2,3,\ldots,N$; $j=1,2,3,\ldots,N$; $i \neq j$). A raid is laid against the network. Each of the pre-assigned separate paths from the ith station to the jth station is examined. If one or more of the pre-assigned paths survive, communication is said to exist between the ith and the jth station. The criterion of survivability used is the mean number of stations connected to each station, averaged over all stations.

Figure 8 shows, unlike the distributed perfect switching case, that there is a marked loss in communications capability with even slightly unreliable nodes or links. The difference can be visualized by remembering that fully flexible switching permits the communicator the privilege of ex post facto decision of paths. Figure 8 emphasizes a key difference between some present day networks and the fully flexible distributed network we are discussing.

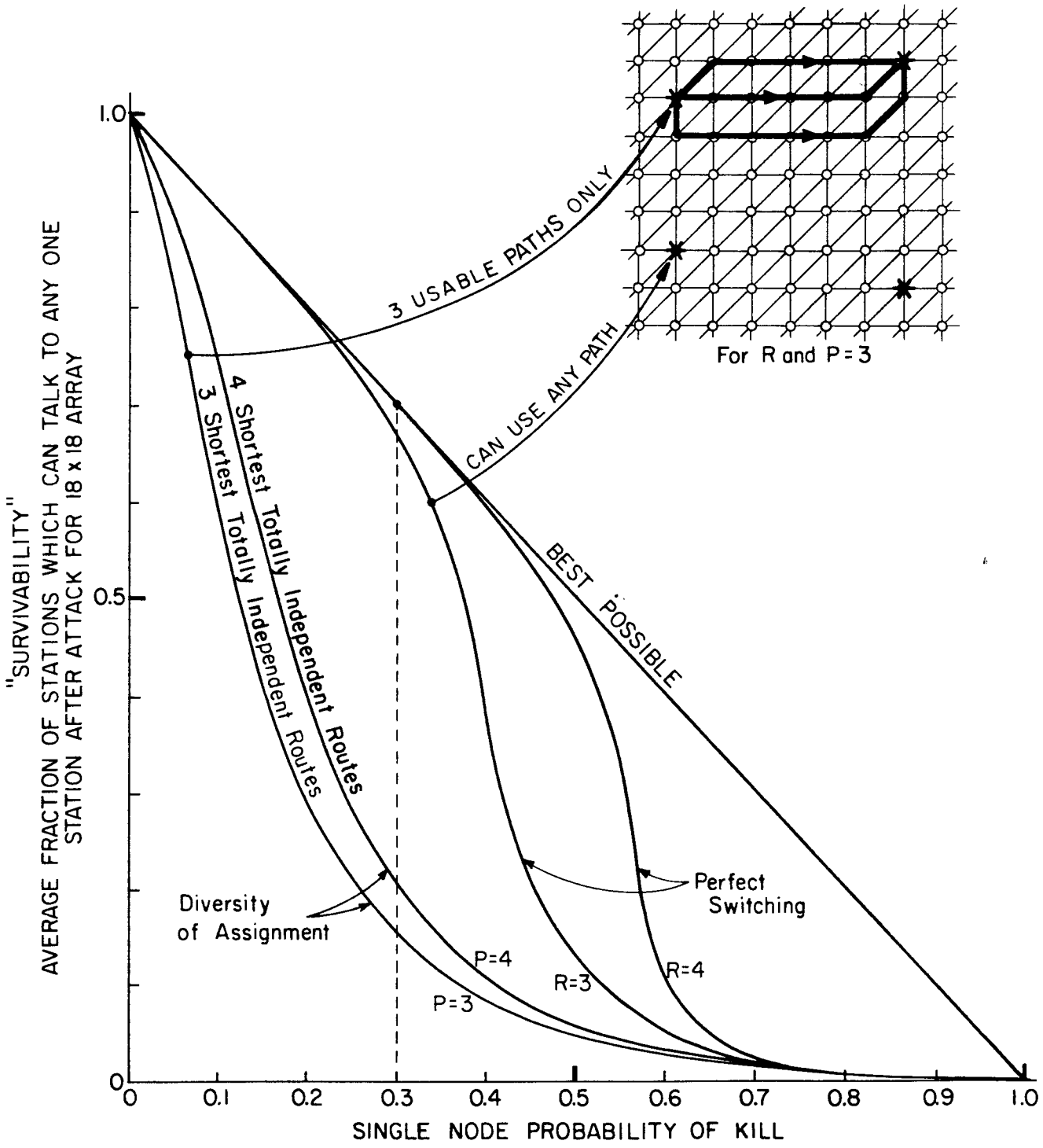

FIG. 8 — Diversity of Assignment vs. Perfect Switching in a Distributed Network.

