NBSIR 80-2149

GRIDNET

R. T. Moore

Center for Computer Systems Engineering
Institute for Computer Sciences and Technology
National Bureau of Standards
U.S. Department of Commerce
Washington, DC 20234

October 1980

Sponsored by
Defense Nuclear Agency
Washington, DC 20305

NBSIR 80-2149

GRIDNET

R. T. Moore

Center for Computer Systems Engineering
Institute for Computer Sciences and Technology
National Bureau of Standards
U.S. Department of Commerce
Washington, DC 20234

October 1980

Sponsored by
Defense Nuclear Agency
Washington, DC 20305

U.S. DEPARTMENT OF COMMERCE, Philip M. Klutznick, *Secretary*

Luther H. Hodges, Jr., *Deputy Secretary*
Jordan J. Baruch, *Assistant Secretary for Productivity, Technology, and Innovation*

NATIONAL BUREAU OF STANDARDS, Ernest Ambler, *Director*

TABLE OF CONTENTS

FIGURES

FLOWCHARTS

GRIDNET

R. T. Moore

ABSTRACT

In work reported in NBSIR 79-1725, Phase II Final Report Computerized Site Security Monitor and Response System, by Moore, et al., a novel communications system called CROSSFIRE was introduced. In this system, a primary station controls communications with a number of associated secondary stations using a link level communications control protocol defined by ANSI X3.66-1979 entitled American National Standard for advanced data communication control procedures (ADCCP), which formed the basis for FIPS PUB 71. Communication takes place over two loops, each carrying the same data. One loop transfers the data in a clockwise direction and the other carries the same data in a counter clockwise direction. As a result of the redundant transmission paths, the delivery of a message to its destination cannot be prevented by cutting both loops at any single geographic location. Further, by the use of a continuous polling mode of operation, the loops are never idle for significant intervals of time. This facilitates the prompt recognition of any attempted injection of a spurious or "spoofing" message.

These CROSSFIRE advantages can be extended to large networks by interconnecting a number of CROSSFIRE loops together using "gateway" type stations at their intersections. Multiple interconnections of the loops and adaptive routing techniques further enhances the possibility of retaining an alternate route between a given pair of stations even in the face of simultaneous interruptions to the continuity of multiple loops. This conceptual approach has been termed GRIDNET.

GRIDNET is intended to provide a communications capability between a relatively large number of stations, perhaps as many as several thousand. It's high survivability is achieved by providing many alternative routes. At the same time, it conserves the available bandwidth by using simple procedures to determine traffic routing.

Key words: Communications networks; CROSSFIRE; link protocols; loop; survivability.

INTRODUCTION

This report describes a highly reliable and survivable digital data communication system. It is based on the multiple interconnection of dual fiber optic loops, with up to about twenty data communications stations being served by each dual loop. The dual loop configuration is called CROSSFIRE. The interconnection of many of these to form a large network is called GRIDNET. The network is highly connected and many alternate routes are available for transmitting a message between two stations located at different points on the network. The intelligence required for the control of communications and the routing of traffic is distributed among a number of data communication nodes called gateway stations. Network survival is not dependent on the survival of any node or link, but requires instead that the network not be fragmented.

The report reviews the operation of CROSSFIRE dual loops and defines a set of desirable characteristics for a network based on their interconnection. Alternative interconnection structures are considered. One is selected that permits growth and still provides an orderly addressing scheme that supports the use of routing algorithms that require the gateway stations to have only limited, localized knowledge of network operability status. Procedures are defined that allow control of communications processes to be rotated between the gateway stations on each loop with graceful degradation of performance occurring as gateways fail so long as at least one gateway retains operational status. Routing algorithms are described and are flowcharted, and the link level communications control protocols and message headings required to support the operation of the network are defined.

CROSSFIRE OPERATION

Because the GRIDNET is the interconnection of a number of CROSSFIRE loops, an understanding of the operational characteristics of the CROSSFIRE is a prerequisite to the understanding of the GRIDNET.

The CROSSFIRE configuration is a double loop in which traffic flows in opposite directions in two independent parallel paths. At any time, the communications are under the control of a primary station. This station sequentially polls each of the secondary stations on the loop, inviting them to send any traffic that they may have originated. The primary also delivers messages originated by others to the secondary to whom they are addressed during the exchange of communications that are involved in a poll.

Optical fibers are used as the communications media for the loops because of their immunity to electromagnetic interference and their relatively high resistance to degradation in performance as a result of most environmental factors, including radiation and moisture. The fibers will support signalling rates of tens to hundreds of megabits per second, however, in the Computerized Site Security Monitor and Response System (CSSMRS) a much lower signalling rate was selected.

On one of the paths of the double loop the data flow is in a clockwise direction, while on the other path, the same data moves in a counterclockwise direction. The data are identical in both timing and content upon leaving the sending station. The sending station, of course, may be either the primary or any one of the secondary stations on the dual loop. Normally, the data arrives at its destination via one path at a slightly different time than the data from the other path because the two paths will have different delay times.

When it is not originating traffic, each secondary station on the double loop functions as a repeater for traffic that is originated elsewhere on the loop. Optical signals arriving over the clockwise fiber are detected and converted to electrical signals that are amplified and reconverted to optical signals with minimum delay. These, in turn, are immediately impressed upon the outgoing clockwise fiber to the

next adjacent clockwise station. In a similar fashion, signals arriving on the incoming counterclockwise fiber are detected, amplified and output to the next adjacent counterclockwise station. The detected bit stream as received over both of the loops is continuously monitored by each secondary station. Figure 1 is a block diagram illustrating the arrangements at a secondary station. Recognition of its own address causes that secondary to take appropriate actions.

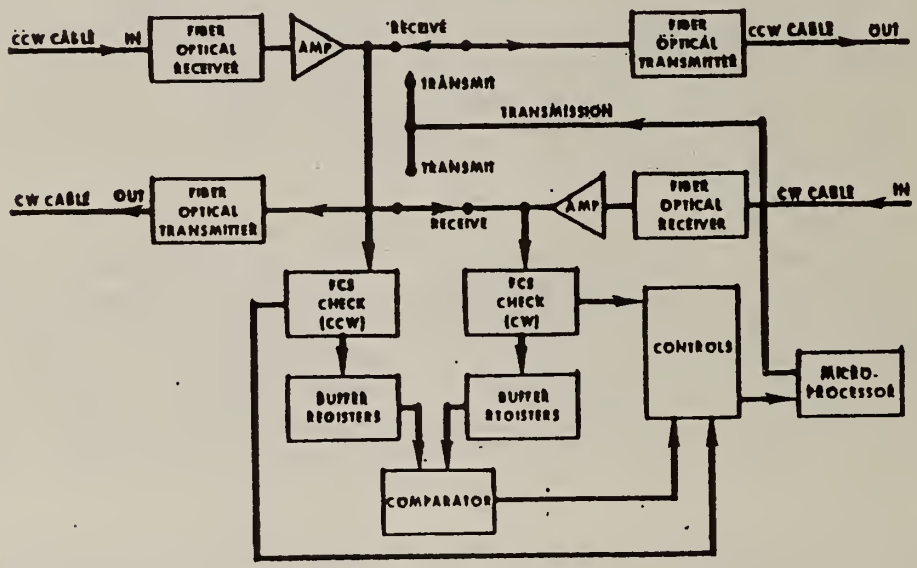


Figure 1. CROSSFIRE control components at a secondary station

The primary station does not retransmit the messages that it receives on either loop, but it does check to see that any traffic that it sends out comes back to it on each of the two incoming loops after a time long enough to accomodate the delay required for propogation and regeneration. Failure of a bit sequence to propagate completely around one or both of the loops is an immediate indication of a problem. The problem can generally be localized when a polling sequence

involving each of the secondary stations has been completed.

The communications control protocol that is used on the CROSSFIRE dual loop configuration is based on to Advanced Digital Data Communication Procedure (ADCCP), American National Standard X3.66 - 1979. Transmissions are in packets of bits called frames. Each frame has a minimum length of 48 bits. It begins with a flag which is a unique bit pattern that serves to indicate both the beginning and the end of a frame. Next is an octet of bits containing local link (in this case loop) address information. Then there is a control octet which contains commands issued by a primary or responses generated by a secondary station. This may be followed by an optional information field which may contain some or all of the classical message components such as preamble, address, text and signature. Following the control octet, or the information field if one is present, there are two octets called the frame check sequence, or FCS. This is an error control feature that has an extremely high probability of detecting any transmission error that might occur. Finally, each frame is terminated with another flag octet.

As indicated above, each secondary station continuously monitors the traffic flowing on the clockwise and counterclockwise loops. When it recognizes its own address on a frame, it stores that frame in buffers and calculates the FCS. It then waits for a period of time long enough to accomodate the difference in propagation and repeater delays on the two loops in order to receive the identical frame that is expected to arrive over the other loop. This frame is also stored in buffers and the FCS is calculated. The frame contents, including information field if present, are also compared, bit by bit, to determine whether they agree.

If a frame is received over one loop, and it has a correct FCS, while on the other loop the frame is either missing or has an incorrect FCS, the copy having the correct FCS is accepted as valid and action is executed accordingly. In its response to the primary station, the secondary station also indicates on which loop the frame was not received, or was received with an incorrect FCS. This information permits the primary to determine the location of any fault that may occur and to initiate an appropriate message to the network maintenance organization.

If both copies of a frame are received with correct FCS's but their contents do not agree, this is a possible indication of sophisticated spoofing. The contents of both frames must be ignored and the event is reported in an appropriate response frame to the primary station.

If both copies of a frame are received with incorrect FCS's,

this is an indication of possible malfunction, noise, or even jamming. The frame contents are ignored and a report of the event is made to the primary station.

If both copies of a frame have correct FCS's and the frame contents are identical, operation is normal. The frame contents are accepted for action and it is not necessary for the secondary to generate an exception report for transmission to the primary.

The operation of the CROSSFIRE configuration is controlled and managed by the primary station. This station polls each of the secondary stations in sequence and on a continuous basis. Thus, there are only very brief gaps in the flow of traffic that occur during the reaction time between the receipt of a command by a secondary and its subsequent response. This makes it very difficult for an adversary to spoof the system by injecting spurious frames as these would have a high probability of colliding with legitimate frames. The primary analyzes the responses of the secondaries when abnormal operation is reported as a result of failure of the FCS or when frame contents do not agree. It determines the location of any adversary attack or equipment malfunction based on the response analysis, and generates appropriate messages to invoke maintenance or security actions. Because of its functions in controlling traffic flow and in network management, the operation of a CROSSFIRE data communications configuration is dependent on the functional survivability of a primary station. In the CSSMRS application, the influence of this dependence on the primary was alleviated by endowing the microcomputers associated with each secondary station the capability of autonomous action in the event that communications with the primary station was lost. This provided a secure, survivable and reliable system for use within a special weapons storage site.

The number of secondary stations that can be accomodated on a CROSSFIRE dual loop depends upon a number of factors. These include circuit lengths, signalling rates and required response times. In the CSSMRS application, there are a maximum of 24 secondary stations on each dual loop that is associated with a primary station. At most, the total loop lengths are only a few kilometers and propagation delay is not a significant consideration. The signalling speed selected was 56 kilobits per second and this, together with the number of secondary stations, was the principal determinant of response time. This signalling rate allows the use of low cost Light Emitting Diode (LED) electro-optical components that are still fast enough to permit regenerative repeating at each secondary station without the need for a one-bit delay for complete reconstitutuion of the signal.

This minimum delay repeating, together with the limited loop dimensions permits a poll of each of 24 stations on a loop to be completed in less than 0.1 second. This was considered to be an acceptable value for the intended application.
To evaluate the effects of other engineering tradeoffs, consider the CROSSFIRE dual loop. Average propagation delay time is equal to approximately 75 percent of the total one-way loop propagation delay. This is because the frame must arrive at its destination via both loops. The minimum time required for this will occur when the destination of the frame is located midway around each loop from the source. Here the delay would be half of the total loop propagation time. The longest delay will occur when the destination of the frame is adjacent to the source and one copy of the frame must travel almost completely around the loop. The average will be approximately midway between these extremes, or about 75 percent of the maximum.

Light will travel in an optical fiber at about two-thirds of its velocity in free space, or about 5.0 microseconds per kilometer. Thus, the propagation delay in a 100 kilometer path would be about 500 microseconds. Assuming that there are 10 secondary stations on a CROSSFIRE loop of this size operating at a signalling rate of one megabit per second, and further assuming a one-bit delay for regenerative repeating at each station, the average propagation delay would be about $10 + 0.75\ (500) = 385$ microseconds.

Using typical bit-oriented communication control procedures, a minimum length poll message and its response are each 48 bits long, assuming that neither contain an information field. Allowing 100 microseconds for checking the FCS and for comparing messages received on the clockwise and counter-clockwise loops, the poll and subsequent response from each secondary would each require a little less than 0.6 millisecond on the average. Thus, both poll and response could be completed in about 1.2 millisecond, and all 10 stations could be polled in about 12.0 milliseconds. Therefore if one of the secondary stations should suddenly need to transmit a message, on the average, it would not have to wait more than half of this time or six milliseconds before being polled and given the opportunity to transmit, assuming there were no other traffic queues.

At the other extreme, if each of the secondary stations on the loop had a 10,000 bit message to transmit, the delay before the last station could begin it's transmission would be over one tenth of a second. Under these circumstances, and assuming that the transmission is error-free, 10,000 bits could be transferred every 11,200 microseconds for a bandwidth utilization of about 89%.

The addition of more secondary stations to the loop described above would cause proportional increases in poll cycle time and message delay time. The impact of the message delay time could be reduced by employing a higher signalling rate, but this would have little effect on poll cycle time, as propagation delay is already the major factor in that aspect of performance.

Perhaps a more significant factor to consider in evaluating the optimum number of stations that might be connected to a CROSSFIRE loop is reliability. The failure of any single regenerative repeater will disable one half of the loop. Thus, the Mean Time Between Failures (MTBF) of each repeater divided by the number of repeaters on each half of the loop provides an estimate of the MTBF for the nodes on that half of the loop. It would probably not be reasonable to expect to have more than about 30 stations on a loop, and smaller numbers would provide faster response times and higher reliability. Offsetting this is the fact that circuit costs, prorated on a per station basis, are reduced as the number of stations on a loop is increased.

Longer loop lengths will result in a nearly proportional increase in poll cycle time when the loop is idle, but will have relatively little impact on message delay times or efficiency under conditions of maximum traffic. More significance is attached to the condition that the distance between each station should not be so great that intermediate repeaters would be required between stations to compensate for attenuation in the fiber optic system. Recent experiments with low loss fibers have indicated that a repeater spacing of about 20 kilometers might be used with 100 Mbit/s transmission at a wavelength of 1.5 micrometers. This would permit a 20 station loop with up to 20 kilometer spacing between stations to have a total length of as much as 400 kilometers.

Given these general characteristics of CROSSFIRE loops, it is now appropriate to consider how multiple crossfire loops could be interconnected so as to provide multiple routes between the stations on one loop and those on another loop. Such an interconnection of CROSSFIRE dual loops is called GRIDNET.

DESCRIPTION OF GRIDNET

Even without a detailed knowledge of the operational requirements that a GRIDNET would be expected to satisfy, it is possible to identify a number of desirable characteristics. Survivability is at the top of this list. We define survivability to represent a capability to permit critical message traffic to flow (perhaps at reduced rates) in spite of the simultaneous disablement of multiple links and nodes within the network and the introduction of jamming and spoofing signals by an adversary. Features that support this definition of survivability include the following:

1. The availability of multiple alternative routes between stations on the network.

2. A routing mechnaism that attempts to select short routes for messages whenever possible, but that automatically reverts to more indirect routing when this is required to permit message delivery.

3. A protocol that supports error control, message accountability, and the prompt detection of malfunction or adversary attack.

4. Distributed intelligence so that network operation does not depend on the survival of a few critical switching or control centers.

5. Resistance to man-made or naturally occurring electromagnetic interference and to all other expected environmental factors.

In addition, the GRIDNET structure should allow for future growth and expansion.

GRIDNET Constraints

The following constraints were adopted for GRIDNET:

1. An addressing scheme that helps guide the routing algorithms.

2. Routing algorithms that require primary stations to have only limited knowledge about operability conditions at remote portions of the network. (For instance, a gateway station that relays messages between adjacent interconnected loops might only be required to know the status of the other stations on its home loop and whether it and the other gateways can communicate successfully with their adjacent interconnected foreign loops).

3. A protocol that supports efficient operation but still permits loop control to be transferred between gateways in order to permit graceful degradation of performance. (Operation of a loop should not depend on the continued operation of any single station).

4. Some method of extinquishing messages that might otherwise circulate indefinitely in a fragmented or badly fractured network.

