

Description

The present invention relates to a distributed communications network, that is to say a communications network in which the control of data transmission is not localised but is distributed between a plurality of nodes in the network.

Hitherto, most networks have been arranged to provide either a connection-oriented service or a connectionless service. In a connection-oriented service, a circuit is established between a source node and a destination node of the network, and all the data transferred between the source node and destination node travels via the links and nodes incorporated in the circuit. The circuit is released only when the call is terminated. Such an arrangement has the advantage that all the data arrives at the destination node in the same order as it is transmitted from the source node.

In a connectionless service, data is transmitted between a source node and a destination node via data packets (generally referred to as datagrams) which each include the address of the destination node and travel along one or more paths through the network to the destination node, the paths chosen generally depending upon the conditions in the network. Accordingly the datagrams from the source node can reach the destination node via a multiplicity of paths through the network, each having different delays so that the order in which the datagrams are received at the destination node does not necessarily correspond to the order in which the datagrams are transmitted from the source node. Accordingly, facilities are normally required to re-order the data received by a destination node of such a network.

For a general discussion of computer networks, reference is made to Tanenbaum "Computer Networks" 2nd Edn. (1989) Prentice-Hall Inc., particularly the OSI reference model shown in Fig 1-7 (p.15). In this model the network is conceptually divided into seven layers. The following description relates mainly to layer 3, the network layer.

Some of the more important networks are summarised below: TYMNET is a commercial value-added network operated by BT Tymnet Inc of California. It was set up in 1971 and currently has approximately 1000 nodes in the USA and Europe. Routes in TYMNET are computed by a central Supervisor node, which has its own representation of the complete network. A route computation is carried out every time a circuit is needed. A path is set up using a NEEDLE packet, which knows its route when it is generated by the Supervisor. This packet travels to the source of the call and then travels to the destination node, setting up a circuit as it goes. The circuit is dismantled using a ZAPPER packet.

Packets are forwarded over the constructed circuit by using logical channel associations set up by the NEEDLE packet. In TYMNET, link costs are assigned by the Supervisor according to the link capacity, transmission facilities, the type of sessions, and so on. To help with this, each node periodically sends status reports to the Supervisor. In TYMNET the link cost function is based partially on the link load; when the load changes sufficiently the Supervisor is informed. On failure all nodes incident to the failure inform the Supervisor, all circuits using failed links are dismantled, and the Supervisor then generates fresh circuits and creates them using NEEDLES.

Thus TYMNET is a centralised communications network which is connection-oriented. In particular, it should be noted that both the NEEDLE and the ZAPPER packets travel in predetermined circuits and are therefore not classified as datagrams.

ARPANET (Advanced Research Projects Agency Network) was the first network in the USA to use packet-switching technology. It was set up as an experimental network to investigate system resource sharing in 1969, and currently connects approximately 60 nodes in the USA and Europe. Routes are generated in ARPANET in a distributed fashion. Each node maintains a network database, which includes a complete network topology and the link costs. Each node broadcasts the link costs of its incident links to all other nodes periodically. ARPANET uses datagrams, so that circuits do not have to be constructed. Each node calculates a Routing Table (RT) from its delay table and topology database; this RT contains the best outgoing link to all other nodes in the network. Packet routing in ARPANET is datagram-oriented. An end-to-end protocol is used to ensure that packets are received at the destination in the correct order.

The link cost is the average packet delay suffered by a packet over the link during a 10-second interval. Thus packet delay includes queuing and processing. In ARPANET the link cost function is based partially on the link load; when the load changes sufficiently all other nodes in the network are informed. Also, since the routing decision is made at the nodes without concern for the session, packets tend to be steered towards areas of low congestion. Network component failure triggers a link cost update and hence a route generation update; this is essentially no different from other link cost changes apart from the topology changing. When a link fails, all packets queued for that link are discarded.

Thus ARPANET is a connectionless distributed data communications network.

DECnet is any network which uses the Digital Network Architecture developed by the Digital Equipment

Corporation. In DECnet routes are computed in a distributed fashion. Whenever a link cost changes, the information ripples out through the network. The link costs are based on the full capacity of the link, and are assigned by the system manager. They are updated only when the transmission facilities change status, e.g. failure. Routing in DECnet is datagram-oriented, so the transport layer orders the packets at the destination. Network component failure triggers a link cost update and hence a route generation update; this is essentially no different from other link costs changes apart from the topology changing. When a link fails, all packets queued for that link are discarded.

Thus the DECnet network is entirely connectionless (i.e. only datagrams are used) and in most respects is distributed, although the link costs are assigned by a central network manager node. MERIT is a small network, intended to connect the computers belonging to the universities and colleges in the state of Michigan, USA. In MERIT routes are generated in a distributed fashion.

Whenever a link cost changes, information ripples out through the network. This information is in the form of NETCHANGE messages sent between neighbours; the receiving node then updates its tables and may itself send NETCHANGE messages. The link cost in MERIT is based on hop-count only.

Failure may cause the hop-count in a NETCHANGE message to be increased to the number of nodes in the network, in which case it is assumed that the appropriate node is unreachable. Routing in MERIT is datagram-oriented.

Thus the MERIT network is a distributed, connectionless network. It will be noted that each of the above networks is either a centralised, purely connection-oriented network or is a distributed, connectionless (i.e. datagram-oriented) network.

In US-A-4 736 363 (Aubin et al) there is disclosed a path oriented routing system for a packet switching network in which each node has a list of the other nodes and an associated list of the outgoing trunk and path ID corresponding to the "shortest" route, where this term means smallest time delay through the network to the destination node.

In the article "A 'Near-Optimum' Multiple Path Routing Algorithm for Spaced-Based SDI Networks" by J.B. Cain et al., 1987 IEEE Military Communications Conference Milcom '87, Washington, October 1987, vol. 2, pages 578-584, IEEE, New York, US, there is disclosed a system in which a call set-up packet is transmitted from a source node to a destination node, the path on which this packet travels becoming the path for that data transmission (a virtual circuit). In general, there are multiple output links from a node and the choice of link is basically random but with a bias determined by the expected capacity required by the new call, the excess capacity on the available links and the amount of capacity on each link already dedicated to virtual circuit service.

In the article "Routing Strategies in Computer Networks" by W. -N. Hsieh et al., Computer, vol. 17, No. 6, June 1984, pages 46-56, IEEE, Long Beach, California, US, there are discussed and compared the routing functions of ARPANET, TYMNET, TELENET, DATAPAK, IBM's SNA, and Digital Equipment Corporation's DNA. Only TYMNET, TELENET and SNA use session-dependent routing to forward packets to the destination nodes. TYMNET, as mentioned herein, is not a distributed network, and route generation and updating is performed by the Supervisor. TELENET is similar in that the routing tables are generated by a network control centre (NCC) using a link cost function based on the link capacity and downloaded to the nodes. The routing tables remain static until the next system reconfiguration performed by the NCC.

In SNA, at the time of system configuration, up to eight explicit routes are defined for each node to each of its destination nodes, a number of virtual routes are mapped onto the explicit routes, and the ordering of the virtual routes is performed.

In both TELENET and SNA the central generation of the routing tables for the nodes involves a set of route selection criteria such as hop count, packet delay, link capacity, route availability. As compared with other systems which derive, from the full routing table of route delay ("distance") for each destination node and each outgoing link, a table of the link corresponding to the "shortest" route for each destination node, TELENET nodes have a routing table of destination nodes against each outgoing link in which the links corresponding to the two "shortest" routes are indicated as Primary and all other links as Secondary, and the node chooses between the two Primary links on the basis of lesser utilisation. SNA nodes contain an ordered list of the virtual routes on each explicit route, these having been established at system configuration, and packets of a new session are sent on the first available virtual route.

A distributed connection-oriented network is known from the article "Routing Techniques and Their Impact on the Capacity Planning Procedure in DATEX-P", K.E. Nüssler et al., Third International Network Planning Symposium, Tarpon Springs, Florida, June 1986, pages 94-97, IEEE, New York, US, in which a solution to "ping-pong" looping (further discussed later) involves generating a routing table update (RTU) message at a node from its updated table for dissemination to its neighbouring nodes. The RTU normally contains a list of the minimum delays of other nodes in the networks to that node, so that the neighbouring nodes can in turn

update their tables. However, it is not permitted to send an RTU to a neighbouring node via a trunk group if that trunk group is specified in the RTU as part of the minimum delay route to a node via that trunk group.