COMPARISON WITH PRESENT SYSTEMS

Present conventional switching systems try only a small subset of the potential paths that can be drawn on a gridded network. The greater the percentage of potential paths tested, the closer one approaches the performance of perfect switching. Thus, perfect switching provides an upper bound of expected system performance for a gridded network; the diversity of assignment case, a lower bound. Between these two limits lie systems composed of a mixture of switched routes and diversity of assignment.

Diversity of assignment is useful for short paths, eliminating the need for switching, but requires survivability and reliability for each tandem element in long haul circuits passing through many nodes. As every component in at least one out of a small number of possible paths must be simultaneously operative, high reliability margins and full standby equipment are usual.

IV. ON A FUTURE SYSTEM DEVELOPMENT

We will soon be living in an era in which we cannot guarantee survivability of any single point. However, we can still design systems in which system destruction requires the enemy to pay the price of destroying n of n stations. If n is made sufficiently large, it can be shown that highly survivable system structures can be built--even in the thermonuclear era. In order to build such networks and systems we will have to use a large number of elements. We are interested in knowing how inexpensive these elements may be and still permit the system to operate reliably. There is a strong relationship between element cost and element reliability. To design a system that must anticipate a worst-case destruction of both enemy attack and normal system failures, one can combine the failures expected by enemy attack together with the failures caused by normal reliability problems, provided the enemy does not know which elements are inoperative. Our future systems design problem is that of building very reliable systems out of the described set of unreliable elements at lowest cost. In choosing the communications links of the future, digital links appear increasingly attractive by permitting low-cost switching and low-cost links. For example, if "perfect switching"* is used, digital links are mandatory to permit tandem connection of many separately connected links without cumulative errors reaching an irreducible magnitude. Further, the signaling measures to implement

---

*See ODC-V. (ODC is an abbreviation of the series title, On Distributed Communications; the number following refers to the volume in the series. See list on p. 35.)

highly flexible switching doctrines always require digits.

FUTURE LOW-COST ALL-DIGITAL COMMUNICATIONS LINKS

When one designs an entire system optimized for digits and high redundancy, certain new communications link techniques appear more attractive than those common today.

A key attribute of the new media is that it permits formation of new routes cheaply, yet allows transmission on the order of a million or so bits per second, high enough to be economic, but yet low enough to be inexpensively processed with existing digital computer techniques at the relay station nodes. Reliability and raw error rates are secondary. The network must be built with the expectation of heavy damage, anyway. Powerful error removal methods exist.

Some of the communication construction methods that look attractive in the near future include pulse regenerative repeater line, minimum-cost or "mini-cost" microwave, TV broadcast station digital transmission, and satellites.

Pulse Regenerative Repeater Line

Samuel B. Morse's regenerative repeater invention for amplifying weak telegraphic signals has recently been resurrected and transistorized. Morse's electrical relay permits amplification of weak binary telegraphic signals above a fixed threshold. Experiments by various organizations (primarily the Bell Telephone Laboratories) have shown that digital data rates on the order of 1.5 million bits per second can be transmitted over ordinary telephone line at repeater spacings on the order of 6,000 feet for #22 gage pulp paper insulated copper pairs. At present,

more than 20 tandemly connected amplifiers have been used in the Bell System T-1 PCM multiplexing system without retiming synchronization problems. There appears to be no fundamental reason why either lines of lower loss, with corresponding further repeater spacing, or more powerful resynchronization methods cannot be used to extend link distances to in excess of 200 miles. Such distances would be desired for a possible national distributed network.

Power to energize the miniature transistor amplifier is transmitted over the copper circuit itself.