GRIDNET Structure

A key factor in the development of an addressing scheme that would be supportive of the routing algorithms is an ordered format for interconnecting loops in a regular manner. That is, given a loop somewhere in the interior of the connected network, to how many other loops should it be connected? Arrangements may be considered in which one interior loop is connected to two, three or to four other loops.

In Fig. 2, each interior loop is shown connected to two other loops. (The figure shows it as a single rather than a dual loop for clarity in illustrating loop connectivity.)

The result is a simple chain like structure. Connectivity is clearly limited, and the net can be fragmented by cutting any single interior loop in only two places. Such a structure would be inadequate to the stated GRIDNET objectives.

Figure 2. Two-connectivity loop structure

If each horizontal loop is connected to three others, a regular structure can be laid out as shown in Fig. 3.

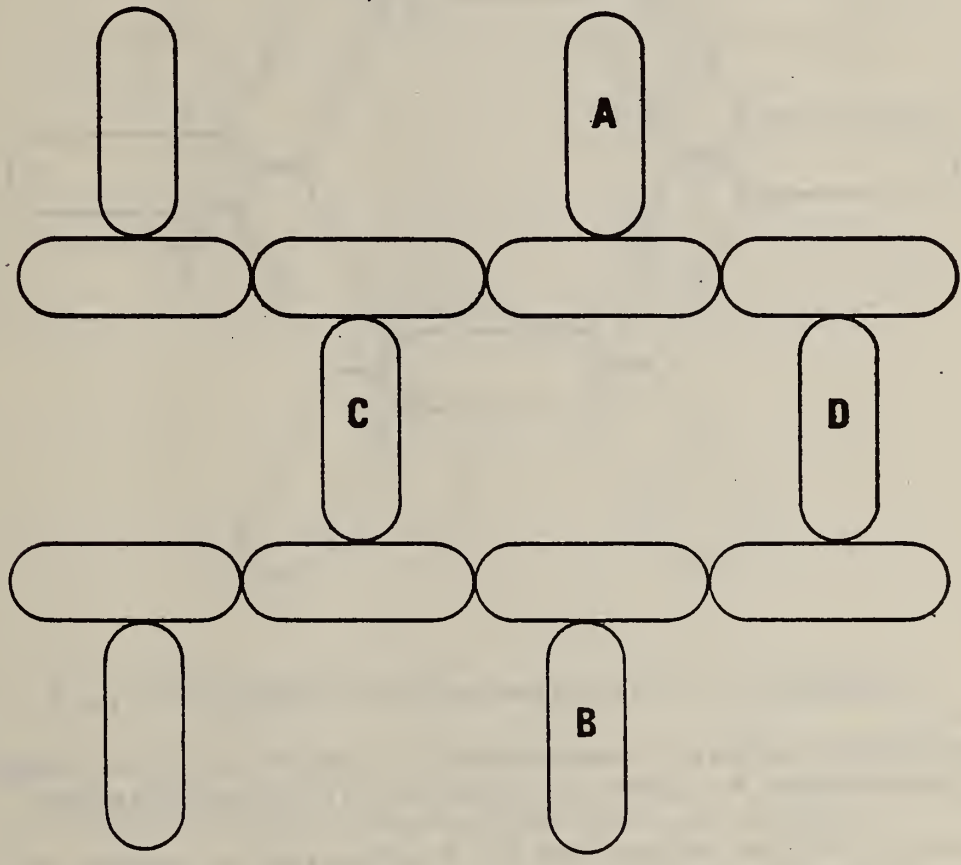


Figure 3. Three-connectivity loop structure

Here, a message entering a loop via one of the connecting points or gateway stations, could leave the loop for a distant destination via either of two other gateways. A message that originated in loop "A" and was destined for loop "B"

would have to traverse five intermediate loops, while a message from "C" to "D" would only have to traverse through three intermediate loops.

When each interior loop is connected to four others, a hexagonal array is formed as shown in Fig. 4.

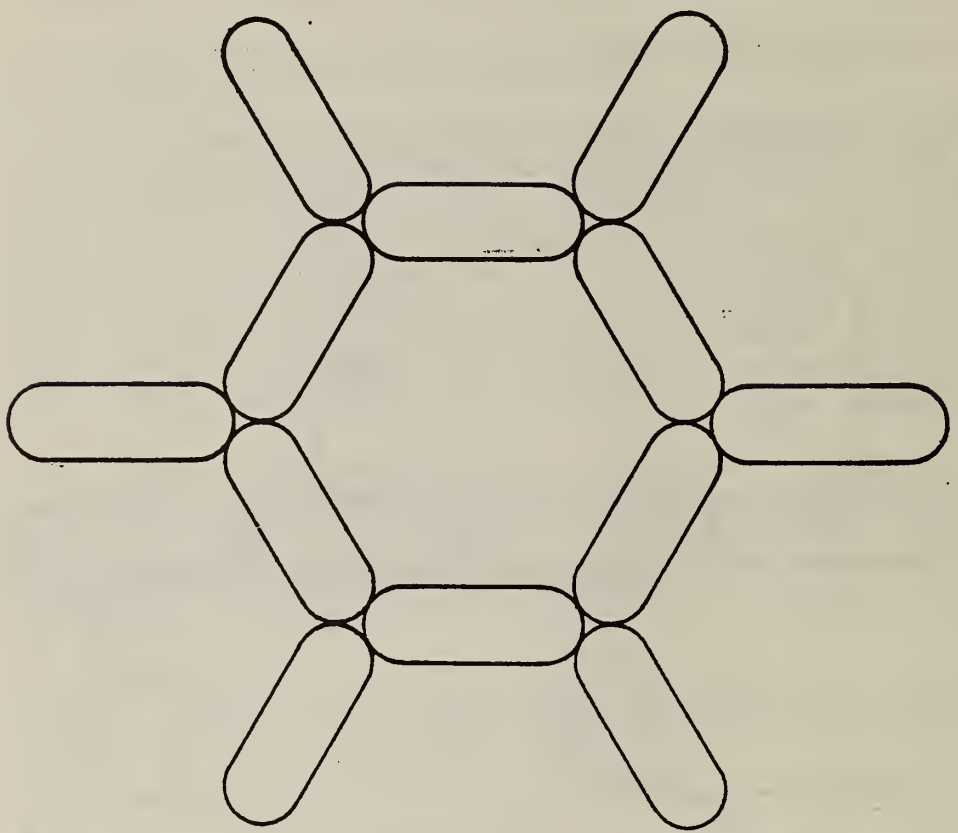


Figure 4. Four-connectivity loop structure

The hexagonal array shown in Fig. 4 could also be viewed as being assembled by joining triplets of loops together.

This array can be distorted by flattening as shown in Fig. 5 to facilitate application of an orthogonal addressing scheme (to be described later).

In this arrangement, a message from "A" to "B" passes through only three intermediate loops, and a message from "C" to "D" traverses only two.

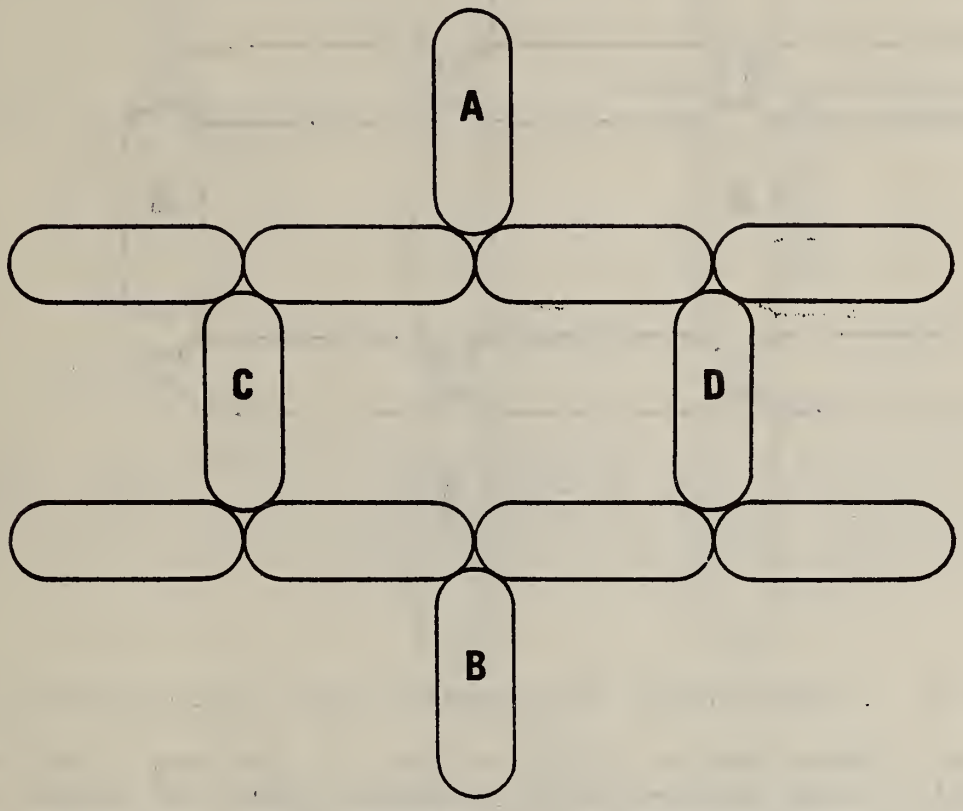


Figure 5. Flattened four-connectivity loop structure

An alternative arrangement in which each interior loop is connected to four others is shown in Fig. 6. In this each have to go through three intermediate enroute loops.

Higher orders of connectivity can be contemplated, but they do not appear to lend themselves to both an orthogonal

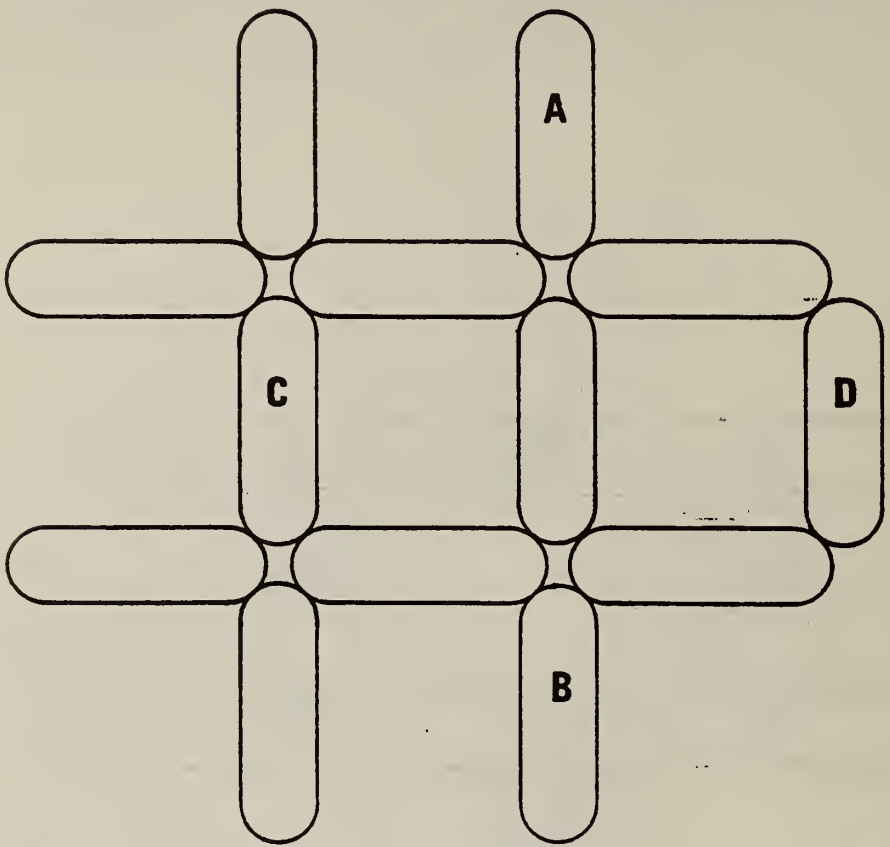


Figure 6. Alternative four-connectivity loop structure

addressing scheme and to a single type of gateway station. For example, a six-connectivity structure could be formed by adding a four-way gateway at the intersection of each row and column of the loops shown in Fig. 6. Then each loop would have four gateway stations that could each communicate with one other loop and two gateways that could communicate with three other loops. While this would provide additional redundancy in numbers of gateways, it would greatly add to the complexity of the routing algorithms.

The flattened four-connectivity structure has been selected as the most reasonable engineering choice to provide adequate redundancy, short routing paths and straightforward routing algorithms.

The basic building block for this structure can be considered to be three loops interconnected by three gateway stations as shown in Fig. 7A, which can be schematically represented in the flattened form in Fig. 7B. The latter

form will be used in illustrating the development of the addressing scheme.

The flattened triple loop structure lends itself well to building large ordered arrays.

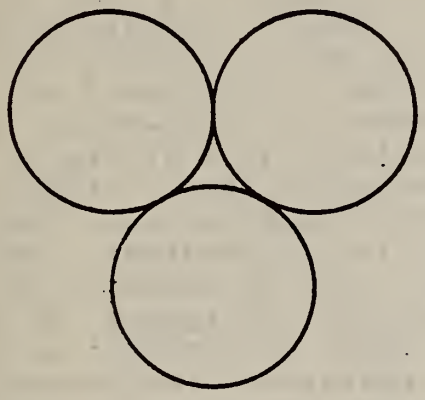


Figure 7A.

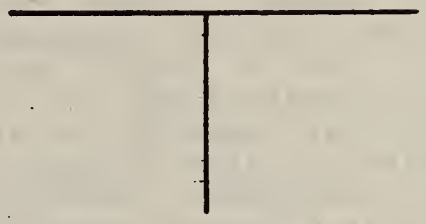


Figure 7B.

In Fig. 8, a number of these building blocks have been assembled into an array forming the flattened four-connectivity loop structure of Fig. 5. This array is the basis for forming the addresses of the GRIDNET stations.

Addresses

The GRIDNET address has three parts designated L, N and S. The first of these, L, represents a "level" counting from the bottom upward on a topological graph of the network. In Fig. 8, the L values for the loops appear along the Y axis of the array. The second address component, N, represents a "loop number". These are shown along the X axis of Fig. 8. The final address component, S, represents the "station number" on a loop. One of the gateway stations on a loop will be assigned the value of one for its S address component, and each of the other stations will be assigned increasingly higher S component address in accordance with their geographic location on the loop as measured in a clockwise direction looking at their graphical representation on the array. The S component of the address will typically range from one to some maximum value as defined for the system. Typically, this maximum will be less than 30. The sequence need not be full, that is, there may be unassigned values, but the assigned values of the S component of the address should increase progressively in a clockwise direction around the loop.

The L and N address components need not be assigned values beginning with the origin of their coordinate system. In fact, it is preferable that they do not, since this will provide room for future expansion of the network in any direction. In a similar fashion, all L,N values need not be occupied, however, total lack of occupancy in a given value of L or of N that would fragment the network is prohibited. In addition, the origin of both L and N should be such that the loops that serve as vertical interconnectors in Fig. 8 have even values for both L and N, while the horizontal interconnector loops have odd values for these address components. The oddness or evenness of the loop addresses is an argument that is used by the routing algorithms.

All of the address components must have integer values greater than zero. These address components are used in developing routing information for messages that flow through the network.