According to the present invention there is provided a distributed communications network comprising a plurality of nodes (N) which are each capable of acting as source (N_S), sub-destination (N_B , N_C) and destination (N_D) nodes, the nodes being interconnected by links (L) and each node having local processing and memory means (M), each node including in its memory means a stored list of destination nodes in association with a first stored list whose entries directly or indirectly represent links to its neighbouring nodes, each destination node being associated with a respective entry in said first stored list, the processing means being arranged to send a datagram between a source node and a destination node via one or more successive sub-destination nodes in response to the source node calling a remote destination node, each sub-destination node being determined (50) by the entry selected from said first stored list in the source node or, as the case may be, in a preceding sub-destination node, the network being characterised in that each node also includes a second stored list whose entries directly or indirectly represent links to its neighbouring nodes, the datagram is arranged to designate at least some of said sub-destination nodes as circuit nodes and to store circuit data in the memory means of the circuit nodes, said circuit data establishing (100) a temporary circuit for data transmission between the source and destination nodes via said circuit nodes, each node (N_E , N_F) traversed by said datagram which is not the source node (N_S), the destination node (N_D) or a circuit node (N_B , N_C) is determined (60) by selecting that entry from said second stored list in the immediately preceding node which is associated with the next designated circuit node in the path of the datagram, and the criteria for determining the entries for said first and second stored lists in each node are distinct and biased towards optimum transmission of said data in said temporary circuit and optimum establishment of said temporary circuit, respectively.

In use, when a given node acts as a source node and calls a remote destination node in the network, a datagram is generated in the source node and the first list in the source node is searched until the entry for the called destination node is found. This entry may be one of the links radiating from the source node (or alternatively one of the neighbouring nodes of the source node since there is a one-to-one correspondence between the radiating links and neighbouring nodes). The datagram then knows which link to traverse, and when it reaches the node at the end of the link it looks up the entry for the called destination node in that node's first list and finds the next link to traverse. This procedure is continued until the destination is reached by the datagram. In a typical embodiment the criterion for determining the entries in each first list is that the entries are those which minimise the hop count between the source and destination nodes.

Thus the information required to guide the datagram from the source node to the destination node is not located in a single node (at the beginning of its journey, the datagram does not know the path it will take) but is distributed throughout the nodes in the network. In a simple embodiment, each sub-destination node traversed by the datagram is designated as a circuit node and accordingly the circuit established by the datagram between the source and destination node is simply the path travelled by the datagram through the network. However this is not preferred, since it is now envisaged that the optimum criteria for guiding datagrams through the network are different from the optimum criteria for selecting circuits. In a simple embodiment, the nodes could each incorporate a permanent list showing the best link or neighbouring node from the node in question to various destination nodes which are selectable by the datagram. However in preferred embodiments which are described below, the lists are updated in response to changes in the network. It will be seen that the communications network of the present invention combines the advantage of a connection-oriented service (in which data is received in the same order in which it is transmitted) with the simplicity of a distributed arrangement, which does not require an expensive central control computer.

A preferred embodiment is shown schematically in Figure 1. The network shown in Figure 1 comprises, eight nodes N_S which are linked by communications links L. Each link L may for example be an optic fibre, a twisted pair of electrical conductors or even a satellite link. The precise nature of the links is immaterial.

Each node N incorporates a memory M, only one of which is shown for the sake of clarity, namely the memory of node N_S .

It will be seen that memory M incorporates a list of each possible destination node which can be reached from node N_S . In association with this list of nodes, there is provided a first list (LIST 1), each entry in this list being one of the two links extending from node N_S . Each list entry in LIST 1 is the link which minimises the hop count from node N_S to that destination node which is entered at the corresponding position in the list of nodes. Furthermore, memory M incorporates a second list (LIST 2), each entry of which is one of the links L leading from node N_S .

Each list entry of LIST 2 is that link which minimises the transmission time for data sent from node N_S to the destination node at the corresponding entry in the list of nodes.

It is assumed that node N_S calls node N_D . Accordingly a datagram is generated (which will subsequently be referred to as a BUILD datagram) and the entry in LIST 1 adjacent node N_D in the list of nodes is examined

to find that link which will enable the BUILD datagram to construct a circuit to node N_D having the minimum hop count. It will be noted that the minimum hop count between N_S and N_D is 3, representing a circuit via node N_B .

This node is the next node in the circuit but before the datagram is actually sent over the link to N_B , LIST 2 is examined to check that the direct link to N_B is the route with the shortest delay.

It is in fact the only route to N_B , and so the datagram is transmitted from node N_S to node N_B via the link L which joins them.

Node N_B then looks up its own first and second lists in its associated memory means (not shown) and finds from its own LIST 1 that the node which will minimise hop count in the circuit to node N_D is node N_C . It then looks up its LIST 2 (not shown) and finds that the quickest route for sending the datagram to node N_C is not the direct link from node N_B to node N_C but the link to intermediate node N_E . Accordingly, the BUILD datagram is sent to node N_E but this subdestination node is not designated as a circuit node. Since node N_E is not designated as a circuit node, it does not need to examine its LIST 1 in its associated memory means (not shown) but merely consults its LIST 2 (not shown) and finds that the quickest route to node N_C is via the direct link L to that node. Accordingly the BUILD datagram is transmitted from node N_E to node N_C . Node N_C is designated by the datagram as a circuit node and accordingly looks up its LIST 1 to find the link L which will minimise the hop count of the circuit to node N_D . It finds that there is a direct link to node N_D and it then looks up its LIST 2 to find the quickest route for sending the BUILD datagram to node N_D . As a result, node N_F is selected and the BUILD datagram is sent to this node but does not designate it a circuit node. Node N_F accordingly merely looks up its LIST 2 and finds that the link with the least delay to node N_D is the direct link joining node N_F to node N_D . Accordingly the datagram is forwarded to node N_D and the latter is designated as the destination node.

It will be seen that the path D traversed by the datagram (and shown in chain dotted lines) is different from the circuit C (shown in dashed lines) from the source node N_S to destination node N_D .

However, the time taken for the datagram to travel over path D is less than the time it would have taken to travel over route C . In general, the circuit path C and datagram path D will be determined by the criteria used to generate LISTs 1 and 2 for each node.

It is advantageous to employ a minimum hop count criterion for establishing the circuit because during data transmission, loading of the network is minimised and the data-processing capacity of the nodes in the network is utilised most efficiently. It should be noted that in practice, a given node may be common to two or more independent data circuits and accordingly, the node will be required to process the incoming data and guide it to the appropriate links.

Such processing will inevitably result in some delay and accordingly in some cases the transmission delay in a circuit between a source node and a destination node will depend mainly on the number of nodes in the circuit, rather than the delay due to the links themselves. Even when this is not the case, the links L will have a finite data-carrying capacity and in general, the data-carrying capacity of the network as a whole will be utilised most efficiently if each circuit employs the minimum possible number of links, i.e. the hop count of each circuit is minimised.

In general, the transmission of data through the circuit established by the BUILD datagram will be optimised by minimising the loading of the network, i.e. by minimising the hop count of the circuit.

The datagrams will not overload the network because the time taken for a datagram to establish a circuit will be much less than the duration of data transmission through the circuit. Instead, the optimum path for a datagram will be that path which minimises the delay in establishing the circuit.

Accordingly, in preferred embodiments the BUILD datagram is guided at each node to that outgoing link which will result in minimum delay to the next circuit node. This delay may for example be calculated and stored in ROM, but in preferred embodiments each node is arranged to measure periodically the delay to each of its neighbouring nodes by transmitting a timing packet to them and waiting for its neighbouring node to return the timing packet. This is illustrated in Figure 1 where node N_A is shown transmitting a timing packet to node N_E which returns it as soon as it has identified it as a timing packet. Assuming that the total delay (including the processing time at node N_E) is $2d$, then the delay recorded by node N_A for the link to node N_E is d . It will be noted that in general, the transmission delay between adjacent nodes will vary in dependence upon the conditions in the network, since the time taken for the processor at any given node to process any incoming message such as a timing packet will depend upon the loading of that processor by other incoming data.

As will subsequently be described, in preferred embodiments the nodes of network are arranged to transmit both delay and hop count information to distant nodes in the network. Changes in hop count between nodes may arise as a result of changes in the topology of the network as a result of addition of nodes or the failure of links, for example.

In a preferred embodiment, each node is arranged to update its first list in response to hop count information received from a neighbouring node which relates to the hop count between that neighbouring node and

a distant node in the network, each node being arranged to increment the hop count in response to receiving such hop count information from a neighbouring node and to communicate the incremented hop count to its other neighbouring node (if any). Preferably, each node is arranged to update its second list in response to delay information received from a neighbouring node which relates to the transmission delay between that neighbouring node and a remote node in the network, each node being arranged to add to said transmission delay, the additional delay associated with receiving the delay information over the link from the neighbouring node, to transmit the resulting total delay information to its other neighbouring nodes (if any), and to measure said additional delay by transmitting and receiving a timing packet over the link via which the delay information was received.