"Mini-Cost" Microwave

While the price of microwave equipment has been declining, there are still untapped major savings. In an analog signal network we require a high degree of reliability and very low distortion for a long string of tandem repeaters. However, using digital modulation together with perfect switching we minimize these two expensive considerations from our planning. We would envision the use of low-power, mass-produced microwave receiver/transmitter units mounted on low-cost, short, guyed towers. Relay station spacing would probably be on the order of 20 miles. Further economies can be obtained by only a minimal use of standby equipment and reduction of fading margins. The ability to use alternate paths permits consideration of frequencies normally troubled by rain attenuation problems reducing the spectrum availability problem.

Preliminary indications suggest that this approach appears to be the cheapest way of building large networks of the type to be described (see ODC-VI).

TV Stations

With proper siting of receiving antennas, broadcast television stations might be used to form additional high data rate links in emergencies.*

Satellites

The problem of building a reliable network using satellites is somewhat similar to that of building a communications network with unreliable links. When a satellite is overhead, the link is operative. When a satellite is not overhead, the link is out of service. Thus, such links are highly compatible with the type of system to be described.

VARIABLE DATA RATE LINKS

In a conventional circuit switched system each of the tandem links requires matched transmission bandwidths. In order to make fullest use of a digital link, the post-error-removal data rate would have to vary, as it is a function of noise level. The problem then is to build a communication network made up of links of variable data rate to use the communication resource most efficiently.

VARIABLE DATA RATE USERS

We can view both the links and the entry point nodes of a multiple-user all-digital communications system as elements operating at an ever changing data rate. From instant to instant the demand for transmission will vary.

---

*Baran, P., Coverage Estimate of FM, TV and Power Facilities Useful in a Broadband Distributed Network (UFOUO), The RAND Corporation, RM-3008-PR, March 1962.

We would like to take advantage of the average demand over all users instead of having to allocate a full peak demand channel to each. Bits can become a common denominator of loading for economic charging of customers. We would like to efficiently handle both those users who make highly intermittent bit demands on the network, and those who make long-term continuous, low bit demands.

COMMON USER

In communications, as in transportation, it is more economical for many users to share a common resource rather than each to build his own system--particularly when supplying intermittent or occasional service. This intermittency of service is highly characteristic of digital communication requirements. Therefore, we would like to consider the interconnection, one day, of many all-digital links to provide a resource optimized for the handling of data for many potential intermittent users--a new common-user system.

Figure 9 demonstrates the basic notion. A wide mixture of different digital transmission links is combined to form a common resource divided among many potential users. But, each of these communications links could possibly have a different data rate. Therefore, we shall next consider how links of different data rates may be interconnected.

STANDARD MESSAGE BLOCK

Present common carrier communications networks, used for digital transmission, use links and concepts originally