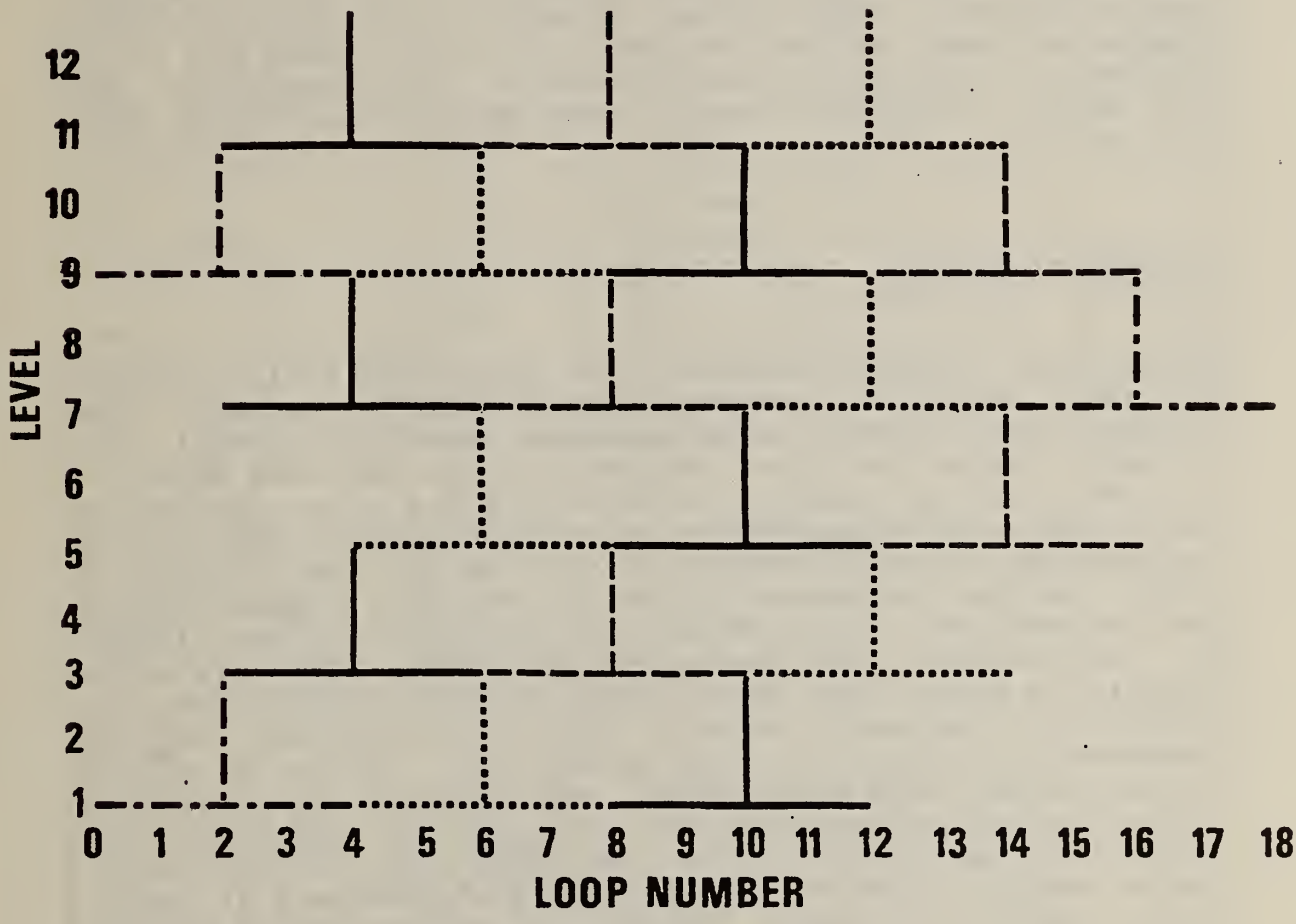


Figure 8. GRIDNET Addressing

The L and N portion of the address of a message are used to determine to which enroute loop it should be forwarded in

order to reduce the distance to its destination L,N. When a message reaches its destination L,N, then the S component designates the station on that loop to which delivery is finally made.
Each loop that is "interior" to the network and fully connected has four gateway stations. Peripheral loops that are only partially connected have fewer gateways, normally two. The gateway is logically two stations, one on each loop. Each of the two logical stations has a different addresses but they can communicate directly with each other by exchanging ownership of buffers containing information that is to be transferred between the two loops. All link control, flow control, and routing functions are performed by the gateways. They also perform certain of the network management functions. Each gateway keeps track of the operational status of all of the stations on its home loop and of the ability of each of the other gateways to communicate with its other (foreign) loop in order accomplish traffic routing.

Gateway Functions

The gateway stations contain the distributed intelligence needed to manage and control the operation of GRIDNET. Each gateway must maintain an independent record of certain information about the other stations on the loop and about the adjacent foreign loops to which it and the other gateways on the loop are connected. This information must be kept current and up to date since it is used by the gateway to make routing decisions, to manage the link communications, and to generate trouble reporting messages that are directed to the network maintenance organization. The information that is of concern to each gateway is the identity and the operational status of each of the other stations, including the other gateways, that are on the loop. It is also concerned with knowledge about the identity and operational status of each of the adjacent foreign loops associated with itself and with each of the other gateways. It maintains a record of the L and N components of the addresses of each of these adjacent foreign loops in a register which has been called the "R" register in the routing algorithm flow charts. There, the i-th entry in the register contains the local loop address of the i-th gateway progressing clockwise around the loop, with self as the zero entry. Each entry associates the local loop address of that gateway with the L and N components of the foreign loop that the gateway can communicate with, and indicates whether or not that foreign loop is operational. In addition to the "R" register data, each gateway maintains a list of all operational stations

that are on its home loop and keeps an indicator of their desire to transmit traffic and their ability to accept traffic directed to them. If there are changes in the operational status, such as failure of a station, or failure of one half of the dual loop, appropriate messages are generated and transmitted in order that proper maintenance actions can be initiated.

Each of the gateway stations on a loop is capable of combined operation. That is, it is capable of functioning as a primary station issuing commands to the other (secondary) stations on the loop, and it is also capable of acting as a secondary station and responding to commands, and it alternates between these two roles. All of the non-gateway stations on the loop are always secondary stations. Gateways are assigned the temporary role of primary station as a consequence of either of two events: 1) The expiration of a predetermined period of time causes primary status to be advanced to the next clockwise gateway that is able to accept this role; 2) if the acting primary, in conducting the normal polling sequence, encounters another gateway that has traffic that it wishes to route to some other station on the loop, it may immediately relinquish primary status to that gateway. The latter procedure permits traffic to be routed forward with the minimum number of hops, while, at the same time, it minimizes message queueing delay and provides a mechanism for resource sharing.

The determination of whether a gateway station will begin operation as a primary or a secondary station is made immediately after power-up. When power is turned on at a gateway, that station starts a timer and clears certain address and status registers. The timer has a duration, t1, that is slightly greater than the maximum propagation delay of the loop plus the maximum response time of a station. During the interval t1, the gateway checks for the presense of any signals on the loop. Signals are an indication that the loop is operational, and that some other gateway station on the loop must be acting as the primary station. The detection of signals during the time interval t1 causes the gateway that has just powered-up to assume the status of a secondary station. It waits for a frame addressed to itself, and responds to that frame in an appropriate manner. Timer t1 is reset whenever signals are observed on the loop. Thus, it serves as a watchdog timer whose expiration initiates possible reassignment of primary status.

If no signals are observed during the duration of the time interval t1, a second timer is started. This timer has a duration, t2, that is zero for the gateway having an S address component of one and increases by an amount equal to the maximum propagation delay time for each gateway having

an increasingly higher value of S address component. For example, if a loop had four gateway stations, with respective S addresses of 1, 5, 9, and 17, and the maximum propagation delay the loop was 500 microseconds, then the t2 value at S1 would be zero, the value at S5 would be 500, and it would be 1,000 at S9 and 1,500 at S17. Each gateway station looks for signals on the loop until the expiration of t2. If any signal is detected, the station assumes the role of secondary. The first station at which t2 expires before signals are observed assumes the role of primary station and immediately begins transmission. This procedure avoids contention among the gateway stations for the initial role of primary station, and assures that the initialization functions will be started by the operable gateway station with the lowest value of S address component. It also provides a mechanism by which loop operation can be resumed if a gateway becomes inoperable while it is functioning as the current primary station.

The gateway at which t2 expires first following initial power-up, or following the expiration of t1 resulting from loss of signals due to malfunction or other failure, assumes the role of primary station and begins transmission. Simultaneously, it starts a third timer, t3, whose nominal duration is one minute. Expiration of this timer is the basis for initiating the orderly transfer to the role of primary station to the gateway having the next higher (modulo m) value of S address component, where m is a constant, fixed, maximum number of stations that any loop can accomodate. Timer t3 is started whenever any gateway becomes the acting primary station, either as a result of normal rotation, or because of request in order to expedite the flow of traffic.

Upon assuming the role of primary, the station immediately begins a polling sequence by transmitting a command addressed to the station having the next higher S address. This is repeated, incrementing the S address component (modulo m) until all possible addresses on the loop have been tried and the sequence is completed. The secondary addresses that are occupied by operable stations respond to this command and provide information about their status. These responses permit the primary station to compile a list of operable secondary station addresses on the loop and their status with respect to traffic. If the secondary station responding to this command is a gateway, the response frame contains the address of the next adjacent foreign loop with which the gateway can communicate, and an indication of whether or not that foreign loop is operational. These additional data elements are entered in a register that is designated the "R" register at the primary station and are used to support routing procedures. The purposes of this initial poll sequence are to fill the "R" register and to

develop a list of all of the occupied, and operable, S addresses on the loop. On subsequent poll sequences during the nominal one minute period that the status of primary station is retained, only the members of this list are polled.

The acting primary station continues to poll each of the stations on the list, accepting or delivering frames containing information fields as appropriate. The polling is sequential, modulo m, until the expiration of timer t3. When this occurs, the poll sequence is continued until any outstanding frames have been acknowledged and the gateway with the next higher (modulo m) S address is reached. Primary station status is transferred to this station. If, for any reason, that station is unable to accept the role of primary station, the polling sequence is continued and transfer of status is attempted at the next gateway that is reached. If no other gateway is able to accept primary status, the gateway that is currently acting as primary station continues to do so, and continues to attempt to transfer this function to the other gateways in turn. After three complete poll sequences have been completed without the transfer of primary station status, and if there are still routing paths available, a trouble message is generated and dispatched to the network maintenance facilities.

If, in the course of the normal polling sequence, the primary encounters another gateway that is holding traffic that should be routed to some station other than the acting primary, then primary status may be transferred immediately to that gateway.

When primary station status is transferred to a new gateway, the t3 timer at that station is started, and if the transfer of primary status was because of waiting traffic, that traffic is immediately dispatched. This is followed by a polling sequence which attempts to reach all possible loop addresses and the accumulation of occupied address lists and "R" register data entries. (This process is begun immediately if primary status transfer was based on timer expiration rather than waiting traffic.) These are compared with the data accumulated on the immediately preceeding assignment to the role of primary station. If there are additions, either to the occupied address list or to the operable foreign loops in the "R" register, these are accepted and no further action is necessary. This would be the expected result of service restoral following an outage, or of the installation of a new station on the loop, or even the initial assignment of primary status following power-up.

If there are deletions in either list, this will indicate that an outage has occured. In this event, a trouble

message is generated and directed to network maintenance. In addition, messages are directed to each of the other gateway stations on the loop instructing them to delete the appropriate items from their lists or registers in order to inhibit them from generating duplicate trouble messages as they, in turn , are designated primary stations.

Gateway stations are provided with seven frame buffers and the other stations have three frame buffers. All of the frame buffers are large enough to accomodate a frame with an information field of 1,000 bytes. When a primary station receives a frame or frames that are destined for another station on the loop, it forwards them to that station as soon as it finishes its communication with the current secondary. Then it returns to the place in the polling sequence where it left off. Messages that are to be routed to another loop by a gateway are transferred immediately to the corresponding gateway partner on the other loop. Stations that are unable to accept a frame because of full buffer conditions indicate their status with the Receive Not Ready (RNR) command/response, and the station that wishes to send the frame will then hold it until the next time around on the polling sequence. If, after three such attempts, the frame still can not be accepted by the preferred next destination station, alternate routing is attempted following the same rules that would be observed if the preferred next destination station were inoperative.

The primary station is responsible for disposition of each frame, and will make retransmissions as appropriate in order to obtain a positive acknowledgement of acceptance by the next enroute secondary. If exception conditions preclude successful acceptance by the preferred or any alternate secondary station, the message is returned to the originator together with indication of its undeliverability. When a message reaches the destination addressee, that station initiates an acknowledgement message that is directed back to the originator as an indication of successful delivery. Failure of the originator of a message to receive either an end to end acknowledgement or a message that has been returned as undeliverable within a system dependent time interval is accepted as evidence that the message was not delivered.

MESSAGE ROUTING

Message routing is accomplished by the gateway stations under the direct control of the acting primary station.

This is done as a result of examination of the L and N message address components and comparing them with the L and N address components of the home loop and those of each of the candidate "foreign" loops with which the other gateways can communicate. If destination "level" minus current, local loop "level" and destination "loop number" minus local "loop number" are both zero, that indicates the message is destined for the local loop, and it is delivered to the proper station on the loop.

If the message address does not correspond with the L and N components of the local loop address as described above, it is destined for a foreign loop. The loop to which it is routed is determined in accordance with the following algorithms.

Routing Algorithms

In selecting the new loop to which a message is forwarded, subtract the L and N components of the current loop address from the L and N components of the destination address. Compare these differences in L and N address components with the corresponding differences that are obtained from each of the foreign loops with which the gateways can communicate. Do not include the foreign loop from which the message came in making this comparison. Do not evaluate any foreign loop address as a candidate for routing when that address has been appended to the appropriate heading item in the information field of the frame as a "no return" indicator.

Call the difference in the L components, Delta L, and the difference in the N component, Delta N, in referencing the following rules:

1. If the current loop address is ODD and its Delta L magnitude is greater than one, select the first foreign loop address encountered that will reduce the magnitude of Delta L.

1a. If 1 is unsuccessful, or if the Delta L from the current loop address has a magnitude that is zero or one, try to find a foreign loop address that will reduce the magnitude of Delta N by two.

1b. If 1a. is unsuccessful, try to find a foreign loop address that will reduce the magnitude of Delta N by only one.

1c. If 1b is unsuccessful, and the magnitude of Delta L is zero or one, append the current loop address to the appropriate header item of the message and try to send it back to the foreign loop where it came from.

1d. If 1b. is unsuccessful, and the magnitude of Delta L is greater than one, try to find a foreign loop address where the magnitude of Delta L is the same and the magnitude of Delta N is the same.

1e. If 1d. is unsuccessful, try to find a foreign loop address where the magnitude of Delta L is the same and the magnitude of Delta N is increased by two.

1f. If 1e. is unsuccessful, append the current loop address and insert the no return header item if it is not already in the message and try to return it to the loop where it came from.

1g. If 1f. is not posssible because that address has already been appended, change the header item so that it will reverse the source and destination addresses and indicate that delivery is not possible and route the message back toward the originator.

2. If the current loop address is EVEN, try to find a foreign loop address that will reduce the magnitude of Delta N while keeping the same, or smaller, magnitude of Delta L.

2a. If 2 is not possible, try to reduce the magnitude of Delta N while only permitting the magnitude of Delta L to increase by one.

2b. If 2a. is unsuccessful, try to find a foreign loop address that will cause equal or smaller magnitude of Delta L.

2c. If 2b. is unsuccessful, try to find a foreign loop address that only increases the magnitude of Delta L by one.

2d. If 2c. is unsuccessful, append the current loop address and reverse the source and destination addresses using the appropriate header items and return to the sender.

3. If rule 1c. or 2d. is executed, set the "no return" flag using the appropriate header item of the message and this will cause the addresses of all loops that it may subsequently traverse to be appended in that header item as long as it remains set. In all cases where a candidate loop address has been selected, do not forward it to that address

if that address has been already appended to the "no return" header item of the message. Always append the current loop address before forwarding the message if the "no return" flag (header item) is set.

3a. Remove the "no return" header item and all appended addresses whenever the message in which it has been previously set enters a loop from which there are three gateways that may be considered as candidates for exit routes under the rules.

4) If there is only a single gateway leading to a foreign loop that is eligible for selection, select it. Append the address of the loop the message is leaving to the frame and set the "no return" indicator in the appropriate header items of the frame to prevent that loop from being re-entered.

5) When a gateway is not able to communicate with its immediately adjacent foreign loop, that loop is excluded from consideration in determining routing.

6) If the station that is selected as the next recipient of a message is unable to accept it because of full buffer conditions, the station holding the message will store it until the next opportunity that it has to forward it, but no more than three poll cycles. If the intended next station is then still unable to accept it because of full buffer conditions the holding station will:

6a) If the selected station is a gateway, the holding station will select alternate routing just as if that gateway had been inoperative.

6b) If the selected station is the final destination (i.e., is on the same loop as the holding station), the holding station will transpose source and destination addresses by setting the appropriate heading item indicator and return the message to the sender as an indication of non-delivery.

7) If a message cannot be routed further because of "no return" appendages, this is interpreted as an indication that it is undeliverable. The gateway making this determination shall then transpose the source and destination addressesby setting the appropriate header item indicator and delete all of the "no return" addresses and their associated header item, and release the message for return to its source as an indication of non-delivery.

Flowcharts

The basic routing rules have been drafted in flow chart form for clarity, using the following notation:

A = (L,N) = Address (L and N components)

L = Loop level component of address

N = Loop number component of address

Subscripts:

d = message destination

c = gateway currently holding the message

i = index of register maintained at each gateway that contains the address, operational status and buffer availability of the foreign loops which that gateway, and the other gateways on the loop can route traffic.

f = operational foreign loop with which a gateway can communicate

o = message originator

Delta = Difference in address components L, N.