In the above embodiments in which the first and second lists maintained by each node are updated by delay, hop count or other information which is relayed from distant nodes, a problem known as "ping-pong looping" can occur this problem arises as a result of inconsistent information held in the lists of respective neighbouring nodes and involves the indefinite exchange of datagrams between adjacent nodes. For example, if in the network of Figure 1 the node N_E considers that the quickest route for transmitting a BUILD datagram to the next circuit node N_C is via the link L to node N_B then the node N_E will return the BUILD datagram to node N_B which in turn will retransmit the BUILD datagram to node N_E .

In principle, such "ping-ponging" could occur indefinitely and prevent establishment of the data circuit. Inconsistencies in the information available to neighbouring nodes can arise for various reasons, e.g. due to link failure which affects the updating of the information in the lists of certain nodes but not others.

This problem has been analysed by Shin and Chen in IEEE Transactions on Computers Volume C 36 No. 2 February 1987 pp 129-137 "Performance Analysis of Distributed Routing strategies free of ping-pong type looping". In this paper Shin and Chen consider the ARPANET network which is a connectionless network. Shin and Chen propose the following solution to the problem of "ping-ponging" in the ARPANET network:

"If the routing message is passed from N_j to N_i which is the second node in the optimal path from N_j to some destination node N_d we replace the delay of the optimal path from N_j to N_d with the delay of its second optimal path (i.e. the path requiring the second shortest delay to N_d among all paths in the network delay table of N_j) in the routing message passed to N_i ."

It has now been found that in preferred embodiments of the present invention in which each node is arranged to communicate changes in its said first and/or second lists to selective neighbouring nodes in dependence upon the updated entries in its said first and/or second lists, the problem of "ping-ponging" can be avoided by ensuring that each amended link/neighbouring node entry is communicated only via those links to neighbouring nodes which do not correspond to that entry. For example, if in the embodiment of Figure 1 node N_B was informed that its link L to node N_C would minimise the hop count in a circuit to N_D (so that this link should be substituted in its LIST 1 (not shown) adjacent " N_D " in the list of nodes), then the information that LIST 1 of node N_B had been updated would be communicated to node N_C not via the direct link to that node but by an indirect route. This is a particularly significant feature of a preferred embodiment of the invention but is not strictly essential to any embodiment, since the problem of "ping-pong-type looping" can be alleviated in other ways. For example, a time limit may be set for the establishment of a circuit, and the source node may be arranged to attempt to establish an alternative circuit if an initial attempt fails as a result of ping-pong type looping or for some other reason.

Further preferred features of the invention are defined in the dependent claims.

Preferred embodiments of the invention are described below by way of example only with reference to Figures 1 to 4 of the accompanying drawings of which:

Figure 1 is a schematic representation of a network in accordance with the present invention (and has already been referred to),

Figure 2 is a flow diagram illustrating the establishment of a data circuit in a network in accordance with the invention,

Figure 3 is a flow diagram illustrating the transmission of updating information between nodes in a network in accordance with the present invention, and

Figure 4 is a schematic representation of a network in accordance with the invention, showing the effects of changes in the network topology during operation.

The flow diagram of Figure 2 applies both to the network of Figure 1 and the network of Figure 4. However the latter network is more sophisticated than the former in that each of its nodes incorporates third and fourth lists of links to neighbouring nodes, each entry in the third list representing second best hop count and each entry in the fourth list representing second best transmission time. As is described below in detail with reference to Figure 3, in the network of Figure 4, each node (N) is arranged to communicate changes in its said first, second, third and fourth lists to selected neighbouring nodes in dependence on the updated entries in its said first, second, third and fourth lists, each amended link/neighbouring node entry of the first and fourth

lists being communicated via that link (L) to a neighbouring node which is the associated entry in the said second list and each amended link/neighbouring node entry of the second and third lists being communicated via that link (L) to a neighbouring node which is the associated entry in the said first list.

Thus the flow diagram shown in Figure 3, which shows how changes in the lists of each node are communicated to neighbouring nodes, applies only to the network of Figure 4 and not to the network shown in Figure 1.

Briefly, the way in which the LISTs 1 and 2 of each node in the network of Figure 1 are updated as follows. As previously noted, each node in the network periodically measures the transmission time to its neighbouring nodes and updates its LIST 2 accordingly.

Similarly, each node periodically informs its neighbouring nodes of its continued existence so that every node knows which local nodes can be reached in a single hop. If as a result of monitoring delay and hop count in this fashion, a node changes any of the entries in its LIST 1 or LIST 2 then these changes are communicated over those links to neighbouring nodes which do not correspond to any updated entries. For example if node N_B in Figure 1 discovered that the transmission time to node N_S had changed, it would update its delay entry in LIST 2 for node N_S and communicate this information over the links L to N_E and N_C . When a node receives such updating information, it amends its own LIST 1 and LIST 2 for consistency with this information, by adding the delay due to the link over which the information was received to the delay information and inserting the resultant delay in its LIST 2 and incrementing the received hop count by 1 and substituting the resultant hop count at the appropriate position in its LIST 1. For example, if node N_E learns from node N_B that the delay between node N_B and node N_S has changed to d' and the delay over the link from node N_B to node N_E is d'' , then node N_E will know that the minimum transmission time between itself and node N_S is d' plus d'' and will insert the value " $d' + d''$ " in its LIST 2 at the entry adjacent node N_S in its memory. This updating information will then be transmitted to the nodes N_A and N_C which will process the information similarly, and in this manner each node will eventually learn the shortest hop count and the minimum transmission delay to every other node in the network.

The establishment of a circuit in the network of Figure 1 will now be described in detail with reference to Figure 2. The procedure begins at step 10, in which a call is initiated from node N_S . In step 20, N_S generates a call identifier (which will identify the circuit established between node N_S and node N_D) and a BUILD datagram which includes the call identifier and also the node addresses of nodes S and D. The BUILD datagram is described below in further detail.

In step 30, the node N_S is designated as a circuit node and a connection structure is stored in the memory of the node. The connection structure includes the call identifier and also identifies the next downstream circuit node.

In order to guide the BUILD datagram to the next circuit node, in step 50 node N_S looks up its LIST 1 to find that link which will minimise the hop count to node D. The link to node N_B is chosen.

The address of node N_D is included in the BUILD datagram and node N_B is nominated as the next circuit node. The link to this node is stored as the downstream link in the connection structure stored in node N_S .

In step 60, node N_S looks up its LIST 2 to find the link with the best delay to the nominated node N_B and finds that the best link is the link leading directly from node N_S to N_B . Accordingly the BUILD datagram is sent over that link to node N_B .

In step 70 the processor in node N_B reads the datagram and determines whether node N_B is nominated as a circuit node by the datagram. In this case it is nominated, but if not, step 60 would be repeated and the datagram sent to the next node as determined by its LIST 2.

In step 80, the call identifier of the BUILD datagram is compared with the call identifiers of any connection structures already stored in node N_B . Each connection structure corresponds to a different circuit which is identified by a unique call identifier.

If it is found that the call identifier of the BUILD datagram is already stored (step 90), then this is due either to a duplicate BUILD datagram already received at that node or alternatively to a loop in the circuit. The node determines which of these two possibilities has occurred by inspecting the upstream link in the stored connection structure containing the duplicate call identifier (step 120). If the upstream link leads to the circuit node from which the BUILD datagram was sent then the datagram is a duplicate BUILD and is discarded (step 140). Otherwise the BUILD datagram is attempting to form a loop in the circuit, so it is converted to a soft DISMANTLE datagram and sent back to the previous circuit node (step 130), i.e. node N_S in the case of Figure 1. The function of the soft DISMANTLE datagram is described in more detail below, but briefly the function of this datagram is to cause the previous circuit node to make a further attempt to establish a circuit.

The BUILD datagram is only duplicated as a side effect of special circumstances involving link or node failure. For example, suppose that node N_S has nominated node N_F (Figure 1) to be the next circuit node, and has sent a BUILD datagram via node N_B . Node N_B would have sent an acknowledgment to node N_S and would

have relayed the BUILD datagram to node N_C . Such an acknowledgment need not be described in detail. Now suppose that the acknowledgment is lost and the link between N_S and N_B fails.

Node N_S has no way of knowing whether the BUILD datagram was successfully transmitted or not, and so transmits another BUILD datagram. In this manner, two copies of the same BUILD datagram will arrive at node N_F .

Accordingly, this possibility is catered for by steps 120 and 140 in the flow diagram of Figure 2.