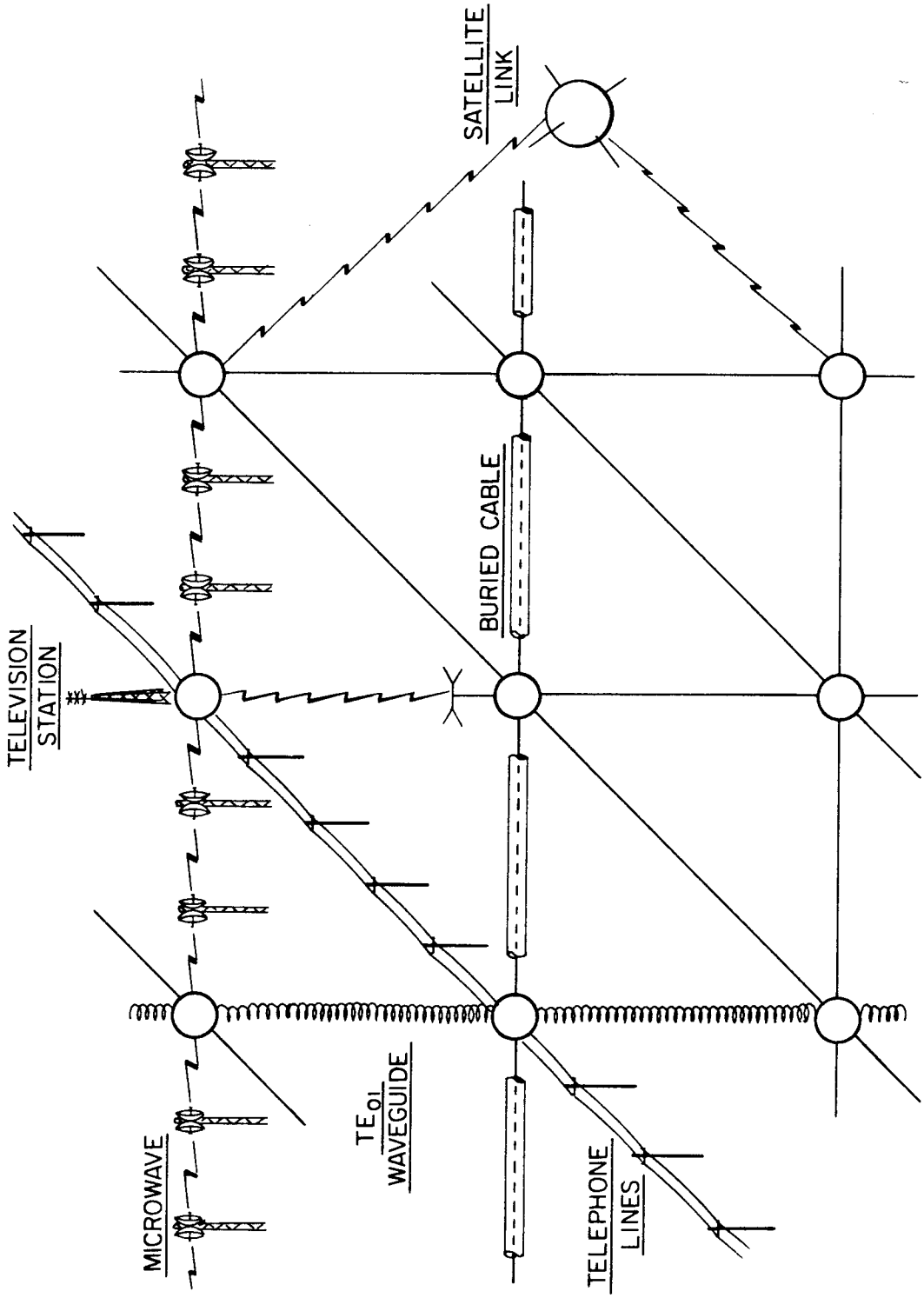

FIG. 9 - All Digital Network Composed of Mixture of Links

designed for another purpose--voice. These systems are built around a frequency division multiplexing link-to-link interface standard. The standard between links is that of data rate. Time division multiplexing appears so natural to data transmission that we might wish to consider an alternative approach--a standardized message block as a network interface standard. While a standardized message block is common in many computer-communications applications, no serious attempt has ever been made to use it as a universal standard. A universally standardized message block would be composed of perhaps 1024 bits. Most of the message block would be reserved for whatever type data is to be transmitted, while the remainder would contain housekeeping information such as error detection and routing data, as in Fig. 10.

As we move to the future, there appears to be an increasing need for a standardized message block for all-digital communications networks. As data rates increase, the velocity of propagation over long links