Delta Lc = Ld - Lc, i.e., refers to current loop

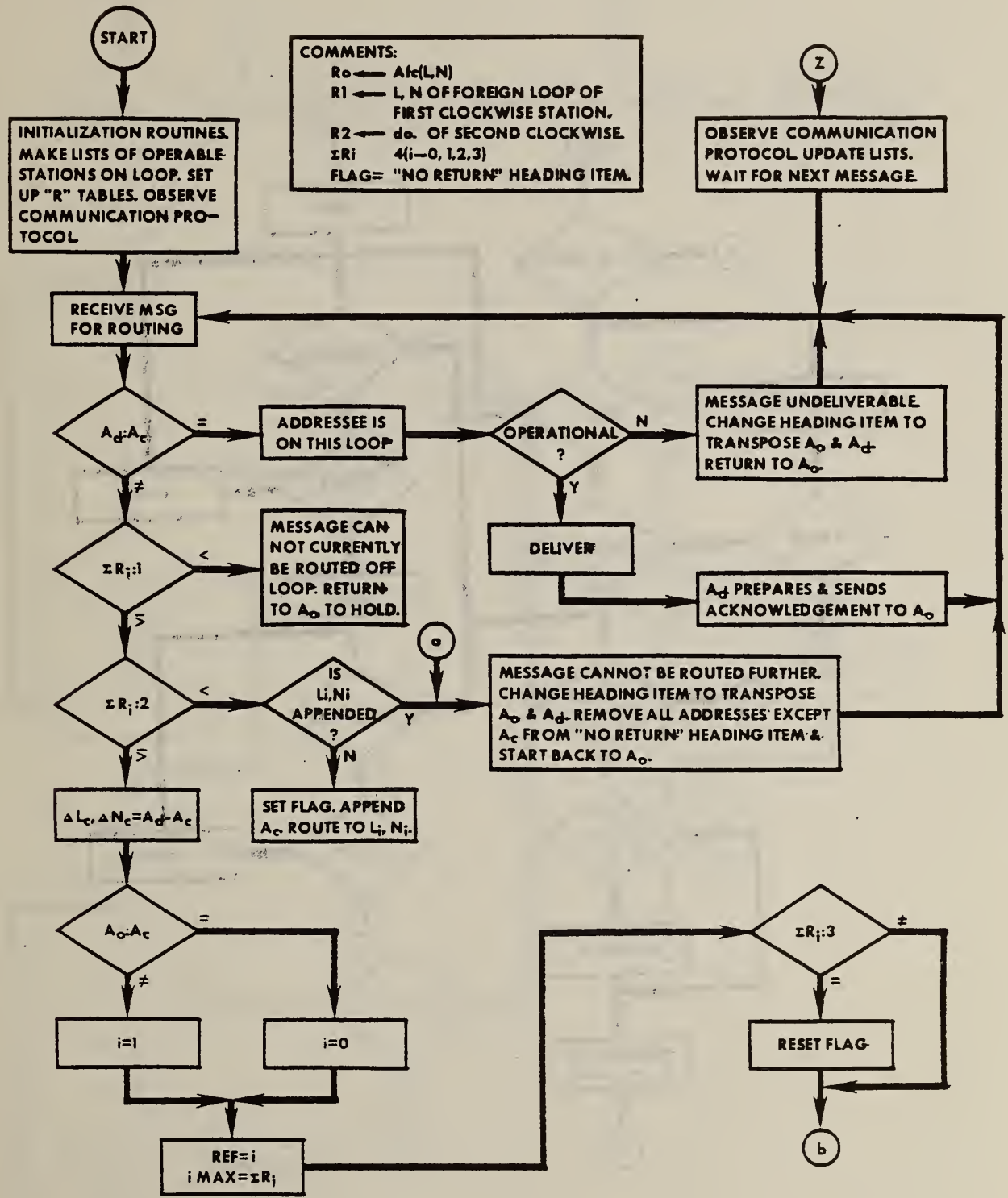
Delta Li = Ld - L of the ith operational foreign loop

Ri = i-th Register containing the foreign loop address of the operational loop with which the ith gateway can communicate. R sub zero contains the foreign loop address of the home gateway (self). R sub one contains the foreign loop address of the next gateway going clockwise around the loop., etc. i may attain a maximum value of three.

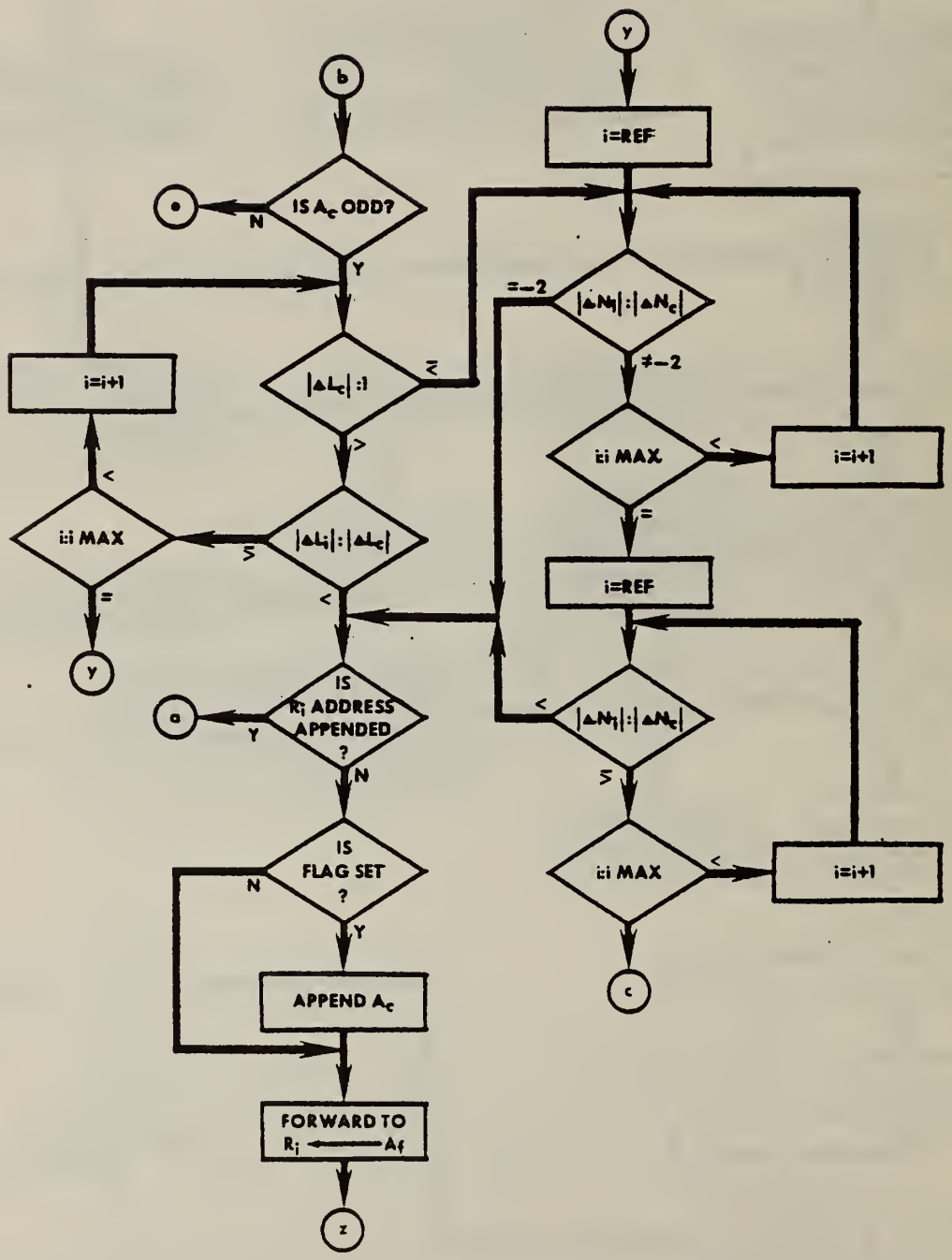
A sub fc is the address of the foreign loop that the gateway currently holding the message can communicate with.

Sigma R sub i is the total number of operational gateways including the one associated with self.

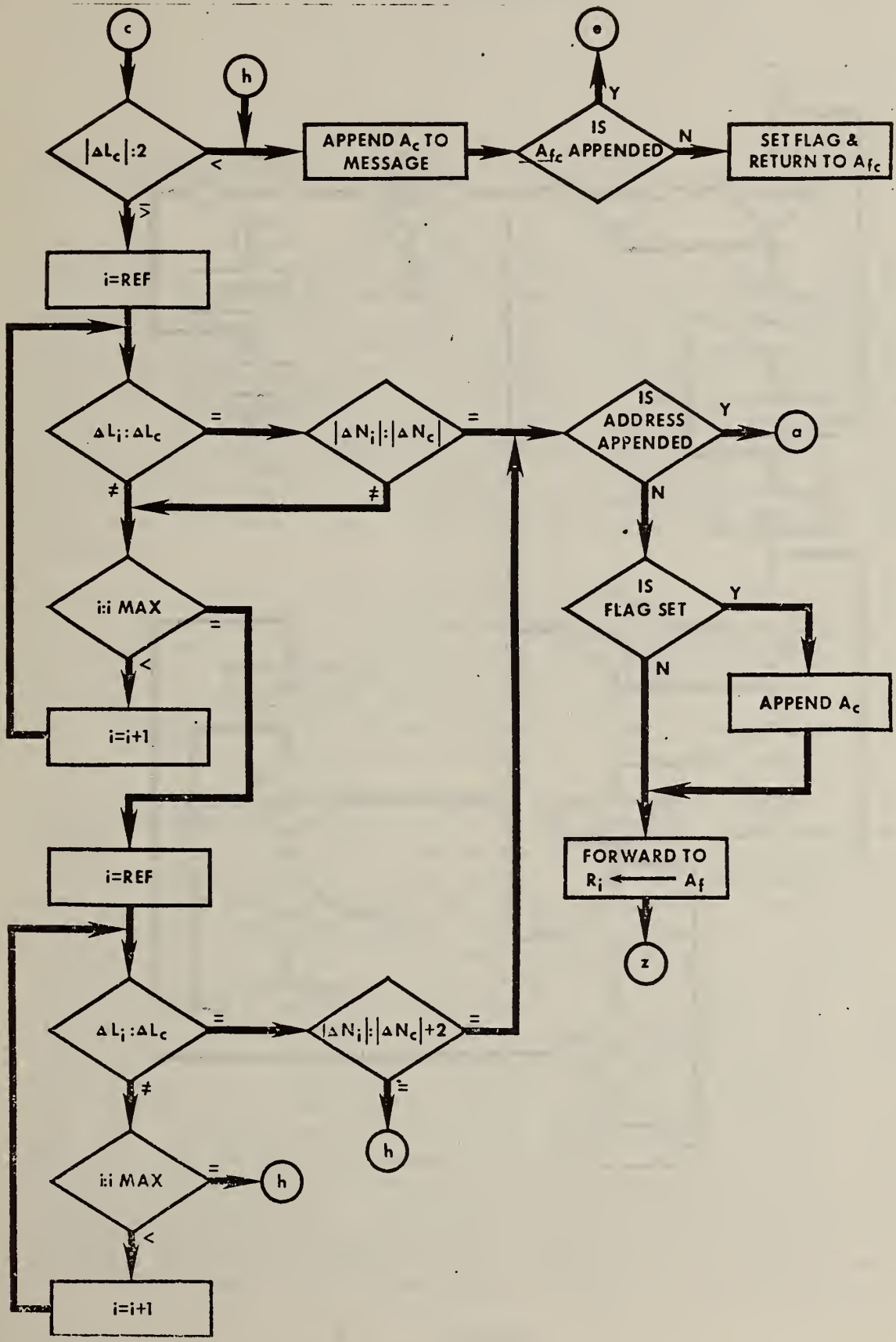
FLOWCHART 1


START
INITIALIZATION ROUTINES. MAKE LISTS OF OPERABLE STATIONS ON LOOP. SET UP "R" TABLES. OBSERVE COMMUNICATION PROTOCOL.
COMMENTS:
Ro ⟵ Afc(L,N)
R1 ⟵ L, N OF FOREIGN LOOP OF FIRST CLOCKWISE STATION.
R2 ⟵ do. OF SECOND CLOCKWISE.
ΣRi 4(i=0, 1,2,3)
FLAG= "NO RETURN" HEADING ITEM.
Z
OBSERVE COMMUNICATION PROTOCOL. UPDATE LISTS. WAIT FOR NEXT MESSAGE.
RECEIVE MSG FOR ROUTING
Ad:Ac
=
ADDRESSEE IS ON THIS LOOP
OPERATIONAL ?
N
Y
MESSAGE UNDELIVERABLE. CHANGE HEADING ITEM TO TRANSPOSE Ao & Ad. RETURN TO Ao.
≠
ΣRi:1
<
MESSAGE CAN NOT CURRENTLY BE ROUTED OFF LOOP. RETURN TO Ao TO HOLD.
DELIVER
Ad PREPARES & SENDS ACKNOWLEDGEMENT TO Ao
≥
o
ΣRi:2
<
IS Li,Ni APPENDED ?
Y
N
MESSAGE CANNOT BE ROUTED FURTHER. CHANGE HEADING ITEM TO TRANSPOSE Ao & Ad. REMOVE ALL ADDRESSES EXCEPT Ac FROM "NO RETURN" HEADING ITEM & START BACK TO Ao.
≥
ΔLc, ΔNc=Ad-Ac
SET FLAG. APPEND Ac. ROUTE TO Li, Ni.
Ao:Ac
=
≠
i=1
i=0
REF=i
i MAX=ΣRi
ΣRi:3
±
=
RESET FLAG
b

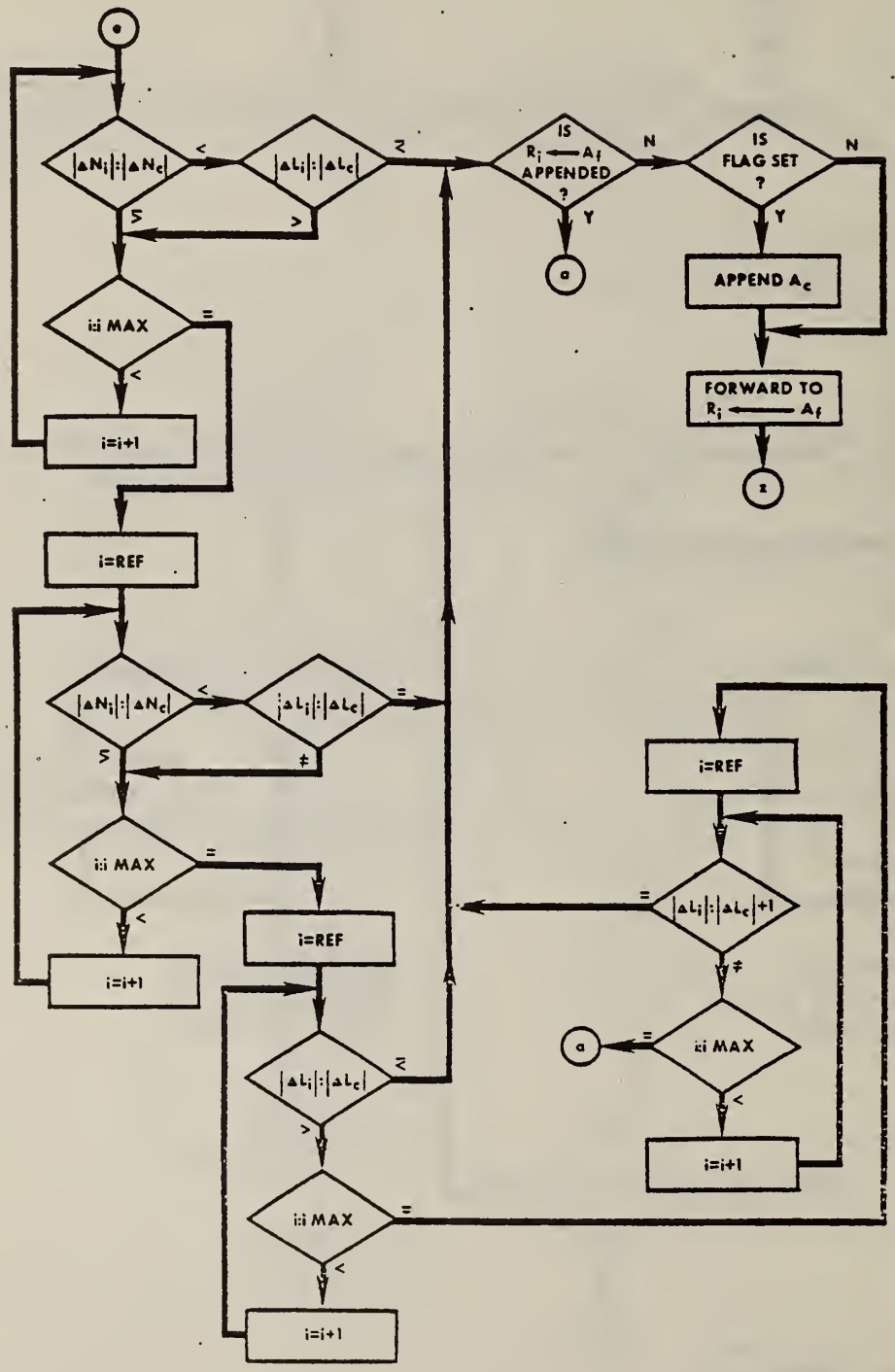
FLOWCHART 2


b
IS A_c ODD?
N
Y
a
i=i+1
$|\Delta L_i| : 1$
=
>
<
i:i MAX
$|\Delta L_i| : |\Delta L_c|$
<
=
>
y
IS R_i ADDRESS APPENDED ?
Y
N
IS FLAG SET ?
N
Y
APPEND A_c
FORWARD TO $R_i \leftarrow A_f$
z
y
i=REF
$|\Delta N_i| : |\Delta N_c|$
=-2
≠-2
i:i MAX
<
i=i+1
=
i=REF
$|\Delta N_i| : |\Delta N_c|$
<
>
i:i MAX
<
i=i+1
c