Returning now to step 90 in Figure 2, it will be assumed that the call identifier of the datagram received at node N_B is not already stored in a connection structure in node N_B . Node N_B then checks whether it is nominated as a destination node by the datagram (step 100). Since it is not nominated, the upstream link, namely the link between node N_B and node N_S is incorporated in the BUILD datagram (step 40) and a connection structure is generated in node N_B (step 30) which includes this upstream link. In step 50, node N_B finds that node N_C is the next circuit node and in step 60, the BUILD datagram is sent over the link to node N_E en-route to node N_C . In step 70, node N_E determines that it is not nominated as a circuit node and accordingly the algorithm returns to step 60 and sends the BUILD datagram over the link to node N_C .

The algorithm then proceeds through steps 80, 90, 100, 40 and 30 and in step 50, node N_C determines that the next circuit node is N_D . In step 60, it finds that the route with the least delay to node N_D is via node N_F . In step 70, the algorithm determines that node N_F is not nominated as a circuit node and accordingly the algorithm returns to step 60 and sends the BUILD datagram to node N_D . The node N_D then implements steps 70, 80, 90 and 100 of the algorithm and since it is indeed the destination node, sends an ACKNOWLEDGE datagram back through the circuit comprising nodes N_D , N_C , N_B , N_S . The function of this datagram is simply to inform the caller at node N_S that a circuit has been established to node N_D , so that data transmission from node N_S to node N_D can begin. Unless conditions in the network have changed, the ACKNOWLEDGE datagram will be guided back through the network over the same path as was traversed by the BUILD datagram.

Accordingly, it is not necessary to describe the transmission of this datagram in detail.

The representation of the algorithm as given in Figure 2 is somewhat idealised in that it is assumed that the links L have sufficient capacity to enable a circuit to be established between node N_S and node N_D or any other pair of source and destination nodes in the network. In practice, this is not always possible and accordingly the network provides for three types of dismantle datagrams, namely SOFT DISMANTLE, EXTENDED SOFT DISMANTLE and HARD DISMANTLE datagrams. The various datagrams incorporate the appropriate information required to process connection structures in the nodes as follows:

BUILD

CALL IDENTIFIER

SOURCE NODE ADDRESS

DESTINATION NODE ADDRESS

SUBDESTINATION NODE ADDRESS

PREVIOUS NODE ADDRESS

ACKNOWLEDGE

CALL IDENTIFIER

SOURCE NODE ADDRESS

DESTINATION NODE ADDRESS

SUBDESTINATION NODE ADDRESS

PREVIOUS NODE ADDRESS

	TIME OF TRANSMISSION	TIME OF TRANSMISSION
	RE-LIST	TIMESLOTS FOR DATAGRAM
5	TIMESLOTS FOR DATAGRAM	HOP COUNT
	HOP COUNT	
10	SOFT DISMANTLE	EXTENDED SOFT DISMANTLE
	CALL IDENTIFIER	CALL IDENTIFIER
	SOURCE NODE ADDRESS	SOURCE NODE ADDRESS
	DESTINATION NODE ADDRESS	DESTINATION NODE ADDRESS
15	SUBDESTINATION NODE ADDRESS	SUBDESTINATION NODE ADDRESS
	PREVIOUS NODE ADDRESS	PREVIOUS NODE ADDRESS
	TIME OF TRANSMISSION	TIME OF TRANSMISSION
20	TIME SLOTS FOR DATAGRAM	TIME SLOTS FOR DATAGRAM
	HOP COUNT	HOP COUNT
	TRY AGAIN FLAG (TRUE OR FALSE)	
25	HARD DISMANTLE	
	CALL IDENTIFIER	
	SOURCE NODE ADDRESS	
30	DESTINATION NODE ADDRESS	
	SUBDESTINATION NODE ADDRESS	
	TIME OF TRANSMISSION	
35	TIME SLOTS FOR DATAGRAM	
	HOP COUNT	

It will be seen that each datagram incorporates the following information besides an indication of its category:

CALL IDENTIFIER
 SOURCE NODE ADDRESS
 DESTINATION NODE ADDRESS
 SUB DESTINATION NODE ADDRESS
 PREVIOUS NODE ADDRESS
 TIME OF TRANSMISSION
 TIME SLOTS FOR DATAGRAM
 HOP COUNT

The call identifier, source node address and destination node address do not require any further explanation. The subdestination node address is the address of the next node in the circuit. The time of transmission is utilised to enable datagrams to be ignored if they have been in existence for so long that their function must have been completed if it is ever going to be completed. The identification of the time slots for the datagram is required by another layer of the network and need not be described in detail.

The hop count is simply the number of links traversed by the datagram since it was generated. If the hop count exceeds a predetermined limit (which must be less than or equal to the total number of nodes in the network) the nodes may be programmed to ignore the datagram on the grounds that it has got lost in the network.

The function of the BUILD datagram has already been described in detail. The function of the remaining ACKNOWLEDGE, SOFT DISMANTLE, EXTENDED SOFT DISMANTLE and HARD DISMANTLE datagrams

is as follows.

When a BUILD datagram arrives at its Destination the circuit has been set up. The BUILD is converted to an ACKNOWLEDGE datagram to be sent back to the source node as an end-to-end acknowledgement.

Whenever a source node transmits a BUILD datagram an ACKNOWLEDGE Timeout is set. When the associated ACKNOWLEDGE datagram is received the Timeout is cleared. The ACKNOWLEDGE Timeout is associated with the corresponding Connection Structure (and hence Call Identifier) so that the correct ACKNOWLEDGE datagram can be recognised. Any ACKNOWLEDGE datagram containing a Call Identifier which does not correspond to a Call Identifier in any of the stored Connection Structures causes a HARD DISMANTLE to be sent to dismantle the erroneous circuit being acknowledged. (This situation can occur if a circuit takes so long to set up that the Source tries to set up an alternative. The ACKNOWLEDGE datagram for the first circuit will subsequently be received and so the first circuit must be dismantled).

If the Timeout expires before the ACKNOWLEDGE datagram is received then the Source node assumes that an error has occurred. The circuit set up so far is cleared by the Source sending a HARD DISMANTLE. The Source then tries again with a new BUILD datagram, and with the ACKNOWLEDGE Timeout reset. A new Call Identifier is written into the Connection Structure and the new BUILD datagram. When the ACKNOWLEDGE Timeout expires and the Source node makes another attempt to set up a circuit, an associated Retry Count is decremented. If this reaches zero then it is assumed that the circuit cannot be constructed, so the Connection Structure is deleted and the higher layers of the network are informed.

The Retry Count will initially be set to some small integer when the Connection Structure is created. The value of this integer could be chosen by a network manager according to the nature of the particular network. As a guide, DOTI Technical Guide TG101/1 "Intercept Recommendations for Local Area Networks According to the Logical Link Control Access Method" suggests that 4 is a reasonable choice for the maximum number of retries, since this would "allow protection against occasional short duration failures, but provide quick detection of serious problems". Note, however, that this is for a Layer 2 protocol. The same parameter in BT's own Packet SwitchStream is 20.

A node may receive a BUILD datagram nominating it as part of a circuit when it has no outgoing link with sufficient capacity to establish the circuit. If this happens then the BUILD is converted into a SOFT DISMANTLE, its Try Again flag is set to FALSE and the datagram is dispatched back to the previous circuit node. The SOFT DISMANTLE datagram dismantles the circuit back to the neighbouring upstream circuit node so that it can choose a different link to continue the circuit.

When a circuit node receives a SOFT DISMANTLE the Try Again flag is inspected. If it is set to TRUE then the same link may be tried again, so the SOFT DISMANTLE datagram is converted to a BUILD datagram and dispatched to the same downstream circuit node that was chosen before. If the Try Again flag is set to FALSE then another link must be chosen (as will be described below) and the SOFT DISMANTLE is similarly converted to a BUILD and dispatched. If there are no links available then the SOFT DISMANTLE datagram is dispatched to dismantle the circuit further.

The SOFT DISMANTLE is used to dismantle part of a circuit with a view to trying to construct an alternative route from an intermediate circuit node. Normally the SOFT DISMANTLE is transmitted from a circuit node back upstream to its predecessor.

There are, however, occasions when it is necessary to dismantle all the way back to the Source to permit another attempt. The EXTENDED SOFT DISMANTLE is used for this.

When a circuit node receives an EXTENDED SOFT DISMANTLE the datagram is dispatched towards the next upstream circuit node and the associated Connection Structure is then deleted.

When the Source receives the EXTENDED SOFT DISMANTLE it makes another attempt to construct a circuit. If the Retry Count reaches zero then the Connection Structure is deleted and higher protocol layers are informed.

Whenever a circuit node has to send a control datagram to another circuit node then the address of the receiving node is written into the Subdestination field before the datagram is dispatched. The address of the transmitting node is written into the PreviousNode field, ("PREVIOUS NODE ADDRESS" in the lists of pages 14 and 15) if the datagram is a BUILD or a SOFT DISMANTLE.