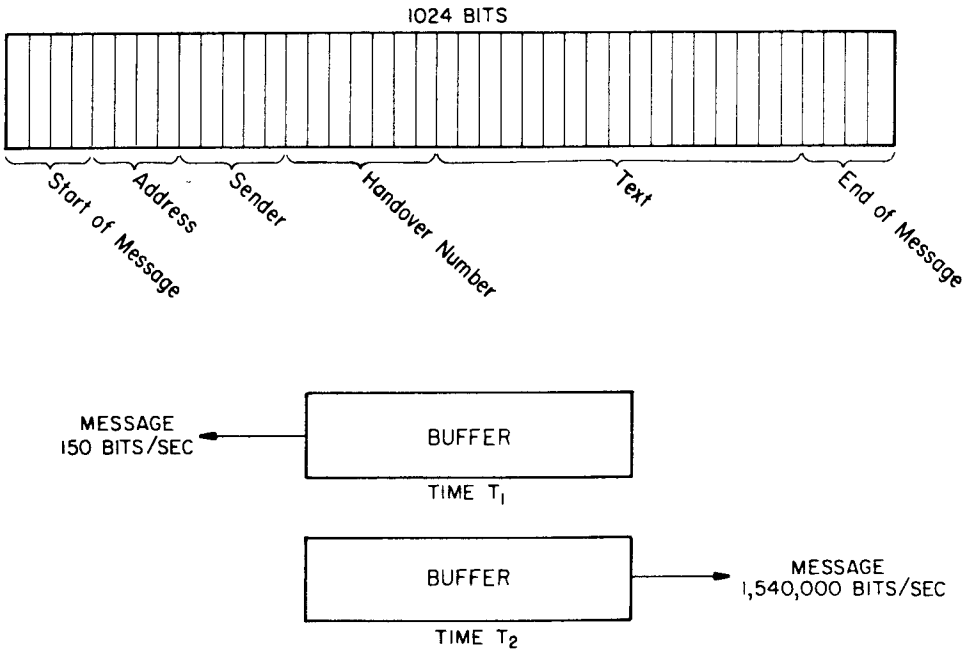

FIG. 10 — Message Block

becomes an increasingly important consideration.* We soon reach a point where more time is spent setting the switches in a conventional circuit switched system for short holding-time messages than is required for actual transmission of the data.

Most importantly, standardized data blocks permit many simultaneous users each with widely different bandwidth requirements to economically share a broadband network made up of varied data rate links. The standardized message block simplifies construction of very high speed switches. Every user connected to the network can feed data at any rate up to a maximum value. The user's traffic is stored until a full data block is received by the first station. This block is rubber stamped with a heading and return address, plus additional housekeeping information. Then, it is transmitted into the network.

SWITCHING

In order to build a network with the survivability properties shown in Fig. 4, we must use a switching scheme able to find any possible path that might exist after heavy damage. The routing doctrine should find the shortest possible path and avoid self-oscillatory or "ring-around-the-rosey" switching.

We shall explore the possibilities of building a "real-time" data transmission system using store-and-forward techniques. The high data rates of the future

---

*3000 miles at $\simeq$ 150,000 miles/sec. $\simeq$ 50 milliseconds transmission time, T.
1024-bit message at 1,500,000 bits/sec. $\simeq$ 2/3 millisecond message time, M.
$\therefore$ T>>M

carry us into a hybrid zone between store-and-forward and circuit switching. The system to be described is clearly store-and-forward if one examines the operations at each node singularly. But, the network user who has called up a "virtual connection" to an end station and has transmitted messages across the United States in a fraction of a second might also view the system as a black box providing an apparent circuit connection across the U.S. There are two requirements that must be met to build such a quasi-real-time system. First, the in-transit storage at each node should be minimized to prevent undesirable time delays. Secondly, the shortest instantaneously available path through the network should be found with the expectation that the status of the network will be rapidly changing. Microwave will be subject to fading interruptions and there will be rapid moment-to-moment variations in input loading. These problems place difficult requirements upon the switching. However, the development of digital computer technology has advanced so rapidly that it now appears possible to satisfy these requirements by a moderate amount of digital equipment. What is envisioned is a network of unmanned digital switches implementing a self-learning policy at each node so that overall traffic is effectively routed in a changing environment--without need for a central and possibly vulnerable control point. One particularly simple routing scheme examined is called the "hot-potato" heuristic routing doctrine and will be described in detail.

Torn-tape telegraph repeater stations and our mail system provide examples of conventional store-and-forward switching systems. In these systems, messages are relayed from station-to-station and stacked until the "best" outgoing link is free. The key feature of store-and-forward transmission is that it allows a high line

occupancy factor by storing so many messages at each node that there is a backlog of traffic awaiting transmission. But, the price for link efficiency is the price paid in storage capacity and time delay. However, it was found that most of the advantages of store-and-forward switching could be obtained with extremely little storage at the nodes.

Thus, in the system to be described, each node will attempt to get rid of its messages by choosing alternate routes if its preferred route is busy or destroyed. Each message is regarded as a "hot potato," and rather than hold the "hot potato," the node tosses the message to its neighbor, who will now try to get rid of the message.

The Postman Analogy

The switching process in any store-and-forward system is analogous to a postman sorting mail. A postman sits at each switching node. Messages arrive simultaneously from all links. The postman records bulletins describing the traffic loading status for each of the outgoing links. With proper status information, the postman is able to determine the best direction to send out any letters. So far, this mechanism is general and applicable to all store-and-forward communication systems.