FLOWCHART 3


c
h
$|\Delta L_c|:2$
<
APPEND A_c TO MESSAGE
o
Y
IS A_{fc} APPENDED
N
SET FLAG & RETURN TO A_{fc}
≥
i=REF
$\Delta L_i:\Delta L_c$
=
$|\Delta N_i|:|\Delta N_c|$
=
IS ADDRESS APPENDED
Y
a
≠
≠
N
i:i MAX
=
<
i=i+1
IS FLAG SET
Y
N
APPEND A_c
i=REF
FORWARD TO $R_i \leftarrow A_f$
z
$\Delta L_i:\Delta L_c$
=
$|\Delta N_i|:|\Delta N_c|+2$
=
≠
=
h
i:i MAX
=
h
<
i=i+1

FLOWCHART 4


|ΔNi|:|ΔNc|
<
|ΔLi|:|ΔLc|
≥
IS Ri ← Af APPENDED ?
N
IS FLAG SET ?
N
Y
Y
a
APPEND Ac
FORWARD TO Ri ← Af
z
≤
>
i:i MAX
=
<
i=i+1
i=REF
|ΔNi|:|ΔNc|
<
|ΔLi|:|ΔLc|
=
≤
≠
i:i MAX
=
<
i=i+1
i=REF
|ΔLi|:|ΔLc|
≥
>
i:i MAX
=
<
i=i+1
i=REF
|ΔLi|:|ΔLc|+1
=
≠
a
i:i MAX
=
<
i=i+1

LINK LEVEL PROTOCOL

The communications control protocol that is used on the CROSSFIRE dual loop configuration is defined in ANSI X3.66-1979, American National Standard for Advanced Data Communication Control Prodedures (ADCCP). This protocol could be employed on GRIDNET as well, but improved performance can be obtained with the use of a modified version that has been optimized for GRIDNET. These modifications retain the basic structure of ADCCP but restrict operation to the use of unnumbered commands and responses in a "select/hold" mode of operation. In addition, certain non-reserved commands and responses are implemented and there are modifications to the interpretation of certain other commands.

The basic frame structure as described in section 3 of ANSI X3.66 is employed, but the address and control fields are not extended and are limited to a single octet. The S component of a station address is used in the address field of a frame while the full address (L, N and S) of the source and destination of a frame is carried in a designated location within the information field.

In X3.66, information frames are sequentially numbered with a modulus equal to eight for the basic control field format. This permits a station to have up to seven outstanding and unacknowledged frames at any given time, but it require that each station maintain certain state variables. Every secondary station maintains a send variable on the I frames it transmits to, and a receive variable on the I frames it correctly receives from, the primary. Each primary station must maintain a corresponding pair of variables for every secondary station with which it communicates. The maintenance and continual resetting of these state variables with each rotation of primary status among the gateway stations of GRIDNET is an undesirable load of overhead traffic. It can be avoided by operating in a "select/hold" mode where the primary station is responsible for insuring that acknowledgements are exchanged with the secondary station on a frame by frame basis. This mode of operation is particularly appropriate to the dual loop, link level configuration of GRIDNET, since the response to a frame is at least partly predicated upon the comparison of FCS and frame contents as

received on both the clockwise and counterclockwise loops. With acknowledgement (either positive or negative) on a frame by frame basis there can never be more than a single outstanding, unacknowledged frame. Therefore, there is no need for frame numbers and the maintenance of state variables to accomodate them. If a primary station is unable to obtain an acknowledgement for a frame, that is an immediate indication of trouble and is cause for the generation and dispatch of a message requesting maintenance action.

Since GRIDNET is always operated in the "select/hold" mode, there is no requirement for the mode setting commands and responses as defined in X3.66, so these are not implemented.

The Poll/Final (P/F) bit is used for flow control in this mode, but in a way slightly different than defined in X3.66. The primary station always sets the P/F bit to one when it expects a response from a secondary and sets it to zero when it does not want a response from the secondary. An example of the latter condition occurs when the primary station acknowledges a final information frame received from the secondary station. The secondary station always sets the P/F bit to one when it has no requirement to transmit another frame and sets it to zero as a request to the primary for permission to transmit another frame.

Commands and Responses

Commands and responses that are implemented are a subset of those identified in ANSI X3.66. Where their definition, use or interpretation differs from that given in X3.66 it is described below.

Exchange Identification (XID). The XID command is used to cause the addressed secondary/combined station to report its identification and status. The status report is contained in the information field of the XID response frame. If the responding secondary is not a gateway, the information field consists of a single octet. The first four bits of this octet are zeros (reserved for future codings). The fifth bit is a one if the frame was received only on the clockwise loop and the sixth bit is a one if it was received only on the counterclockwise loop. The seventh bit is a one if the station has traffic waiting to be transmitted, otherwise it is zero. The eighth bit is a one if the station is unable to accept a frame containing an information field (e.g. full buffers), otherwise it is a zero. If the responding secondary is a gateway, the information field contains three

octets. The first three bits of the first octet are zeros (reserved for future codings). The fourth bit is a one if the adjacent foreign loop associated with that gateway is inoperable. (This inhibits data entry in the "R" register of the primary.) The fifth through eighth bits convey the same meaning as for non-gateway secondaries. The second and third octet are the binary representations of the L and N address components of the adjacent foreign loop associated with the gateway. The P/F bit is always set to one in the XID command and in the response frame the P/F bit is zero only when bit seven of the first octet is one.

Unnumbered Poll (UP). The UP command is used by the primary to solicit response frames from a secondary. It is sent with the P/F bit set to one. If the secondary station has no traffic it will respond with a UA. If it has traffic for that primary it will respond with a UI and include an information field in the response frame. The P/F bit will be set to zero if the station has more traffic, otherwise it will be set to one. If it is a gateway and is holding traffic that it wishes to route to another gateway that is not currently the primary, it will respond with a PI with the P/F bit set to one.

Primary/Information (PI). PI is a nonreserved response having the bit pattern 11011000. It is used by a gateway that is also a secondary and is currently holding traffic that is to be routed to another gateway that is not the current primary. It signifies a desire to be assigned the role of primary at the earliest opportunity so that the traffic can be forwarded directly to the required gateway without having to be relayed through the current primary.

Primary Status (PS). The PS command directs the addressed gateway secondary to assume the role of primary. The PS response is the acceptance of the role of primary. The PS is another of the nonreserved command/responses and has the bit pattern 11011001.

Unnumbered Information (UI). The UI command directs the addressed station to accept a frame containing an information field. The information field will contain header information subfields that must be examined by the station in order to make appropriate disposition of the traffic. The response to a UI command is UA for positive acknowledgement or REJ for negative acknowledgement. The UI response is used as a response to the UP command when the addresses secondary has traffic for the primary. Under these circumstances the primary will accept frames using UA or require their repetition using the REJ. It will control the flow of frames with the P/F bit.

Unnumbered Acknowledgement (UA). The UA is a response used for positive acknowledgement.

Reject (REJ). The REJ command/response is a negative acknowledgement and request that the frame be retransmitted.

Frame Reject (FRMR). The FRMR response is used to report an error condition that is not recoverable by retransmission of the identical frame, such as receipt of a control field that is invalid, or receipt of a UI frame with an information field that exceeds the maximum length.

Receive Not Ready (RNR). RNR is used only as a response by a station to indicate a "busy" condition; i.e., the temporary inability to accept additional frames containing information fields. It is a positive acknowledgement of the immediately preceding frame received by the station.

Receive Ready (RR). The RR command/response is used by a station to indicate that it is ready to receive a frame with an information field. An RR frame is one way to report the end of a station busy or buffers full condition.

Heading Items

Heading items are communication control data that are contained within the information field of frames. They contain destination and source addresses and the addresses of loops to which the frame should not be returned. Their contents are examined, and changed where required by the routing rules. The general arrangement of frames is shown in Figure 9 which shows the sequence of fields and heading items.

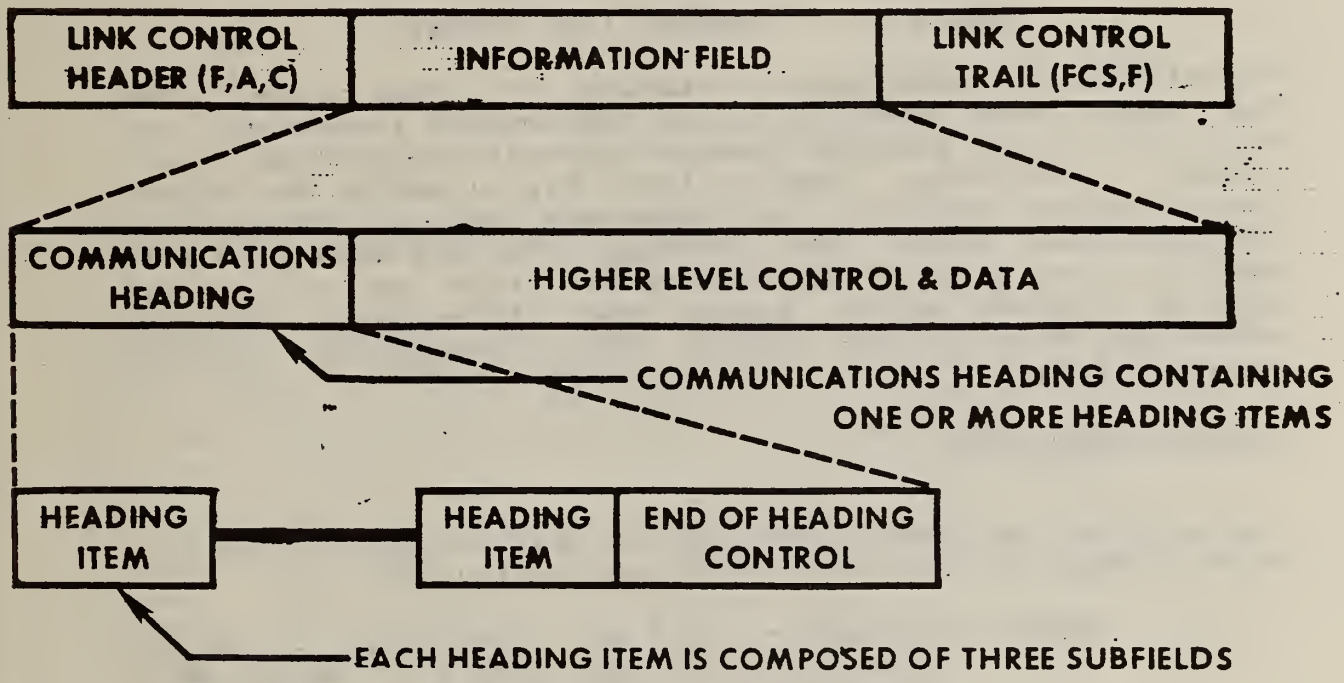


Figure 9. Sequence of Fields and Heading Items

The heading items that are defined herein are assembled in octet units and are portrayed in the left to right sequence in which the octets would be transmitted. Numeric values (binary) within each octet are shown with the high order bit on the left and the lowest binary weight bit (transmitted first) on the right.

Each heading item is composed of three subfields. The first of these is an octet containing a unique heading item code that identifies the nature of the information that is contained in the item. The second is an octet that defines the length in octets of the immediately following, variable length parameter subfield. The third is the parameter subfield itself. The arrangement is illustrated in Fig. 10.

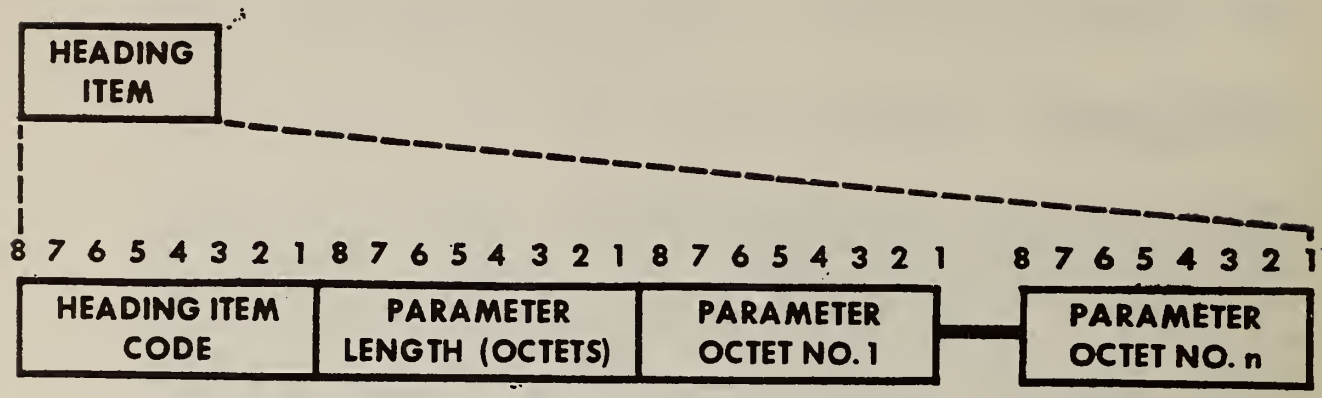


Figure 10. Heading item format.

Frames that are addressed to stations on loops other than the local loop and containing information fields will require one or two heading items to support the routing algorithms. Additional heading items may be useful for higher level control functions, but these are not a subject for consideration here. The heading items are separated from the following data in the information field by an end of heading control octet having the bit sequence 10000000, where the zeros are low order, and transmission is from right to left.

Heading Item Codes

The heading item code, subfield one, may use any of the following sequences and associated interpretations:

11000001 Parameter field will contain L, N, and S components, in that order, of the ultimate destination(s) of the frame. Values are binary, one octet long. The last three octets are the L, N, and S components of the address of the source of the frame.

This is the basic information required for forward routing of a frame. For single destination frames the parameter field is six octets long.

11000010 This frame has been determined to be undeliverable to the destination(s). The parameter field contains the same data elements listed above, but the last three octets (originally the source) are now the destination to which this

frame is to be returned as an indication of non-delivery.

11000011 The parameter field contains the L and N components (only) of loop addresses to which this frame should not be returned. These addresses are not candidates for further forward routing. Under some circumstances this field may increase in length by two octets each time the frame is advanced to a new loop.

PERFORMANCE DETERMINATION

In order to develop reasonable estimates of the performance of GRIDNET, a computer simulation of the network is planned. The objectives of this simulation are to validate the protocol and routing algorithms that have been developed and to determine what patterns of loop or node outages would be required in order to defeat the alternate routing capabilities of the system. In addition, the simulation will be used to develop estimates of the system performance in terms of transfer rates, capacity and message delays as a function of network size and signalling rate. The simulation will also provide insight regarding the adequacy of the buffer capacities proposed for the GRIDNET stations.

The model used in the simulation will reflect the following assumptions:

1) Uniform loop length of 100 kilometers.

2) Population of 20 stations per loop.

3) Inter-loop rather than intra-loop traffic.

4) Uniform message information field size of 1,000 bytes.

5) Seven buffers at each gateway station.

6) Three buffers at non-gateway stations.

The simulation will be written in a higher level programming language such as FORTRAN. This will insure a reasonable degree of portability and enhance its potential for future utility. Pending results from this computer based simulation, limited manual testing of the routing algorithms has been accomplished. These tests have disclosed, that with a fully operable network, a message is routed from a source to a destination loop via the most direct intermediate path. As loops and nodes are disabled, the algorithms cause a message to follow more circuitous routes around the trouble locations in order to reach their destination. The extent and distribution of outages that can be tolerated while still maintaining a degraded level of service can not be readily determined using this mode of test.

LOOP NETWORK INTERCONNECTION

The Appendix to this report is a study conducted by Hemant Kanakia of Sytek, Incorporated, entitled "Interconnection of Loop Networks: A Survey" which reviews and evaluated a number of interconnection schemes for loop networks which have been reported in the literature. Their respective strengths and weaknesses are identified, and the report concludes that the GRIDNET routing scheme does better than other routing schemes on most counts.

BIBLIOGRAPHY AND REFERENCES

Moore, R. T., et al, "Computerized Site Security Monitor and Response System," NBSIR 77-1262. National Bureau of Standards, June 1, 1977, NTIS PB 269 346

Moore, R. T., et al, "Phase II Final Report Computerized Site Security Monitor and Response System," NBSIR 79-1725, National Bureau of Standards, March 1979, NTIS PB 294 343.