All of these control datagrams travel from one circuit node to another by the shortest route with respect to delay, not necessarily by the direct link joining the two nodes.

The above description of the construction of data circuits in the network of Figure 1 by means of the algorithm of Figure 2, utilising the BUILD, ACKNOWLEDGE and various DISMANTLE datagrams assumes that each node incorporates in its memory means a list of other nodes in the network and a LIST 1 of the best hop count links to each of those nodes and a LIST 2 of the best delay links to each of those nodes. Such information can either be stored permanently when the network is set up or it can be updated by arranging for each node to communicate changes in its first and/or second lists to neighbouring nodes. If the nodes are arranged to

communicate changes in their lists to neighbouring nodes, it is a preferred feature of the invention that each amended list entry of a given node is communicated only via those links to neighbouring nodes which do not correspond to that entry. As noted above this feature alleviates the problems of "ping-pong" type looping, for the reasons explained mathematically in the Shin and Chen paper referred to above.

5 However this feature can be explained in simple terms with reference to Figure 1. It will be recalled that the route with the shortest delay from node N_C to N_D is via node N_F . Suppose that N_C has just learnt this information from a neighbouring node and has updated its LIST 2 accordingly. It has to decide whether to communicate this updating information to its neighbouring nodes, including node N_F (which is assumed not to know the route with the least delay between N_C and N_D). However if node N_C does communicate this information
10 (namely that this link is the best delay link for communications between N_C and N_D) directly to node N_F over their common link then node N_F may incorrectly deduce that the best delay route from N_F to N_D is over the link to N_C , back in the reverse direction to N_F , and then over the link to N_D . If this incorrect deduction is made, then any datagram transmitted from node N_C to node N_F *en-route* to node N_D will ping-pong indefinitely between nodes N_C and N_F .

15 In normal circumstances, node N_F will know that the quickest route to node N_D is via the direct link L to that node. However in some circumstances that link may not be available, e.g. due to failure or lack of capacity.

It should be noted that an analogous argument applies to the communication of changes to LIST 1, the list of links with the best hop count to various destination nodes.

20 In the embodiment of Figure 1, the problem of "ping-pong" type looping can be avoided simply by ensuring that any amended list entry (i.e. a link) is not transmitted over that link.

In the embodiment of Figures 3 and 4, a more sophisticated procedure is employed and each node is arranged to maintain a routing table in its memory which incorporates, besides the list of best hop count links and list of best delay links (corresponding to LISTs 1 and 2 respectively of Figure 1), a third list of second best hop count links and a fourth list of second best delay links.

25 The algorithm for communicating changes in the first, second, third and fourth lists of each node will first be described with reference to Figure 3, before describing an example of such outdating with reference to Figure 4.

Changes in the lists of each node are communicated in the form of routing information packets (RIP's) each containing one or more routing elements (RE's) of the form:

30 (node address, hop count, delay).

The node address in any RE transmitted by a given node is the address of a destination node, which will typically be a distant node in the network. The hop count and delay in the RE are the lowest hop count and shortest delay between the node in question and the destination node. It will be noted that the RE's do not contain any explicit information which identifies the link with the best hop count and the link with the best delay
35 which lead to the destination node whose address is given in the RE. However, as will become clearer, this information can be deduced from the information already present in the routing table of the node which receives the RE in conjunction with the identity of the link over which the RE is received.

In Figure 3, the first step 150 involves the reception of a RIP containing one or more RE's from an adjacent node. It should be noted that each RE contains a different node address.

40 In step 160, the algorithm considers the routing table of the node which receives the RIP and in respect of each RE determines whether the routing table contains a destination node entry corresponding to the node address given in the RE. The algorithm then proceeds to step 180, creating a new row in the routing table with the required destination node entry (step 170) if necessary.

In step 180 the algorithm increments the hop count in the RE, adds the delay due to the link over which the RE was received to the delay given in the RE and substitutes the resulting values in link columns of the routing table. Examples of such link columns are shown as the three right hand columns in table 2 on p. 22
45 (which will be described in more detail below) but at this stage it should be noted that each link column is a list of hop, delay combinations over a given link from the node in which the routing table is stored to the various destination nodes listed in the left hand column of the table. As will subsequently be described, these link column entries are utilised to update lists 1, 2, 3 and 4.

50 The algorithm then determines whether any of the best or second best hop or delay entries have changed as a result of the new information stored in the routing table (step 190) and if no changes have been made, exit from the process (step 300). In this case, no updating information is transmitted as a result of the receive outdating information in the RIP.

55 If however any of the best or second best hop or delay entries have changed, the algorithm considers the links to adjacent nodes (step 200) and in respect of each link, determines whether that link is both the best hop link and the best delay link to the destination node whose node address is given in the receive RE (step 210). This is to ensure that "best hop" information is not sent over the best hop link and that "best delay" in-

formation is not sent over the best delay link, in order to avoid "ping-pong" type looping. (If a link is both the best hop link and the best delay link (i.e. the link over which a datagram should be sent from the node which has received the RE in order to minimise both hop count and delay) then the routing table of the node is examined and a new RE with the second best hop count and second best delay to the destination node is sent over that link (step 230).

If the link is not both the best hop and best delay link, the algorithm then considers whether the link is merely the best delay link (step 220) and if it is, since it knows from step 210 that it is not the best hop link, it is free to send an RE with the best hop and second best delay over that link (step 240).

If the link examined by the algorithm in step 220 is not the best delay link the algorithm then examines the routing table to determine whether the link being examined is the best hop link (step 250). If it is, then the algorithm knows that it cannot send an RE with the best hop over that link and accordingly it sends an RE with the second best hop and best delay (step 260) to the destination node of the received RE.

If the outcome of step 250 is negative, then the algorithm knows that the link it is examining is neither the best hop link nor the best delay link to the destination node of the received RE. However it may be that only the second best hop and delay entries in the routing table have changed in which case there is no new best hop, best delay information to be sent. This is checked in step 270, and if the answer is affirmative, the algorithm then proceeds to examine the next link radiating from the node which has received the RIP (step 290).

If the answer to the above query is negative, then an RE with the best hop and best delay is sent over that link. The algorithm considers each link in turn in this manner.

Because of the specific manner in which the RE's are guided to the outgoing links from each node, any node which receives an RE can deduce from the link over which that RE is received and the information in its routing table which of its own links will provide the shortest delay and lowest hop count to any given destination node. This will now be explained by way of example only with reference to Figure 4.

Figure 4(a) shows a network comprising nodes N1, N2, N3, N4 and N6, with node N5 about to be connected. The following description will concentrate on the routing table of node N6, whose links L to neighbouring nodes N1, N3 and N4 are numbered 1, 2 and 3 respectively.

As in the network in Figure 1, each node periodically transmits a timing package to its neighbouring node in order to measure the delay to those neighbouring nodes. Additionally, each node keeps a record of the spare data-carrying capacity of each of its links.

This information is stored in a link value table in each case. The link value table for node N1 is as follows:

TABLE 1

Link	1	2	3
Delay	X	X	X
Spare	X	X	X
Capacity			
Neighbour	N1	N3	N4

An entry of X denotes an unknown quantity.
The routing table of node N6 is as follows:

TABLE 2

Desti- nation node	Best Hop Hop (List 1)	2nd Hop Hop (List 3)	Best Delay (List 2)	2nd Delay (List 4)	Hop, Delay for Link		
					1	2	3
N1	1 (1)	3 (2)	X (X)	X (X)	1,X	3,X	NULL
N2	2 (1)	2 (2)	X (X)	X (X)	2,X	2,X	NULL
N3	1 (2)	3 (1)	X (X)	X (X)	3,X	1,X	NULL
N4	1 (3)	(5 (1)) (5 (2))	X (X)	X (X)	5,X	5,X	1,X

The routing table comprises a list of destination nodes N1, N2, N3 and N4 which can be reached from N6 and four lists showing the best available hop count and best available delay to each of these destination nodes. The values in brackets refer to the link radiating from node N6 which must be traversed in order to achieve the indicated best or second best hop or delay value. For example the lowest hop count from node N6 to node N2 is 2 which may be achieved either by traversing link 1 or by traversing link 2 from node N6. Since both these links lead to the same hop count, the former is listed in the best hop count column (LIST 1) on the grounds that the route which includes this link has a lower delay than the route which includes link 2. For the sake of clarity however the delays and links corresponding to these delays are not indicated specifically in the routing tables but are merely represented as X entries. It will be noted that the second best hop count from node N6 to N4 is 5 and that this may be achieved either by exiting from link 1 to node 1 and then travelling to nodes 2, 3, 6 and 4 in succession, or by exiting from node 2 and then travelling to node 3, 2, 1, 6 and 4 in succession. Since both these routes will have the same delay, both entries are bracketed together in LIST 3. But they will not both be stored in the real LIST 3.