Assuming symmetrical bi-directional links, the postman can infer the "best" paths to transmit mail to any station merely by looking at the cancellation time or the equivalent handover number tag. If the postman sitting in the center of the United States received letters from San Francisco, he would find that letters from San Francisco arriving from channels to the west would come in with later cancellation dates than if such letters had arrived in a roundabout manner from the east. Each letter carries an implicit indication of its length

of transmission path. The astute postman can then deduce that the best channel to send a message to San Francisco is probably the link associated with the latest cancellation dates of messages from San Francisco. By observing the cancellation dates for all letters in transit, information is derived to route future traffic. The return address and cancellation date of recent letters is sufficient to determine the best direction in which to send subsequent letters.

Hot-Potato Heuristic Routing Doctrine

To achieve real-time operation it is desirable to respond to change in network status as quickly as possible, so we shall seek to derive the network status information directly from each message block.

Each standardized message block contains a "to" address, a "from" address, a handover number tag, and error detecting bits together with other housekeeping data. The message block is analogous to a letter. The "from" address is equivalent to the return address of the letter.

The handover number is a tag in each message block set to zero upon initial transmission of the message block into the network. Every time the message block is passed on, the handover number is incremented. The handover number tag on each message block indicates the length of time in the network or path length. This tag is somewhat analogous to the cancellation date of a conventional letter.

The Handover Number Table. While cancellation dates could conceivably be used on digital messages, it is more convenient to think in terms of a simpler digital analogy--a tag affixed to each message and incremented every time

	LINK NUMBER							
	1	2	3	4	5	6	7	8
	HANDOVER NUMBER ENTRIES							
A	22	∞	12	10	9	9	8	13
B	5	3	2	2	4	5	12	2
C	7	8	13	9	22	10	7	8
D	21	23	19	21	12	10	12	13
E	7	10	12	14	12	13	13	15
F	7	10	12	13	14	21		
G	6	4	10					
Z	15	20	7	3	10	8	5	10

BEST CHOICE				
1st	2nd	3rd	4th	5th
LINK NUMBER for DECISION CHOICE				
7	5	6	4	3
3	4	8	2	5
1	7	2	8	4
6	5	7	8	3
1	2	3	5	6
1	2	3	4	8
5	2	1	6	
4	7	3	6	5

Each Switching Node contains a table used to record handover numbers of traffic en route. The entries on the table represent the lowest recently measured handover numbers from stations A, B, C, etc., on each of the eight links feeding the node. For example, station E's traffic had a handover number of 7 on link number 1. The table to left orders the preference of the routes to the stations shown. Thus, if traffic were addressed to station E, link number 1, the shortest measured route, will be the first choice. If link 1 is busy or destroyed, the next highest handover number is found on link 2, which then becomes the preferred choice.

FIG. 11 — The Handover Number Table

the message is relayed. Figure 11 shows the handover table located in the memory of a single node. A row is reserved for each major station of the network allowed to generate traffic. A column is assigned to each separate link connected to a node. As it was shown that redundancy levels on the order of four can create extremely "tough" networks and additional redundancy brought little, only about eight columns are really needed.

Perfect Learning. If the network used perfectly reliable, error-free links, we might fill out our table in the following manner. Initially, set entries on the table to high values. Examine the handover number of each message arriving on each line for each station. If the observed handover number is less than the value already entered on the handover number table, change the value to that of the observed handover number. If the handover number of the message is greater than the value on the table, do nothing. After a short time this procedure will shake down the table to indicate the path length to each of the stations over each of the links connected to neighboring stations. This table can now be used to route new traffic. For example, if one wished to send traffic to station C, he would examine the entries for the row listed for station C based on traffic from C. Select the link corresponding to the column with the lowest handover number. This is the shortest path to C. If this preferred link is busy, do not wait, choose the next best link that is free.