ANSI X3.66-1979, American National Standard for Advanced Data Communication Control Procedures (ADCCP), American National Standards Institute, Inc. 1430 Broadway, New York, N. Y. 10018.

FIPS-71 Advanced Data Communication Control Procedures (ADCCP).

FIPS-78 Guideline for Implementing Advanced Data Communication Control Procedures.

APPENDIX

INTERCONNECTION OF LOOP NETWORKS: A SURVEY

PREPARED
FOR:

NATIONAL BUREAU OF STANDARDS
INSTITUTE FOR COMPUTER SCIENCES AND TECHNOLOGY
WASHINGTON, D. C. 20234

UNDER
CONTRACT NUMBER NB0NA-AE4360

PREPARED BY:
HEMANT KANAKIA
SYTEK, INCORPORATED

CONTENTS

Interconnection of loop networks: Survey

Sytek, Incorporated

1. INTRODUCTION

Networks in which nodes are connected in a ring-like topology are known as loop networks. Connecting large number of geographically apart nodes using a single loop can lead to very long response times. Another disadvantage to using a single loop in this situation will be reduced reliability.

Interconnected loop networks is a better design for connecting large number of geographically distributed nodes. Briefly stated, the advantages of interconnected loop networks are increased fault-tolerance, more efficient services, and making possible the division of operating authority.

In this paper we will survey various interconnection schemes proposed for loop networks. We will also mention some other relevant schemes for interconnecting different types of networks. Since the problems faced by these other schemes are similar in nature, these schemes may be found useful in evaluating various interconnection schemes.

The interconnection schemes are evaluated within a framework defined in section 2.2 of this paper. An attempt is made to define a framework which is more relevant to GRIDnet environment.

In section 2.0 we will discuss general issues involved in interconnecting any type of networks and will state the desirable characteristics of interconnections of loop networks.

In section 3.0 we will classify various routing schemes. Unlike the previous work done in this area, this classification is able to include loop switching methods based on hamming's distances and GRIDnet type of routing schemes.

In section 4.0 we will describe all the known loop interconnection schemes. Among these are Pierce's hierarchical rings, DCS multiring extension, and other destination address based algorithmic routing schemes.

The routing schemes used in other types of networks or interconnections of other type of networks can be conceivably used for interconnecting loop networks. Hence in section 5.0 we describe some of the routing schemes used in other types of networks.

The discussion is summarized in section 6.0 and supplemented with a reference list. For ease of referencing, reprints of some of the important methods have been collected as an appendix.

2. NETWORK INTERCONNECTIONS

In this section we will discuss general characteristics of interconnection schemes, define terms normally used, and enumerate performance criteria to be used in evaluating these schemes. We will discuss individual schemes in more detail in sections 3.0 and 4.0.

2.1 Issues Involved in Network Interconnections

Interconnection of networks implies more than providing just physical interconnection links between networks. The minimum capability expected is being able to move messages from a source node to a given destination node across network boundaries. This is referred to as a routing process. In addition to this minimum service, the scheme may choose to have reliable delivery of messages by using end-to-end acknowledgements and retransmissions for lost messages. The other possible services that a scheme may choose to provide can be that of delivering successive messages in sequence, and providing end-to-end flow controls.

A node where messages are removed from one network and inserted in the other network is known as a gateway. Instead of being just one physical node on both networks, a gateway may also be a pair of nodes on each network connected by a link.

2.1.1 Flat vs. hierarchical addressing Addresses of network entities (e.g. nodes, processes etc.) are used in selecting a route for transferring messages across networks. There are two basic types of addressing structures; flat and hierarchical.

A flat address implies that it has no meaningful subparts and the routing process should be able to handle all addresses without segmenting them in parts. There are some advantages in this approach. The addressing is logically simple and has the potential to optimize the particular route between two network entities. But there are some costs: upon adding a new address one must ensure that it does not conflict with those already used, and the routing now has to be able to recognize large number of addresses.

Hierarchical addressing has two distinct advantages. It helps to simplify creation of unique addresses, and to aid in routing. Partitioning of the address space allows freedom to create and assign new addresses within a portion of the hierarchy. Partitioning also reduces the number of addresses any particular gateway has to know for routing messages.

2.1.2 Intranet routing and internet routing The issues involved in routing messages between networks are similar to the ones involved in routing messages within a single network. Hence the routing schemes used in intra-network routing can be successfully used in inter-network routing. In describing individual schemes

(section 4.0 and 5.0) we will alternate between internetwork or single network environment without any loss of generality. Basically each routing scheme is described in an environment for which it was originally proposed. Reading networks for nodes and gateways for links will help in translating intra-network routing schemes to the similar ones useful in internetwork environment.

It should be noted that this does not imply that for a given network one should use the same routing scheme for intranetwork routing as well as internetwork routing.

2.1.3 Short and long range topology changes The topology changes may of two types, long term and short term changes. Long term changes occur because of expansion plans. One may decide to add new networks and new gateways to the existing networks or one may need to eliminate or replace existing ones. These long term changes can be dealt with easily by upgrading in the field routing tables or by effecting the same by loading tables from the central process.

Short range changes involve some of the links or nodes being taken out of operation due to failure of some elements or enemy attack. Taking care of these changes in the routing scheme to ensure correct operation in a matter of seconds is a more difficult problem. Some of the known solutions to tackle the problem are described in sections 4.0 and 5.0.

2.1.4 Survivability and the Network Partitioning The survivable network interconnection is defined as the one in which changes in the connectivities of networks or failures of few networks or gateways do not affect the communication between a pair of nodes which may have remained physically connected by at least one path. To achieve survivability, one should have large enough number of alternate paths between any pair of nodes and one should also have an adaptable routing scheme. The adaptable routing scheme should be able to detect a break in a path and switch over to alternate paths between any pair of nodes without user attention.

Thus for network interconnections in a tactical environment, the survivability requires building enough alternate paths between any two networks, providing flexibility to handle changes in the topology, and not using few critical elements for the operation of the interconnection scheme.

It is possible to use interconnections to increase survivability of individual networks. To understand how this can be done, first let us define a network partition. A partition is a group of nodes which can not be reached from other nodes on the same network using a path totally within the network. If network interconnections to the partitioned network exists and they are still alive than one may be able to route a message out of the network to some other networks and back to the partition which is

unreachable otherwise. The problems associated with designing internetwork schemes which handle network partitioning are described at length in [CERF79] and [PERL79].

2.2 Goals of Interconnection schemes for Loop Networks

The interconnection schemes for loop networks should provide:

- Selection of shortest possible path in a reasonable amount of time,
- Adaptability of choosing alternate paths in case of failures,
- Minimum wastage of time in switching to an alternate route,
- Minimum wastage of bandwidth in exchanging route control information.
- Minimal packet processing for routing individual packets.
- Load shedding and switching of routes to avoid creating a bottleneck at any single gateway.
- Distributed route control to avoid having a central critical and hence vulnerable element.

The extent to which various routing schemes meet these goals is discussed in rest of this document.

3. CLASSIFICATION OF ROUTING SCHEMES

3.1 Introduction

A routing path is defined as the ordered set of links a packet has to travel to reach the destination. The routing process at any node, other than the destination node, involves selection of an outgoing link on which the packet is to be transmitted to eventually reach the destination. The selection of the next link (or hop) is based on the information contained in the packet header as well as the information about the network status. Depending on the type of routing used, the routing process may perform binding of address string to the route, and exchange control messages to keep network status information up-to-date.

Previous classification and survey work of [SHOC78], [RUDI75], [FULT71], and [CYSP80] did not easily accommodate GRIDnet [MOOR80] and some other loop switching methods. Hence we will first offer slightly differing view of categorizing various routing schemes.

3.2 Classification

Various types of routing methods used can be described within the framework of three dimensions:

(1) Packet header information related to route selection,

(2) Type of mapping used to bind an address string to a route to take,

(3) Update mechanism used to exchange current information on network status.

Fourth possible dimension for comparison can be the time at which the path is established. The path established may be permanent or temporary for the duration of a connection, session, or a message delivery period.

Packet Header Information

Route-related packet header information is one of the inputs required for selecting the next outgoing link at any intermediate node. The header may contain:

- destination address,
- explicit path information,
- source address or unique path ID, or

• appropriate combination of the above three.

3.2.1.1 Destination address based routing requires that the address-to-route binding be performed at every intermediate node. Normally (but not necessarily - see section 3.4) this is accomplished by a table look-up procedure. The method requires more processing and memory at the intermediate routing nodes than the other three methods. Due to its incremental nature of decision making, the method is also known as a hop-by-hop or incremental routing method.

3.2.1.2 Explicit path routing have each packet specify all intermediate points in a route to be taken. This simplifies routing process at the intermediate points. Intermediate points select the outgoing link depending on the next node specified in a packet. The number of nodes reachable next from any node are generally much smaller than the possible number of destination addresses. Hence this type of routing will result in reduced memory and processing at the intermediate points.

The address-to-route binding is performed here by the source node where a packet is formed. Source nodes either have to possess network topology and status information or have to request central route control facility to give a route information.

3.2.1.3 Unique path IDs This method requires association between the route and the unique ID to be known to all intermediate nodes. This association either can be predetermined (say for broadcast messages) or can be preassigned for the duration of a session or virtual connection. This routing method is useful for providing multidestination packet protocols across the networks.

3.2.1.4 Hybrid methods A packet may contain some of the intermediate node addresses but not all nodes along the route. This then is a type of routing method where source as well as intermediate nodes participate in determining the route for a packet. Similarly one can find some more routing methods which combine appropriate features of the above three methods.

Type of Mapping

Binding destination address or unique path ID to a route requires some knowledge of the network topology. Such a binding may be performed either at the source node or at any intermediate node. For example in explicit path control method source will perform the translation of destination address to the path and then form the packet. Whereas in destination address based routing method, each intermediate node in the path to determine the next hop of the path.

The bindings or address-to-route mappings are distinguished according to how the knowledge of network topology and status

is used. The type of mappings are:

- Static
- Adaptive
- Algorithmic

3.2.2.1 Static mapping uses initially provided network topological information but does not not bother to update this information to reflect current loading or transient failures of links or nodes. The method can implemented as a simple directory lookup procedure. The directory lookup is used to determine the route or next hop of the route. Since the method does not use current network topology and status information, it is useful only for networks where link or node failures are rare and survivability is not the main concern.

3.2.2.2 Adaptive mapping The routing methods which use current knowledge about the network status are better able to cope with the problems of transient link or node failures and temporary congestion or loading of parts of the network. The mapping again is effected as a simple table lookup procedure. But these tables are updated to reflect the current conditions of the network.

Due to the time lag involved in distributing network status information, it is possible to have at any time slightly different and possibly inconsistent information in various nodes. This may lead to problems of packet looping indefinitely or unnecessarily meandering through the networks. Most routing methods of this type take care to eliminate such packets by using hop counts or lifetime fields in a packet header.

This type of routing methods do ensure more survivability than the static methods but we pay for the increased survivability in terms of the bandwidth wasted in exchanging current network status between nodes.

3.2.2.3 Algorithmic mapping is defined as the type of mapping in which a route selected depends on a destination address and an algorithm which has a built-in knowledge of the topology or the addressing structure of the network. The routing method using this type of mapping at any routing node in a path will send a packet on the link on which it has the highest probability of reaching the destination eventually. This type of methods do not use any global network connectivity or status information to ensure the delivery of a packet. Instead the routing algorithm is based on the addressing structure or general constraints on the topology to ensure that a packet is eventually delivered. The routing methods based on hamming's distances [GRAH72] and that proposed for GRIDnet [MOOR80] are examples of this type of routing schemes.

These schemes do not waste any bandwidth exchanging update messages and still provide better survivability than other methods.

Update Mechanism

The update messages are sent to all nodes which require current network status information to perform routing functions. The status information may be regarding link failures, node failures, link or node coming up again, or simply reflecting loading conditions at each link or nodes. By using only the knowledge about immediately connected links or nodes in the routing method, it is possible to localize the update message transmissions.

One more aspect of the update mechanism is the source of update messages. In a centralized scheme all updates will be handled by a Network Control Center (NCC). The NCC will monitor status of the network and send update messages to routing nodes when required. In a distributed mechanism each node either periodically or due to some failures, sends an update message to nodes in the near vicinity. At the discretion of these nodes These changes may or may not be propagated further down. One more possibility is having entirely isolated node updates. In this case each node by trial and error corrects its connectivity map but does not send any updates to any other node.

The schemes using distributed updating are inherently more robust than the ones depending on central updating. The cost of each node having a global picture in distributed scheme can be quite high due to large number of updates floating in the network. Hence the ARPAnet routing scheme uses only localized exchange of update messages [McQU74].

4. SURVEY OF LOOP INTERCONNECTION SCHEMES

In this section we will consider the known proposals for dealing with interconnections of loop networks. In each subsection we first briefly describe the method. Only a central idea is presented here. The details can be obtained from the references cited. Next we evaluate the method in light of the goals discussed in section 2.4. In section 6.0 we will summarize and compare these methods.

4.1 DCS ring

Distributed Computer System (DCS) is a ring network built at the University of California, Irvine [FARB75]. It is a six node 2 megahertz ring network designed to economically connect mini and midi computers. One of the unusual feature of the scheme is the soft addressing structure used to allow easy migration of processes between host systems. Each packet contains process name (16 bit long) ID rather than actual destination node addresses. In other words packet header does not give any help in routing the message. Every packet passing by a ring interface is compared against the list of processes residing on the corresponding host and copied over if a match occurs. The comparison process is facilitated by using 16 bit wide associative memory.

A proposal to interconnect DCS rings was made by Farber and Vittal [FARB73]. The rings are allowed to be connected any general structure. There may be more than one path between any two rings. Rings are connected with a simple gateway between them. A gateway can connect only two rings but there is no restriction on the number of gateways a ring can have.

Each ring is named and the ring name is added to the process address to uniquely identify the message destination. A source node provides explicit path information using variable length address fields to list out all the gateways to be traversed on the way to the destination address.

Each message has

- Message status bits
- Source process name
- Destination process name
- List entry count
- Variable length list of addresses.

Message status bits are used to indicate acknowledgment by a destination in the current ring. No end-to-end acknowledgement

is provided but if the message is undeliverable than it is turned around back to the source node.

Each node does not keep information about where processes are located. Instead a path to a desired process is located by broadcasting a request to the process. Pathstamping of the broadcast message allows the intended destination to indicate not only the location but the desirable path in an ack sent back to the source. The same method can be used to locate an alternate path when any intermediate gateway fails.

To the best of the author's knowledge, this multiring extension for DCS ring was never implemented.

4.1.1 Discussion The protocol is quite complicated and unfortunately the interconnection protocol has not been defined in great detail. The explicit path routing can waste large number of bits in a message if thousands of rings are connected. Broadcast requests and problems arising out of required acks to these requests are not discussed in detail.

The survivability is basically provided by sending negative acknowledgements for undeliverable messages. On receiving an undeliverable message back source node can reenter the broadcast procedure to locate another path to the destination.

The broadcasting of messages to locate a path or an alternate path will waste significant amount of the available communication bandwidth.

4.2 Pierce's Hierarchical Loops

One of the earliest proposal for interconnecting loop networks was made by Pierce [PIER72]. He envisioned a hierarchy of local, regional, and national loops to provide data network services. The structure he proposed was a strict hierarchy with each local loop connecting only one regional loop and all regional loops connected only via a national loop. The structure is shown in figure 4.2-1.

Under the restriction of the strict hierarchical structure one can implement an easy scheme to transfer messages from one loop to another until it reaches the destination. Destination address (and source address) has three parts, local loop number, regional loop number, and the node number on the local loop. A gateway on the local loop transfers a message to the regional loop if the destination's local loop number does not match with that of the source. Similarly a gateway on the regional loop transfers the message to national loop if regional loop numbers do not match. A similar process takes place in removing messages from national and regional loop to ultimately reach the destination local loop.