It is now assumed that node N5 is added to the network by adding a link between N5 and N4 as shown in Figure 4(b). Node 4 sends the RE (4, 0, 0) to nodes N6 and N5 to indicate its existence. Similarly, node N5 sends the RE (5, 0, 0) to node N4. It will be recalled that each RE is in the form (node address, hop count, delay). Node N4 then knows that node N5 is one hop away from it so it sends the RE (5, 1, X) to node N6. It will be recalled that the value 5 represents the address of node N5, the value 1 is the hop count from node N4 to node N5 and the value X (which will be ignored for the time being) is the delay from node N4 to node N5. Node N6 then updates its routing table to the following:

TABLE 3

5

10

15

20

25

30

35

40

45

50

55

Desti- nation node	Best Hop Hop (List 1)	2nd Hop Hop (List 3)	Best Delay (List 2)	2nd Delay (List 4)	Hop, Delay for Link		
					1	2	3
N1	1 (1)	3 (2)	X (X)	X (X)	1,X	3,X	NULL
N2	2 (1)	2 (2)	X (X)	X (X)	2,X	2,X	NULL
N3	1 (2)	3 (1)	X (X)	X (X)	3,X	1,X	NULL
N4	1 (3)	5 (1) 5 (2)	X (X)	X (X)	5,X	5,X	1,X
<u>N5</u>	<u>2 (3)</u>	<u>-1(-)</u>	<u>X (3)</u>	<u>-1(-)</u>	<u>NULL</u>	<u>NULL</u>	<u>2,X</u>

Changes in the routing table are underlined. It should be noted that the entries of -1 in the second hop and second delay columns (Lists 3 and 4 respectively) for destination node N5 are taken to indicate that the second best hop and second best delay to N5 are infinite, i.e. there is only one route from node N4 to N5.

Node N6 then sends the RE (5, 2, X) to nodes N3 and N1, which then notify node N2. Node N2 then knows it can reach node N5 in four hops via node N3, and therefore sends the RE (5, 4, X) to node N1. The same RE is also sent to node N3. Node N1 then sends the RE (5,5,X) to node N6, whose routing table is as follows (changes being underlined):

TABLE 4

Desti- nation node	Best Hop Hop (List 1)	2nd Hop Hop (List 3)	Best Delay (List 2)	2nd Delay (List 4)	Hop, Delay for Link		
					1	2	3
N1	1 (1)	3 (2)	X (X)	X (X)	1,X	3,X	NULL
N2	2 (1)	2 (2)	X (X)	X (X)	2,X	2,X	NULL
N3	1 (2)	3 (1)	X (X)	X (X)	3,X	1,X	NULL
N4	1 (3)	5 (1) 5 (2)	X (X)	X (X)	5,X	5,X	1,X
N5	2 (3)	6 (<u>1</u>)	X (X)	X (X)	6,X	NULL	2,X

Then node N3 sends the RE (5, 5, X) to node N6, whose table then becomes:

TABLE 5

Desti- nation node	Best Hop Hop (List 1)	2nd Hop Hop (List 3)	Best Delay (List 2)	2nd Delay (List 4)	Hop, Delay for Link		
					1	2	3
N1	1 (1)	3 (2)	X (X)	X (X)	1,X	3,X	NULL
N2	2 (1)	2 (2)	X (X)	X (X)	2,X	2,X	NULL
N3	1 (2)	3 (1)	X (X)	X (X)	3,X	1,X	NULL
N4	1 (3)	5 (1) 5 (2)	X (X)	X (X)	5,X	5,X	1,X
N5	2 (3)	6 (<u>1</u>) 6 (<u>2</u>)	X (X)	X (X)	6,X	<u>6,X</u>	2,X

Now suppose that the link L from node N6 to node N1 fails. As a result, the network shown in Figure 4 (b) is converted to the form shown in Figure 4 (c).

Since each node is arranged to monitor each of its links to a neighbouring node, node N6 detects the failure of link 1.

Accordingly the routing table of node N6 is updated as follows:

TABLE 6

Desti- nation node	Best Hop Hop (List 1)	2nd Hop Hop (List 3)	Best Delay (List 2)	2nd Delay (List 4)	Hop, Delay for Link		
					1	2	3
N1	<u>3</u> (<u>2</u>)	<u>-1</u> (-)	X (X)	X (X)	3,X	NULL	
N2	<u>2</u> (<u>2</u>)	<u>-1</u> (-)	X (X)	X (X)	2,X	NULL	
N3	1 (2)	<u>-1</u> (-)	X (X)	X (X)	1,X	NULL	
N4	1 (3)	<u>5</u> (<u>2</u>)	X (X)	X (X)	5,X	1,X	
N5	2 (3)	<u>6</u> (<u>2</u>)	X (X)	X (X)	<u>6,X</u>	2,X	

Changes in the list entries are underlined and it will be seen that the best hop entries to nodes N1 and N2 have changed and that the second best hop entries to nodes N1, N2, N3, N4 and N5 have changed. In accordance with step 190 of Figure 3, these changes are communicated to the nodes neighbouring N6, namely nodes N3 and N4.

It should be noted that in general, changes in the list entries of a node which result from the receipt of an RE via a given link radiating from that node are not transmitted back over that link since the node at the other end of that link will already be aware of the relevant information.

The transmission of the RIP's from N6 to N3 and N4 is governed by steps 200 to 290 of Figure 3.

The following RE's are transmitted to node N3:

- (1, -1, -1) Node N6's best route to node N1 is now via node 3, and the second best route would normally be sent and there is not one now.
- (2, -1, -1) Node N6's best route to node N2 is now via node N3.
- (3, -1, -1) Node N6's best route to node N3 was via link 2, and still is. The second best information has changed and would therefore be the information sent to node 3. However there is now no second best and so the negative RE is sent.

Node 6 also sends a RIP to node N4 containing the following RE's:

- (1, 3, X)
- (2, 2, X)
- (4, 5, X)
- (5, 6, X).

It should be noted that the RE's sent to node N4 via link 3 from node N6 do not include the node address of node N3. This is because only the second best hop count from N6 to N3 has changed which must therefore be sent over the best hop count link to node N3 (in accordance with step 260 of the algorithm shown in Figure 3) which is not link 3 but link 2.

Finally, N1 will also have learnt that its link to node N6 has failed and will update its routing table accordingly and communicate any changes to node N2 in accordance with the algorithm shown in Figure 3.

It is a preferred feature of the invention that each node is arranged to store in its memory means, in respect of each of its links, a list of pointers to stored connection structures in which that link is an upstream link in a data transmission circuit passing through that node, and a list of pointers to stored connection structures in which that link is a downstream link in a data transmission circuit passing through that node, the node being arranged to detect failure of any of its links, to search the lists of pointers relating to any failed link, and to send datagrams over each remaining link of the connection structures indicated by the pointers relating to the failed link, each connection structure including an upstream link which is directed towards a source node and a downstream link which is directed towards a destination node, each said datagram which is sent via an upstream link being arranged to instruct the associated source node to attempt to establish an alternative circuit and each said datagram which is sent via a downstream link being arranged to cancel the downstream portion of the circuit which interrupted by the link failure.

This feature will now be described by way of example with reference to the network of Figure 4 (a). It is assumed that there are 4 connection structures stored in the memory means of node N6, namely connection structures 10, 20, 30 and 40, each defined by a different combination of upstream and downstream link. These are shown in Table 7 (a):

TABLE 7

CONNECTION STRUCTURE LIST
FOR LINK 1 of NODE N6

CONNECTION STRUCTURE	UPSTREAM LINK	DOWNSTREAM LINK	UPSTREAM POINTER	DOWNSTREAM POINTER
10	1	2	10	30
a) 20	1	3	b) 20	
30	3	1		
40	3	2		

Table 7 (b) shows the upstream and downstream pointers for link 1 of node N6. Thus link 1 is an upstream link in connection structures 10 and 20 and is a downstream link in connection structure 30. Besides the LISTS 1 to 4 shown in the routing tables of tables 2 to 5, the memory of each node in the network of Figure 4 incorporates a list of all the links other than the best hop, second best hop, best delay and second best delay links which can be used to reach each listed destination node. It so happens that in the case of the network of Figure 4 the additional list would be empty in the case of each node, because the nodes have insufficient links. However if the topology of the network changed, for example as a result of the creation of a link between node N6 and node N5 then the additional list in the memory of node N6 would be filled. The purpose of such an additional link is to enable an alternative circuit to be established in the event of the failure of a best or second best hop link. For example, on detecting the failure of link 1 from node 6 in Figure 4 (c) the processor of node N6 would search the list of upstream pointers of Table 7 (b) and would then be directed to connection structures 10 and 20 in Table 7 (a). The processor would then send a HARD DISMANTLE datagram down its downstream link 2 (in respect of connection structure 10) and would send a HARD DISMANTLE datagram down downstream link 3 (in respect of connection structure 20) in order to dismantle the downstream portions of the respective circuits leading from these two connection structures. Similarly, the processor would search the list of downstream pointers of Table 7 (b) and would thereby locate downstream pointer 30 in Table 7 (a). It would then send an EXTENDED SOFT DISMANTLE datagram up through upstream link 3 to the upstream portion of the circuit corresponding to connection structure 30.