Digital Simulation

This basic routing procedure was tested by a Monte

Carlo simulation of a 7x7 array of stations. All tables were started completely blank to simulate a worst-case starting condition where no station knew the location of any other station. Within ½ second of simulated real world time, the network had learned the locations of all connected stations and was routing traffic in an efficient manner. The mean measured path length compared very favorably to the absolute shortest possible path length under various traffic loading conditions. Preliminary results indicate that network loadings on the order of 50 per cent of link capacity could be inserted without undue increase of path length. When local busy spots occur in the network, locally generated traffic is intermittently restrained from entering the busy points while the potential traffic jams clear. Thus, to the node, the network appears to be a variable data rate system, which will limit the number of local subscribers that can be handled. If the network is carrying light traffic, any new input line into the network would accept full traffic, perhaps 1.5 million bits per second. But, if every station had heavy traffic and the network became heavily loaded, the total allowable input data rate from any single station in the network might drop to perhaps 0.5 million bits per second. The absolute minimum guaranteed data capacity into the network from any station is a function of the location of the station in the network, redundancy level, and the mean path length of transmitted traffic in the network. The "choking" of input procedure has been simulated in the network and no signs of instability under

overload noted. It was found that most of the advantage of store-and-forward transmission can be provided in a system having relatively little memory capacity. The network "guarantees" very rapid delivery of all traffic that it has accepted from a user (see ODC-II, -III).

FORGETTING AND IMPERFECT LEARNING

We have briefly considered network behavior when all links are working. But, we are also interested in determining network behavior with real world links--some destroyed, while others are being repaired. The network can be made rapidly responsive to the effects of destruction, repair, and transmission fades by a slight modification of the rules for computing the values on the handover number table.

Learning

In the previous example, the lowest handover number ever encountered for a given origination, or "from" station, and over each link, was the value recorded in the handover number table. But, if some links had failed, our table would not have responded to the change. Thus, we must be more responsive to recent measurements than old ones. This effect can be included in our calculation by the following policy. Take the most recently measured value of handover number; subtract the previous value found in the handover table; if the difference is positive, add a fractional part of this difference to the table value to form the updated table value. This procedure merely implements a "forgetting" procedure--placing more belief upon more recent measurements and less on old measurements. This device would, in the case of network damage, automatically modify the handover

number table entry so as to exponentially and asymptotically approach the true shortest path value. If the difference between measured value minus the table value is negative, the new table value would change by only a fractional portion of the recently measured difference.

This implements a form of sceptical learning. Learning will take place even with occasional errors. Thus, by the simple device of using only two separate "learning constants," depending whether the measured value is greater or less than the table value, we can provide a mechanism that permits the network routing to be responsive to varying loads, breaks, and repairs. This learning and forgetting technique has been simulated for a few limited cases and was found to work well (see ODC-II, -III).

Adaptation to Environment

This simple simultaneous learning and forgetting mechanism implemented independently at each node causes the entire network to suggest the appearance of an adaptive system responding to gross changes of environment in several respects, without human intervention. For example, consider self-adaptation to station location. A station, Able, normally transmitted from one location in the network, as shown in Fig. 12(a). If Able moved to the location shown in Fig. 12(b), all he need do to announce his new location is to transmit a few seconds of dummy traffic. The network will quickly learn the new location and direct traffic toward Able at his new location. The links could also be cut and altered, yet the network would relearn. Each node sees its environment through myopic eyes by only having links and link status information to a few neighbors. There is no central control;