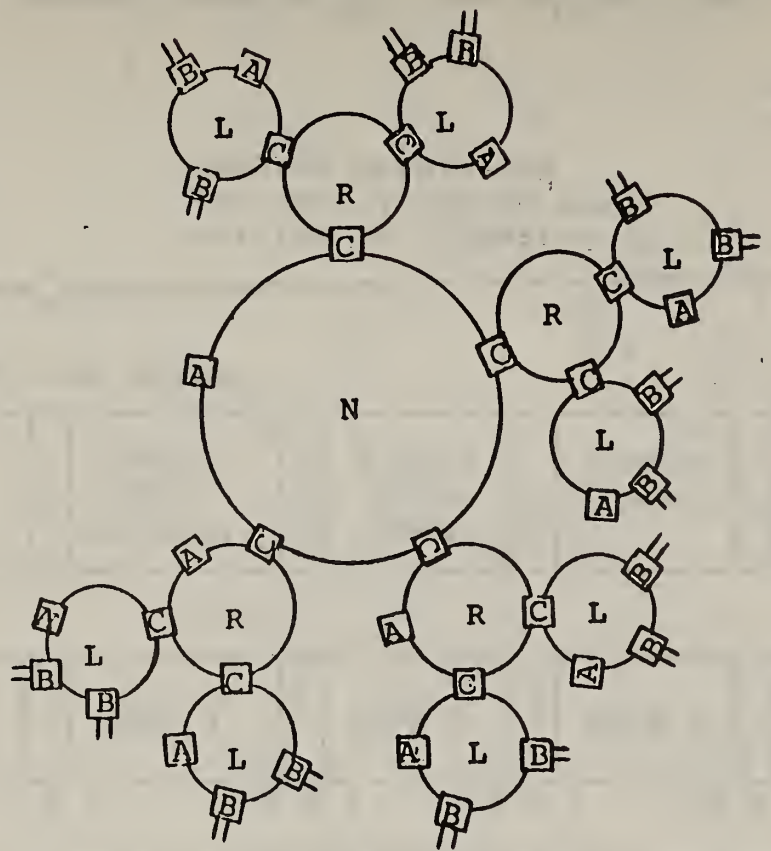
Another slightly more complicated but essentially same scheme was proposed by Kropfl [KROP72]. He extended the message format to include current loop destination and source addresses in addition to the source and destination's local and regional loop addresses (see figure 4.2-2 and 4.2-3). The scheme he proposed basically allowed him to return an undeliverable message back to the source node. Any returning message is thrown away if it has any difficulty reaching the source node.

Pierce recognized the need for providing alternate paths for reliability and providing special trunk loops to accommodate heavy traffic between two loops. He proposed special types of gateways designed to transfer traffic to the alternate loops or to trunk loops. The scheme can handle limited amount of switching at local loop level but will be difficult to use for switching between regional loops.

For experimental purposes a loop using two node interfaces was designed and built but gateways and multiring hierarchical rings were not implemented.

4.2.1 Discussion The hierarchical structure simplifies the routing procedure without wasting too many bits in the message header. The disadvantage is a restrictive nature of the network topology. Even two geographically close local networks may end up talking to each other through a regional or even national loop. This will increase the response time much more than the case where local loops are allowed interconnections. Unfortunately the scheme proposed for trunk loops is not general enough to economically handle extensive local loop switching.

The hierarchical structure is inherently a vulnerable target in hostile environment. Disabling the national loop (or regional loops) can easily isolate large number of loop networks. The provision of alternate loops can help to retain connectivity between various portions, but to make such a hierarchical network survivable one may have to build large number of alternate and bypass loops. This can be quite costly and impractical approach in providing a reliable network in a tactical environment.



N - National Loop
R - Regional Loop
L - Local Loop
A - Loop Repeater Box
B - Node Interface Bo:
C - Gateway

Figure 4.2-1 Hierarchy of Loops - Pierce

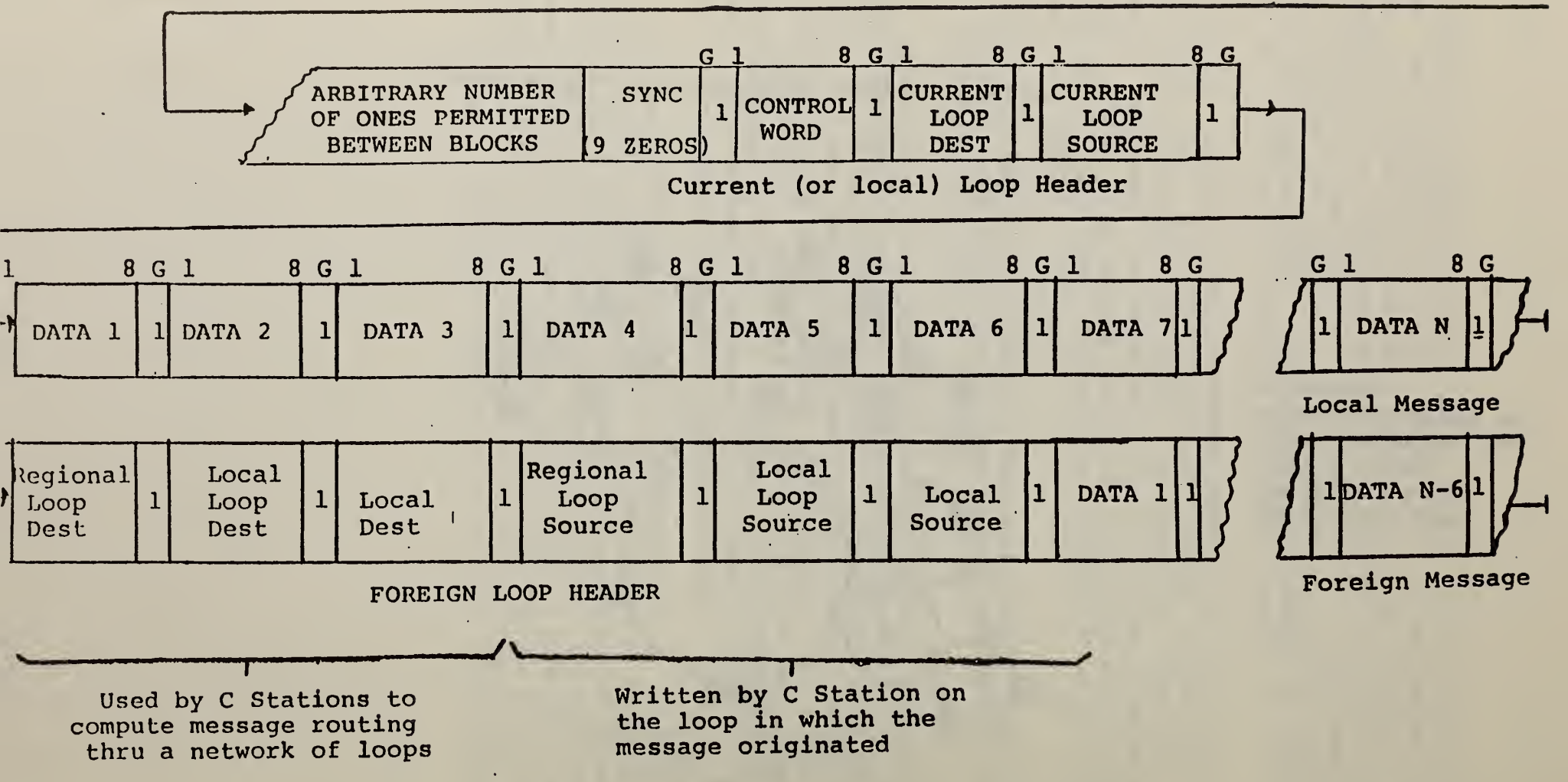
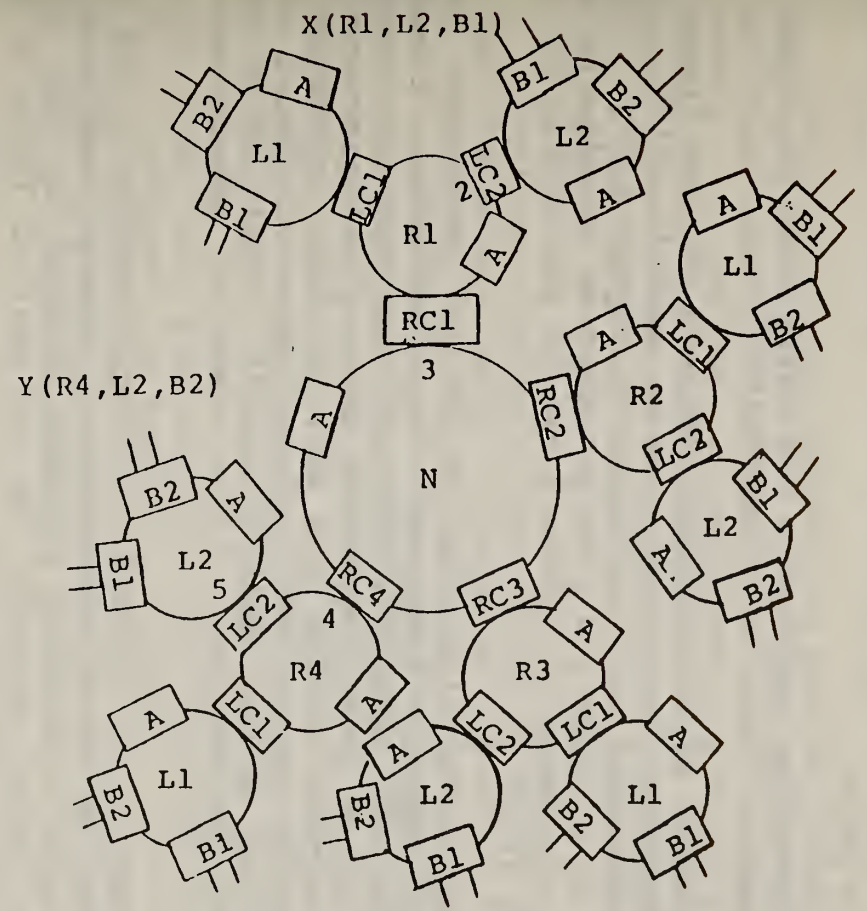


Figure 4.2-2 Message formats for hierarchical loops Kropfl's scheme.



FOREIGN MESSAGE HEADER

SYNC	DBOW	CLD	CLS	RLD	LLD	LD	RLS	LLS	LS
STEP									
1	X TO L	LC	.B1.	R4	L2	B2	---	---	---
2	L TO R	.RC.	.LC2.	R4	L2	B2	.RC1.	.LC2.	B1
3	R TO N	R4	.RC1.	R4	L2	B2	RC1	LC2	B1
4	N TO R	L2	.RC.	R4	L2	B2	RC1	LC2	B1
5	R TO L	B2	.LC.	R4	L2	B2	RC1	LC2	B1

Figure 4.2-3 An example of routing using Kropff's scheme

4.3 Bell Labs work on Loop switching

Following the pierce's work on loop networks, Graham and Pollack [GRAH71] proposed an addressing scheme which has following attractive features:

1. It permits a simple routing scheme to be used by messages to reach their destinations.

2. A message using this routing scheme will always take the shortest possible path.

3. The method of addressing applies to any collection of loops, no matter how complex their interconnections.

4. The method allows for alternate paths between any two loops.

The method is best explained by referring to an example. Let us consider six loops connected in a way shown in figure 4.3-1. If each loop is treated as a point and each gateways as an edge in a graph, than this interconnected network can be abstracted as a graph shown in figure 4.3-1.

Graham et al. devised an algorithm to assign addresses to each loop in a arbitrarily connected loop networks. These addresses are such that the Hamming's distance between any two addresses is exactly equal to the number of edges (i.e. gateways) in a shortest possible path between this pair of loop networks. The Hamming's distance is defined as the number of places in which two address strings differ. More precisely, a pair of 01 or 10 contributes one to the Hamming's distance and all others contribute zero.

In the example shown, the Hamming's distance between A and B is 2 which is equal to the actual distance between two loops. Graham et al. have proved that it is possible to find a set of addresses satisfying this property for any pair of loops in a arbitrarily connected loops.

With this type of addressing scheme the following simple routing scheme is possible. A message is forwarded by a gateway only if in doing so it is likely to reduce the Hamming's distance between the destination loop and the current loop. If several gateways do the same job then each one will lead to an equally short optimal path from sending loop to the receiving loop.

4.3.1 Discussion One of the obstacle in using this scheme is the large number of bits required for addressing loops. Loop addresses are strings with ternary value symbols. This requires that we use 2 bits per digit in the address string. One possible coding can be 0 = 00, 1= 11 and d= 01.

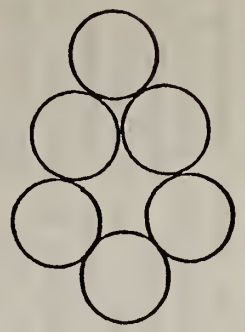
Graham and Pollack have given a construction which gave at best an addressing scheme with 2(N-1) bits per address for all

cases they tried it on. They conjectured that this is true for any arbitrary collection of loops but managed to prove this to be true only for some special cases where topology is like a tree or a cycle of path 2. The address of 2(n-1) may be acceptable when N, number of loops connected, is quite small. But if N is in the range of thousand loops than the address length will be prohibitively large. For instance if N = 1000 loops than the destination loop address alone will be 1998 bits long. In a message which has 512 bytes of text this gives at least 33% header overhead.

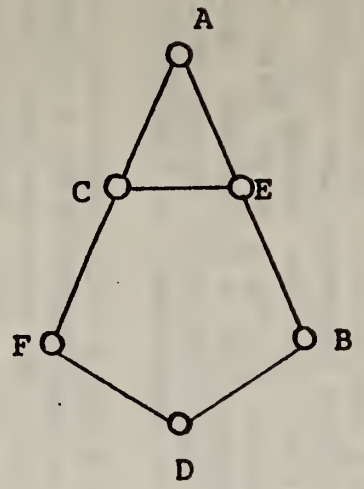
In an effort to reduce the address length required Brandenberg et al. [BRAD71] proposed an algorithm which gives more efficient coding of addresses. But the method requires large memory at gateways to implement the routing method. The implementation of the method requires table lookup procedure at gateways. Hence the method also requires updates to account for failed gateways and in general has same shortcomings as other table lookup methods.

Another problem of the scheme is a low survivability. The simple routing scheme proposed by Graham et al. is not sufficient to take care of problems that can arise when some of the links are inoperable. For instance in the example given consider a case where B wants to send a message to A. In normal operation the message will pass through E and reach A in the next hop. But if say a link between A and E is broken than a message for A has to be routed on to C even if this does not decrease the Hamming's distance. More difficulties are caused if both links AE and CE are broken. There still exists a path between B and A but to take that path a message has to be routed first to D which actually increases the Hamming's distance. The worst case where A is totally isolated can be found only after traversing at least once through all nodes reachable from B. This can be a considerable strain on the communication system.

Future growth of the network will create some more problems for this scheme. The addition of a node will require addresses to be recomputed to preserve the correspondence of hamming's distances to actual distances. This field upgrading of node addresses will lead to low maintainability of the system. Graham and Pollack [GRAH72] suggested combining of the hierarchical structure with this method to alleviate the problem of reassignment of node addresses. In this modification, each level of the hierarchy uses this method for switching at the same level but uses the scheme suggested by Pierce [PIER72] to handle switching between levels. An addition of a node, when using this method of addressing, will only require a subset of node addresses to be changed leaving other addresses unchanged. This modification will also result into reducing the number of bits required for a node address. Unfortunately even with this modification the number of bits used up in addressing are still unacceptably large for connecting thousands of loops.



with Abstract Graph



A - 1111d
B - 001dd
C - 11d0d
D - 10dd1
E - 10dd0
F - 010dd

Figure 4.3-1 Interconnection of six loops - An Example

4.4 GRIDnet scheme

The interconnection scheme designed by Ray Moore at National Bureau of Standards [MOOR80], is also based on addresses assigned to loops. In this method loop interconnections are constrained to be in a particular structure. If we represent each loop network as a point and interconnections between them as edges on the graph, the structure is constrained to look like a grid as shown in figure 4.4-1. Each loop is assigned a two-part address, level and loop number. The odd level numbers are used for horizontal loops and even level numbers are used for vertical loops. The same is true in assigning loop numbers.

The routing scheme consists of trying to reduce either the difference in level numbers or loop numbers between the current loop address and destination loop address of the the message. If a message cannot proceed to its destination along a path than it is returned one step backwards. The rerouting of a returned message is handled much in the same way an "intelligent" mouse would attempt to do. The detailed algorithm is described in [MOOR80].

4.4.1 Discussion The routing method is simple and does not use table lookup. This makes it easy to implement with a simple interface between rings. Since the method does not use any network status knowledge, it does not waste any bandwidth in exchanging status messages. The method guarantees that in normal operations it will take the shortest path possible between two loops. In case of failures in the connections, the message exhausts all possible alternate paths before giving up.

Unlike the previous method of Graham et al. the routing method uses very few bits in the message as a destination loop address. In the worst case arrangement of loops, the number of bits required are of the order of square root of N, where N is the number of loops. The algorithm does require more number of steps than just a simple Hamming's distance calculation but still can be easily computed by using a cheap microprocessor.

The disadvantage of the method stems from the wastage of bandwidth that will occur in finding an alternative path after a failure. First packet, which is not delivered, cannot avoid threading a path by "trial and error" method to succeed. But repeating the same process for other packets going between the same pair leads to large number of meandering packets. These packets will eventually reach the destination but in doing so they unnecessarily waste communication capacity. This fact can be used in devising an insidious form of attack whose sole purpose may be to significantly degrade the internetwork communications. This problem of meandering packets can be avoided by either using some form of explicit path control routing or by retaining memory of the failures at each intermediate gateway.