In this way, data circuits which rely on a failed link are dismantled. It may also be necessary to dismantle a data circuit because a link in the circuit has insufficient capacity.

As a consequence of the dismantling of a circuit employing best hop count links, it may become necessary to construct an alternative circuit by searching the stored list of all possible links which can be utilised to reach a given destination node. This is reflected in the connection structure format of preferred embodiments of the invention, which is as follows:

Connection Structure

CALL IDENTIFIER

UPSTREAM LINK

DOWNSTREAM LINK

5 BEST HOP LINK

BEST HOP LINK TRIED? (Y OR N)

2ND BEST HOP LINK

2ND BEST HOP LINK TRIED? (Y OR N)

LINK LIST

10 LINK LIST TRIED? (Y OR N)

NODE TYPE (SOURCE, DESTINATION OR INTERMEDIATE CIRCUIT NODE)

The call identifier identifies the circuit and the upstream and downstream link entries indicate the link over which the data is being received at the node in question and the link via which the data of that circuit is transmitted from the node in question. The best hop link is the link appearing in LIST 1 of the node in question for the destination node of the circuit. The connection structure also records whether this best hop link has been tried in an attempt to construct the circuit. Similarly the second best hop link (appearing in LIST 3) is recorded and it is indicated whether this second best hop link has been tried. The link list entry is simply a list of all possible links from the node in question which could be utilised to establish the circuit identified by the call identifier.

20 The variable "LINK LIST TRIED?" indicates whether it has been necessary to search through the link list in an attempt to construct the circuit and finally the variable "NODE TYPE" indicates whether the node in question is a source node, a destination node or an intermediate circuit node. A failure in the network may mean that there is no longer a route from a given source node to a destination node and a circuit cannot be constructed. The source node will eventually learn of this in one of three ways:

25 (i) The source receives a RIP containing a negative RE for the destination node.

(ii) The source node receives a SOFT DISMANTLE datagram and the number of retries at establishing the circuit has reached a predetermined limit. A counter is provided in order to count the number of attempts at establishing a given circuit in each node.

30 (iii) No ACKNOWLEDGE datagram is received by the source node within a predetermined time limit after transmission of a BUILD datagram and the maximum number of retries has been reached.

For the sake of clarity, the additions to the Routing Tables to accommodate the above-described connection structure format have not been described. However they will be apparent to those skilled in the art from the preceding description.

35

Claims

1. A distributed communications network comprising a plurality of nodes (N) which are each capable of acting as source (N_S), sub-destination (N_B, N_C) and destination (N_D) nodes, the nodes being interconnected by links (L) and each node having local processing and memory means (M), each node including in its memory means a stored list of destination nodes in association with a first stored list whose entries directly or indirectly represent links to its neighbouring nodes, each destination node being associated with a respective entry in said first stored list, the processing means being arranged to send a datagram between a source node and a destination node via one or more successive sub-destination nodes in response to the source node calling a remote destination node, each subdestination node being determined (50) by the entry selected from said first stored list in the source node or, as the case may be, in a preceding sub-destination node, the network being characterised in that each node also includes a second stored list whose entries directly or indirectly represent links to its neighbouring nodes, the datagram is arranged to designate at least some of said sub-destination nodes as circuit nodes and to store circuit data in the memory means of the circuit nodes, said circuit data establishing (100) a temporary circuit for data transmission between the source and destination nodes via said circuit nodes, each node (N_E, N_F) traversed by said datagram which is not the source node (N_S), the destination node (N_D) or a circuit node (N_B, N_C) is determined (60) by selecting that entry from said second stored list in the immediately preceding node which is associated with the next designated circuit node in the path of the datagram, and the criteria for determining the entries for said first and second stored lists in each node are distinct and biased towards optimum transmission of said data in said temporary circuit and optimum establishment of said temporary circuit, respectively.

Connection Structure

CALL IDENTIFIER

UPSTREAM LINK

DOWNSTREAM LINK

5 BEST HOP LINK

BEST HOP LINK TRIED? (Y OR N)

2ND BEST HOP LINK

2ND BEST HOP LINK TRIED? (Y OR N)

LINK LIST

10 LINK LIST TRIED? (Y OR N)

NODE TYPE (SOURCE, DESTINATION OR INTERMEDIATE CIRCUIT NODE)

15 The call identifier identifies the circuit and the upstream and downstream link entries indicate the link over which the data is being received at the node in question and the link via which the data of that circuit is transmitted from the node in question. The best hop link is the link appearing in LIST 1 of the node in question for the destination node of the circuit. The connection structure also records whether this best hop link has been tried in an attempt to construct the circuit. Similarly the second best hop link (appearing in LIST 3) is recorded and it is indicated whether this second best hop link has been tried. The link list entry is simply a list of all possible links from the node in question which could be utilised to establish the circuit identified by the call identifier.

20 The variable "LINK LIST TRIED?" indicates whether it has been necessary to search through the link list in an attempt to construct the circuit and finally the variable "NODE TYPE" indicates whether the node in question is a source node, a destination node or an intermediate circuit node. A failure in the network may mean that there is no longer a route from a given source node to a destination node and a circuit cannot be constructed. The source node will eventually learn of this in one of three ways:

25 (i) The source receives a RIP containing a negative RE for the destination node.

(ii) The source node receives a SOFT DISMANTLE datagram and the number of retries at establishing the circuit has reached a predetermined limit. A counter is provided in order to count the number of attempts at establishing a given circuit in each node.

30 (iii) No ACKNOWLEDGE datagram is received by the source node within a predetermined time limit after transmission of a BUILD datagram and the maximum number of retries has been reached.

For the sake of clarity, the additions to the Routing Tables to accommodate the above-described connection structure format have not been described. However they will be apparent to those skilled in the art from the preceding description.

35 Claims

1. A distributed communications network comprising a plurality of nodes (N) which are each capable of acting as source (N_s), sub-destination (N_B, N_C) and destination (N_D) nodes, the nodes being interconnected by links (L) and each node having local processing and memory means (M), each node including in its memory means a stored list of destination nodes in association with a first stored list whose entries directly or indirectly represent links to its neighbouring nodes, each destination node being associated with a respective entry in said first stored list, the processing means being arranged to send a datagram between a source node and a destination node via one or more successive sub-destination nodes in response to the source node calling a remote destination node, each subdestination node being determined (50) by the entry selected from said first stored list in the source node or, as the case may be, in a preceding sub-destination node, the network being characterised in that each node also includes a second stored list whose entries directly or indirectly represent links to its neighbouring nodes, the datagram is arranged to designate at least some of said sub-destination nodes as circuit nodes and to store circuit data in the memory means of the circuit nodes, said circuit data establishing (100) a temporary circuit for data transmission between the source and destination nodes via said circuit nodes, each node (N_E, N_F) traversed by said datagram which is not the source node (N_s), the destination node (N_D) or a circuit node (N_B, N_C) is determined (60) by selecting that entry from said second stored list in the immediately preceding node which is associated with the next designated circuit node in the path of the datagram, and the criteria for determining the entries for said first and second stored lists in each node are distinct and biased towards optimum transmission of said data in said temporary circuit and optimum establishment of said temporary circuit, respectively.