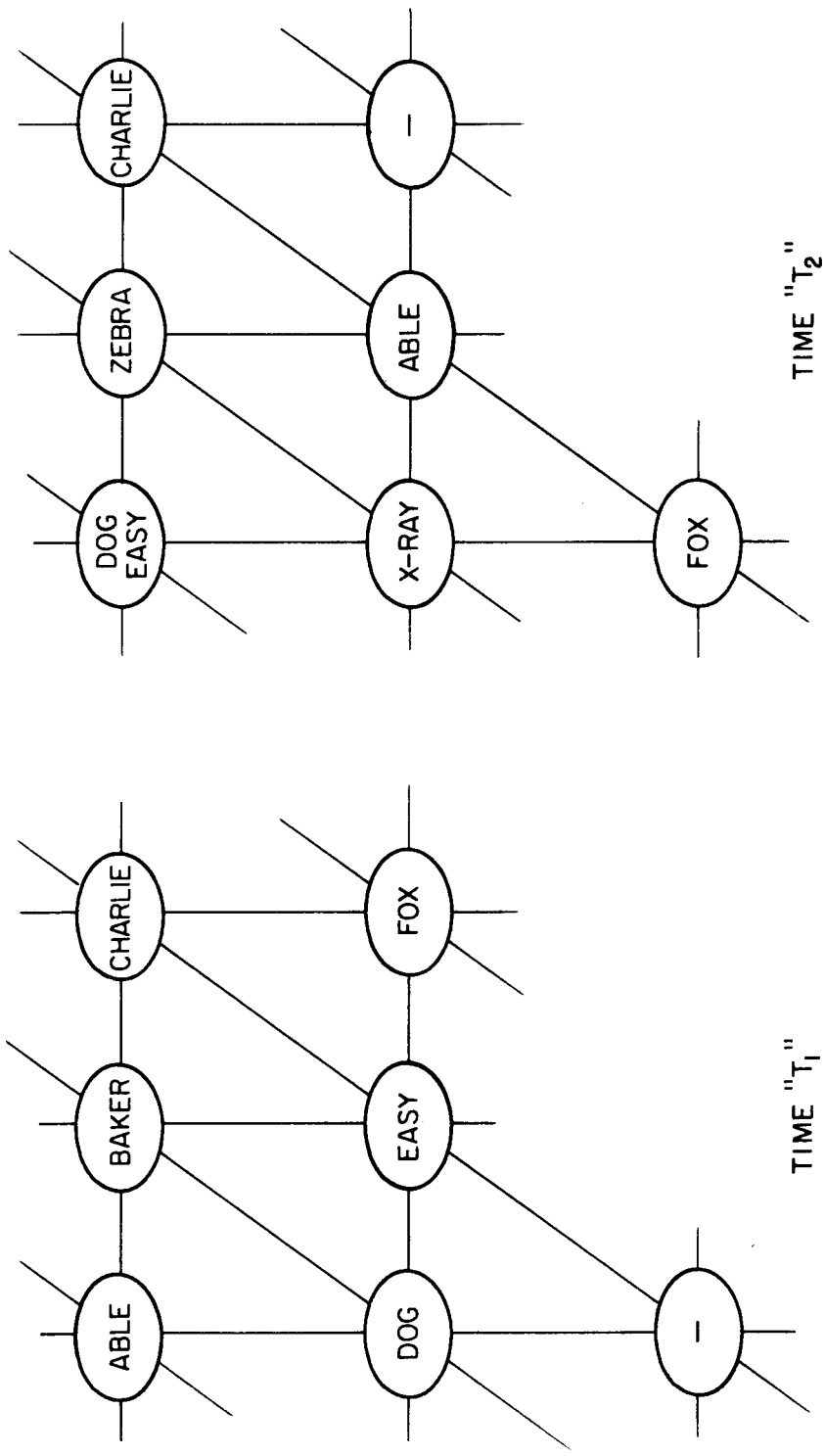

FIG. 12 – Adaptability to Change of User Location

only a simple local routing policy is performed at each node, yet the overall system adapts.

LOWEST-COST PATH

We seek to provide the lowest-cost path for the data to be transmitted between users. When we consider complex networks, perhaps spanning continents, we encounter the problem of building networks with links of widely different data rates. How can paths be taken to encourage most use of the least expensive links? The fundamentally simple adaptation technique can again be used. Instead of incrementing the handover by a fixed amount, each time a message is relayed, set the increment to correspond to the link-cost/bit of the transmission link. Thus, instead of the "instantaneously shortest non-busy path" criterion, the path taken will be that offering the cheapest transportation cost from user to user that is available. The technique can be further extended by placing priority and cost bounds in the message block itself, permitting certain users more of the communication resource during periods of heavy network use.

V. WHERE WE STAND TODAY

Although it is premature at this time to know all the problems involved in such a network and understand all costs, there are reasons to suspect that we may not wish to build future digital communication networks exactly the same way the nation has built its analog telephone plant.

There is an increasingly repeated statement made that one day we will require more capacity for data transmission than needed for analog voice transmission. If this statement is correct, then it would appear prudent to broaden our planning consideration to include new concepts for future data network directions. Otherwise, we may stumble into being boxed in with the uncomfortable restraints of communications links and switches originally designed for high quality analog transmission. New digital computer techniques using redundancy make cheap unreliable links potentially usable. A new switched network compatible with these links appears appropriate to meet the upcoming demand for digital service. This network is best designed for data transmission and for survivability at the outset.

It is the purpose of the other volumes in this series to consider this new direction in more detail. The reader may wish to review ODC-XI as a more recent overview before reading the intervening papers.

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 thermo-electric 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.