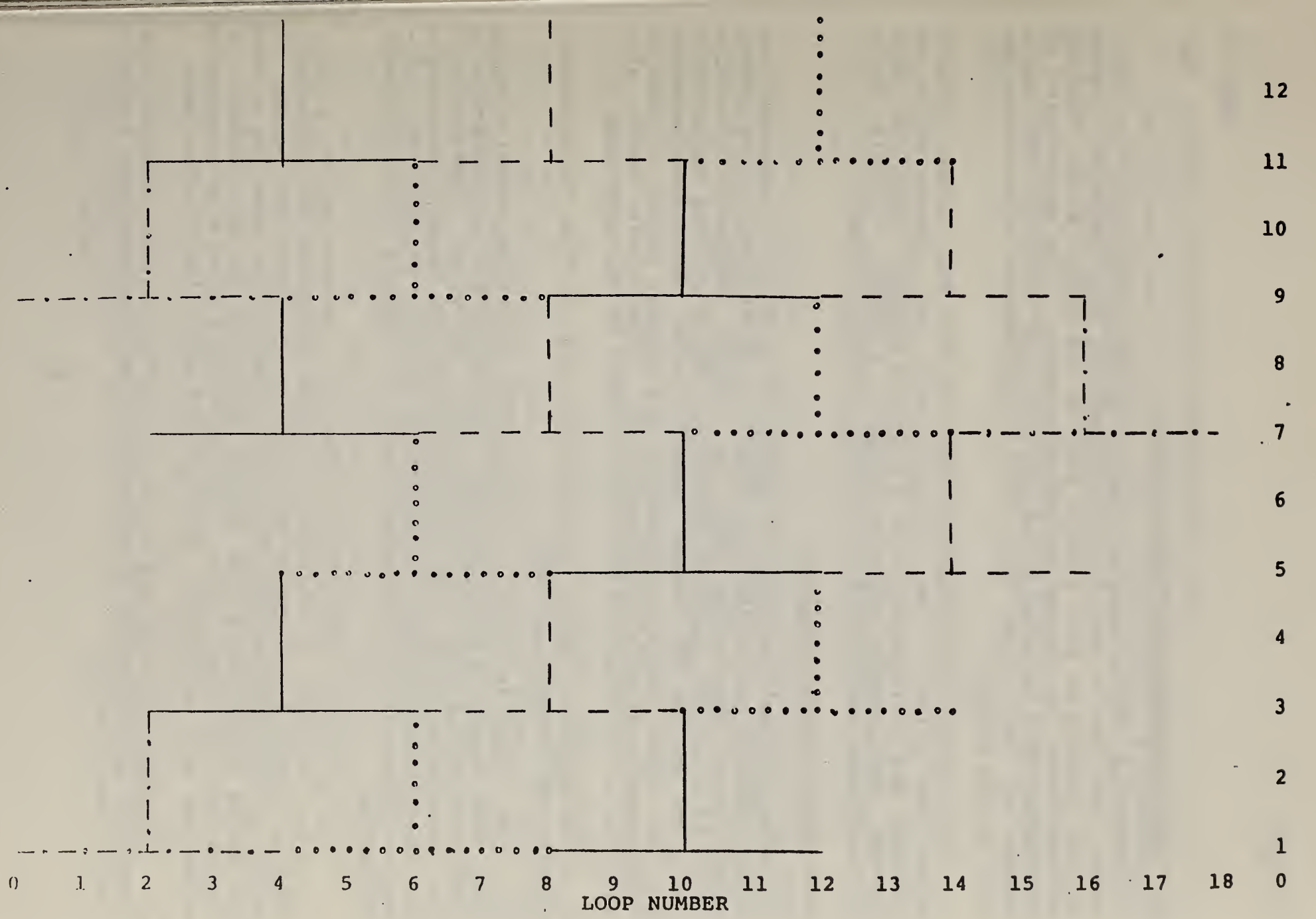


Figure 4.4-1 GRIDNET addressing scheme for interconnection of loops

5. ROUTING SCHEMES IN OTHER NETWORKS

Routing schemes used in connecting any other type of networks can also be used for interconnecting loop networks. Hence in this section we will explore some of the other known routing methods used in packet switching networks. The explanation of the method will be quite brief and will only outline the basic nature of the scheme. This will be supplemented with an evaluation of the method.

No attempt will be made to list out all the known schemes which are quite a few. Instead only the schemes which are relevant and representative of a class of methods will be discussed. One should refer to [FULT71], [CYSP80], [RUDI78] for a more complete enumeration of possible routing schemes.

5.1 Explicit Path Control

The proposal for explicit path routing made by Jueneman and Kerr [JUEN76] is a simple extension to traditional table lookup approach. At any routing node, the search argument into the table is still the destination address. But a routing table entry now gives not one but several possible paths. An index (referred as Next Node Index) contained in the message header is used to select among the possible alternatives. The table entry also provides a new value for NNI which is used at the next node.

The method is being considered for use in SNA multidomain network environment [CYSP80].

The method is basically a hybrid between incremental routing and the source routing method. The method has an advantage in that each node knows of alternate paths to other destinations. In case of a failed link or an undeliverable message the origin can be advised to select an alternate path for sending the message. This is the proposal made in [JUEN76] to handle route switching.

The alternate path selection can also be handled in a distributed fashion by addition of update procedure to the method. The idea would be to inform the previous node that the message it sent is not deliverable. This would result into changing stored NNI value with a value from another entry for the same node. If there is no other entry than the change should be propagated still one step backwards. This effectively selects a new path.

The main disadvantage of the method is that it requires more memory at routing nodes as compared to that required for a simple table lookup method.

5.2 ARPAnet's Distributed Adaptive Routing

The ARPAnet uses tables at each node to determine the next step in the route of a packet. This routing table is dynamically modified to reflect changing traffic patterns. Such a technique is known as an adaptive routing technique.

The algorithm used to determine changes to the routing table may be based on information locally available plus information from

(1) adjacent nodes only (for example queue sizes),

(2) all other nodes in the network, or

(3) primarily the adjacent nodes, plus some information on the state of the network along the route to the destination.

Global algorithms, based on information from all nodes require that either all nodes exchange information with every other node, or all nodes exchange information with a central node. The former involves a very high overhead in traffic devoted to traffic optimization techniques. The latter introduces the vulnerability of the central routing center.

The current ARPAnet algorithm [McQU74] is of type 3 above and is the best known example of distributed adaptive routing.

In the ARPAnet, the objective is to route each packet in such a way as to minimize its delay, with no regard as to how that might affect the delay of other packets. The routing tables are updated by computations that take place in each network node called interface message processors (IMPs). Each IMP periodically exchanges information with its adjacent IMPs on the delay experienced in sending messages via the adjacent IMP to each destination. This is estimated based on the number of links involved and the recent queue lengths on each link. On the basis of this information, each IMP periodically recomputes a new shortest-delay next link to use for each possible destination. The routing control messages are exchanged between any_pair of adjacent IMPs at least every 640 msec.

Gallager [GALG76] has proposed another distributed adaptive algorithm for establishing routing tables. The proposed mode is to keep the routes from each source to destination fixed over long periods of time, and to change them only when necessary to reduce the long-term average delay per packet. The objective, unlike the one in ARPAnet, is to route each packet in such a way as to minimize overall delay in the network. Accordingly, long-term average delays are the criterion and routes need not be changed frequently.

All distributed adaptive methods require some traffic overhead to achieve traffic optimization. This may be a significant percentage of the traffic when multiple failures are likely as in a tactical environment.

5.3 Perlman's Internet Routing Proposal

The scheme proposed by Perlman [PERL79] for internetwork routing is based on each gateway keeping a global knowledge about the network status. The method provides a way to detect partitioning of networks and finding alternate path for a packet to be routed to one of the partitions.

Each gateway keeps tables showing functioning gateways, networks reachable through them, and possible communication paths to these gateways. Each gateway broadcasts information to all gateways information regarding changes in the status of its connections to networks or any connecting links.

The method uses more bandwidth than that used by ARPAnet style internetwork routing method. Keeping global status at each gateway will also increase the memory required. The difficulty of ensuring consistency between network status available at any moment to various gateways is as severe as that in the ARPAnet algorithm. The only possible advantage is the ease of providing parallel paths between any two destination pairs.

5.4 Delta Routing Method

Rudin has proposed a combination of a centralized and distributed adaptive routing scheme [RUDI75]. The scheme uses a centralized route selection center which sends to each node all possible routes according to the current global load conditions. The NRC receives periodic updates form all nodes reporting network status. This information is used by NRC in selecting most appropriate routes for each node. Each node, depending on the local load conditions, will than choose the best alternative from the routes provided by NRC.

The method allows global as well as local optimization of delays in the network. The method also allows optimal route switching procedures. But the NRC is a critical element in this scheme, which reduces the survivability of the method in a tactical environment.

5.5 Source Routing in Internet Protocol

Cohen and Postel [COHE79] in a short memo proposed a source routing option, that can be provided in Internet Protocol [POST80]. The preferred method by them is to specify a partial route and a destination address in a packet. Hence, a source routed message can pass through intermediate gateways and nodes before reaching the next specified point in the route. The

gateways specified are the ones which handle source routing option and other intermediate node or gateways need not have any capability to handle this option. This has the advantage of reducing the work required to provide source routing and still force the message to pass through points which are important for:

(1) reaching obscure destinations,

(2) security, and

(3) measurements.

They have also proposed a way of stamping the path taken by a message using Construct-Return-Route option. This helps the destination of a stamped packet in replying with a source-routed message.

The method works well as long as all the specified points are working in order. But no alternative route selection method is defined in cases where any of the specified intermediate point has failed. Hence, the method has low survivability in a tactical environment.

6. SUMMARY

High survivability of interconnected loops can be achieved by having an adaptable internetwork routing scheme. The desirable characteristics are: not requiring any critical element or elements for correct functioning, having an ability to find an alternate path if there exists one, and adapting easily to long range changes in the topology.

The side-effects of such an adaptable routing on network performance depends on many factors. These factors are: number of bits required for route control in every message header, processing and memory requirements at gateways and intermediate nodes, bandwidth wasted in exchanging current network status information, optimality of path taken in normal as well as failure mode, and communication resources wasted in rerouting after detecting a failed route.

The GRIDnet routing scheme does better than other routing schemes on most counts. The scheme is adaptable, has distributed control, requires small address space, requires simple processing at gateways, does not require any significant memory space at gateways, does not need to waste any bandwidth for exchanging status information, and guarantees selection of shortest path at all times. A simulation may be helpful in confirming that the method will work for any combination of failures. Such a simulation can also be used to determine how optimal is the path being selected and how much communication resources are wasted in rerouting of messages. A critical evaluation supplemented with a simulation may be expected to help in suggesting some modifications to reduce the communication resources used in rerouting of messages.

REFERENCES

[BRAN72a] L. H. Brandenberg, B. Gopinath, and R. P. Kurshan, "On the addressing problem of loop switching", Bell System Technical Journal, 50:7, September 1972, pp. 1445-1469.

[BRAN72b] L. H. Brandenberg and B. Gopinath, "A table look up approach to loop switching", Bell System Technical Journal, 51:9, November 1972, pp. 2095-2099.

[CANT74] D. G. Cantor and M. Gerla, "Optimal routing in a packet switched computer network", IEEE Transactions on Computers, C-23, October 1974, pp. 1062-1069.

[CERF79] V. Cerf, "Internet addressing and naming in a tactical environment", Internet group documentations, IEN no. 110, August 1979.

[COHE79] D. Cohen and J. Postel, "Source Routing", IEN no. 95, USC-Information Sciences Institute, May 1979.

[CYSP78] R. J. Cysper, Communication Architecture for Distributed Systems, Systems Programming Series, Addison Wesley Inc., 1978.

[FARB73] D. J. Farber and J. J. Vittal, "Extendability considerations in the design of the Distributed Computer System (DCS)", National Telecommunications Conference (NTC 1973), Atlanta, November 1973, pp. 15E.1-15E.6.

[FARB75] D. J. Farber, "A ring network", Datamation, February 1975, pp. 44-46.

[FULT71] G. L. Fultz and L. Kleinrock, "Adaptive routing techniques for store-and-forward computer-communication networks", Proc. 1971 Int. Conf. on Communications, June 1971, Montreal, pp. 39.1-39.8.

[GALL76] R. G. Gallager, "Local routing algorithms and protocols", Proc. IEEE 1976 Communications Conf. Philadelphia, June 1976, pp. 20-17-20-20.

[GRAH71] R. L. Graham and H. O. Pollak, "On the addressing problem for loop switching", Bell System Technical Journal, 50:8, October 1971, pp. 2495-2519.

[JUEN76] R. J. Jueneman and G. S. Kerr, "Explicit path routing in communications networks", Proc. Third International Conf. on Computer Communications,

Toronto, August 1976.

* [KROP72] W. J. Kropfl, "An experimental data block switching system", Bell System Technical Journal, 51:6, July-August 1972, pp. 1147-1165.

[McQU74] J. M. Mcquillan, "Adaptive routing algorithms for distributed computer networks", Report no. 2831, Bolt Beraneck and Newman, Cambridge, Mass., May 1974. Also available from National Technical Information Service, AD 781467.

[MOOR80] R. T. Moore, Personal communications, National Bureau of Standards.

* [PERL79] R. Perlman, "Internet routing and the network partition problem", IEN no.120, Internet group's notes, October 1979.

* [PIER72] J. R. Pierce, "Network for block switching of data", Bell System Technical Journal, 51:6, July-August 1972, pp. 1133-1145.

[POST80] J. Postel, "DoD standard Internet Protocol" IEN no.128, Jan. 1980.

[RUDI75] H. Rudin, "On routing and Delta routing: Techniques for packet switching networks", Proc. IEEE 1975 Communications Conf., San Francisco, June 1975, pp. 41-20-41-24.

[SHOC78] J. F. Shoch, "Inter-network naming, addressing, and routing", IEEE Computer Society International Conf. (COMPCON Fall 78), pp.

[SUNS77] C. A. Sunshine, "Source routing in computer networks", Computer Communication Review of the ACM 7, No. 1, January 1977.

[YAO78] A. C. Yao, "On the loop switching addressing problem", SIAM Journal of Computing, 7:4, November 1978, pp. 515-523.

---

* Appendix II contains reprints of these papers.

APPENDIX -I

The references are grouped here according to the main topic covered by them.

(a) General papers on network interconnections

[CERF78]

(b) General papers on routing schemes

[CYSP80]

[FULT71]

[RUDI75]

[SHOC78]

(c) Interconnection of loop networks

[BRAN72a]

[BRAN72b]

[FARB73]

[GRAH71]

[KROP72]

[MOOR80]

[PIER72]

[YAO78]

(d) Other miscellaneous routing schemes

[CANT74]

[CERF79]

[COHE79]

[GALL76]

[JUEN76]

[McQU74]

[PERL79]

[RUDI75]

[SUNS77]

NBS-114A (REV. 2-8C)

U.S. DEPT. OF COMM. BIBLIOGRAPHIC DATA SHEET (See instructions)	1. PUBLICATION OR REPORT NO. NBSIR 80-2149	2. Performing Organ. Report No.	3. Publication Date October 1980

4. TITLE AND SUBTITLE

GRIDNET

5. AUTHOR(S) Raymond T. Moore

6. PERFORMING ORGANIZATION (If joint or other than NBS, see instructions)

NATIONAL BUREAU OF STANDARDS
DEPARTMENT OF COMMERCE
WASHINGTON, D.C. 20234

7. Contract/Grant No.

8. Type of Report & Period Covered
Interim

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (Street, City, State, ZIP)

Defense Nuclear Agency
Washington, D. C. 20305

10. SUPPLEMENTARY NOTES

☐ Document describes a computer program; SF-185, FIPS Software Summary, is attached.

11. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here)

This report describes a highly reliable and survivable digital data communication system. It is based on the multiple interconnection of dual fiber optic loops, with up to about twenty data communications stations being served by each dual loop. The dual loop configuration is called CROSSFIRE. The interconnection of many of these to form a large network is called GRIDNET. The network is highly connected and many alternate routes are available for transmitting a message between two stations located at different points on the network. The intelligence required for the control of communications and the routing of traffic is distributed among a number of data communication nodes called gateway stations. Network survival is not dependent on the survival of any node or link, but requires instead that the network not be fragmented.

12. KEY WORDS (Six to twelve entries; alphabetical order; capitalize only proper names; and separate key words by semicolons)

Communications networks; CROSSFIRE; link protocols; local survivability.

13. AVAILABILITY

☒ Unlimited

☐ For Official Distribution. Do Not Release to NTIS

☐ Order From Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402.

☒ Order From National Technical Information Service (NTIS), Springfield, VA. 22161

14. NO. OF PRINTED PAGES
73

15. Price
$7.00