2. A distributed communications network as claimed in Claim 1, wherein the criterion for determining the entries in each said first stored list is that the entries are those which minimise hop-count in the circuit between source (N_S) and destination (N_D) nodes.
- 5 3. A distributed communications network as claimed in Claim 2, wherein the criterion for determining the entries in each said second stored list is that the entries are those which minimise the time taken for the datagram to travel between successive circuit nodes (N_B , N_C).
- 10 4. A distributed communications network as claimed in Claim 3, wherein each node (N) is arranged to update its first stored list in response to hop-count information received (150) from a neighbouring node which relates to hop-count between that neighbouring node and a distant node in the network, each node being arranged to increment the hop-count information from a neighbouring node and to communicate (200-290) the incremented hop-count to its other neighbouring nodes (if any).
- 15 5. A distributed communications network as claimed in Claim 3, wherein each node (N) is arranged to update its second stored list in response to delay information received (150) from a neighbouring node which relates to the transmission delay between that neighbouring node and a remote node in the network, each node being arranged to add (180) to said transmission delay the additional delay (d) associated with receiving the delay information over the link from the neighbouring node, to transmit (200-290) the resulting total delay information to its other neighbouring nodes (if any), and to measure said additional delay by transmitting and receiving a timing packet over the link (L) via which the delay information was received.
- 20 6. A distributed communications network as claimed in any one of claims 1 to 3, wherein the nodes (N) are responsive to conditions in the network to update their first and/or second stored lists (180), each node being arranged to communicate changes (200-290) in its first and/or second stored lists to neighbouring nodes and to modify (180) its first and/or second stored lists in response to updating information received from its neighbouring nodes.
- 25 7. A distributed communications network as claimed in Claim 6, wherein each node (N) further includes third and fourth stored lists whose entries directly or indirectly represent links to its neighbouring nodes, each destination node (N_D) in the stored list of destination nodes being associated with respective entries in said third and fourth stored lists, the criteria on which the entries in said third and fourth stored lists are based being distinct and being biased towards near optimum transmission of said data in said temporary circuit and near optimum establishment of said temporary circuit, respectively.
- 30 8. A distributed communications network as claimed in Claim 7, wherein the entries in each said third stored list represent sub-optimal hop-count between circuit nodes (N) and the entries in each said fourth stored list represent sub-optimal time taken for the datagram to travel between successive sub-destination nodes (N_B , N_C).
- 35 9. A distributed communications network as claimed in any of claims 4 to 8, wherein each node (N) is arranged to communicate changes (200-290) in its said first and/or second stored lists to selected neighbouring nodes in dependence upon the updated entries in its said first and/or second stored lists, each amended entry being communicated only via those links to neighbouring nodes which do not correspond to that entry.
- 40 10. A distributed communications network as claimed in Claim 9, as dependent upon Claim 7 or Claim 8, wherein each node (N) is arranged to communicate changes (200-290) in its said first, second, third and fourth stored lists to selected neighbouring nodes in dependence on the updated entries in its said first, second, third and fourth stored lists, (230, 240, 260, 280) each amended entry of the first and fourth stored lists being communicated (240) via that link (L) to a neighbouring node which is the associated entry in the said second stored list and each amended entry of the second and third stored lists being communicated (260) via that link (L) to a neighbouring node which is the associated entry in the said first stored list.
- 45 50 11. A distributed communications network as claimed in any of claims 4 to 10, wherein each node (N) is arranged to communicate updating information (200-290) to neighbouring nodes for updating their said first and second stored lists, said updating information being communicated within routing information packets containing at least one routing element which includes a node address (A,B,C,D,E,F,G,S:1,2,3,4,5,6,)
- 55



and the entries for that node address in said first and second stored lists, there being at most one routing element corresponding to each node address in a single routing information packet.

- 5 12. A distributed communications network as claimed in Claim 11, wherein each new node (N_5) is arranged to indicate its existence to its neighbouring nodes (N_4) by sending them an identification routing information packet comprising its own node address (5) and entries of zero for the first and second stored lists.
- 10 13. A distributed communications network as claimed in any preceding claim, wherein said circuit data stored by the datagram which is sent from the source node (N_S) includes connection structure information comprising a call identifier which is unique to the circuit being established by the datagram and also the upstream and downstream links to the circuit nodes which are respectively upstream and downstream of the circuit node currently being visited by the datagram, said currently visited circuit node being arranged to check (80, 90) that the call identifier currently being stored by the datagram does not correspond to an identical call identifier already stored in that circuit node, the datagram being arranged to update its connection structure information as it traverses successive nodes in the circuit (N_S, N_B, N_C, N_F, N_D).
- 15 14. A distributed communications network as claimed in Claim 13, wherein, in the event that said currently visited circuit node (N) determines that the call identifier currently being stored by the datagram does correspond to an identical call identifier already stored in that circuit node, the currently visited circuit node is arranged to check (120) the stored connection structure associated with the already-stored call identifier and in the event that the upstream link of the stored connection structure does not lead to the circuit node last visited by the datagram, is arranged to cancel its status as a circuit node and to return the datagram to the last-visited circuit node, (13) said last-visited circuit node being arranged to re-send the returned datagram to an alternative sub-destination node in an attempt to construct a circuit to said destination node (N_S).
- 20 25 15. A distributed communications network as claimed in any preceding claim, wherein each node (N) is arranged to monitor the spare data-carrying capacity of each of its links and in the event that it receives a datagram designating it as a circuit node but the downstream link in the circuit being established by the datagram has insufficient data-carrying capacity, is arranged to transmit a datagram to the source node to instruct it to attempt to establish an alternative circuit to the destination node (N_D).
- 30 16. A distributed communications network as claimed in any preceding claim, wherein each node (N) is arranged to monitor its links (L) to neighbouring nodes and is arranged, in the event that such a link fails, to update its first and/or second stored lists accordingly and to communicate to all its remaining neighbouring nodes any changes in its first and/or second stored lists which result from link failure.
- 35 17. A distributed communications network as claimed in Claim 16, wherein each node (N) is arranged to store in its memory means (M) in respect of each of its links (L), a list of pointers to stored connection structures in which that link is an upstream link in a said data transmission circuit passing through that node, and a list of pointers to stored connection structures in which that link is a downstream link in a said data transmission circuit passing through that node, the node being arranged to detect failure of any of its links, to search the lists of pointers relating to any failed link, and to send datagrams over each remaining link of the connection structures indicated by the pointers relating to the failed link, each connection structure including an upstream link which is directed towards a source node (N_S) and a downstream link which is directed towards a destination node (N_D), each said datagram which is sent via an upstream link being arranged to instruct the associated source node to attempt to establish an alternative circuit and each said datagram which is sent via a downstream link being arranged to cancel the downstream portion of the circuit which is interrupted by the link failure.
- 40 45 18. A distributed communications network as claimed in any of claims 14 to 17, wherein each node includes in its memory means (M) a third stored list as defined in Claim 7 and, when instructed by a datagram to attempt to establish an alternative data circuit, does so by selecting that entry in the third stored list which corresponds to the desired destination node (N_D).
- 50 55 19. A distributed communications network as claimed in Claim 18, wherein each node (N) is arranged to record failed attempts to establish an alternative circuit and to make further attempts to establish a circuit by sending datagrams via its untried neighbouring nodes in sequence, if any.



20. A distributed communications network as claimed in any preceding claim, wherein any node (N), on receiving a datagram which designates it a destination node (N_D) is arranged to send an acknowledgement datagram (110) back to the corresponding source node (N_S).
- 5 21. A distributed communications network as claimed in claim 20, wherein each source (N_S) node is arranged to send a further datagram to attempt to establish a circuit in the event that it receives no acknowledgement datagram within a predetermined period after sending the first-mentioned datagram.
- 10 22. A distributed communications network as claimed in any preceding claim, wherein all information carried by said links (L) is in time-division multiplexed form.
- 15 23. A distributed communications network as claimed in any of claims 4 to 22, wherein the nodes (N) are arranged to give the transmission of data for updating said stored lists (200-290) priority over the transmission of datagrams (60) governing the establishment and release of circuits and to give priority to the transmission of said datagrams over the transmission of data in the circuits established by such datagrams.

Patentansprüche

- 20 1. Verteiltes Kommunikationsnetzwerk mit mehreren Knoten (N), die alle entweder als Quellknoten (N_S), Zwischenzielknoten (N_B , N_C) oder Zielknoten (N_D) dienen können, wobei die Knoten durch Verbindungen (L) miteinander verbunden sind und jeder Knoten eine örtliche Verarbeitungs- und Speichereinrichtung (M) hat, wobei jeder Knoten in seiner Speichereinrichtung in Verbindung mit einer ersten gespeicherten Liste, deren Eintragungen direkt oder indirekt Verbindungen zu den Nachbarknoten darstellen, eine ge-
25 speicherte Liste von Zielknoten aufweist, wobei jeder Zielknoten einer jeweiligen Eintragung in der ersten gespeicherten Liste zugeordnet ist, wobei die Verarbeitungseinrichtung dazu ausgelegt ist, auf den Aufruf eines entfernt gelegenen Zielknotens durch den Quellknoten hin über einen oder mehrere aufeinanderfolgende Zwischenzielknoten ein Datenpaket zwischen einem Quellknoten und einem Zielknoten zu schicken, wobei jeder Zwischenzielknoten aufgrund der Eintragung, die aus der ersten gespeicherten Li-
30 ste im Quellknoten oder gegebenenfalls in einem nachfolgenden Zwischenzielknoten ausgewählt wird, bestimmt (50) wird, wobei das Netzwerk dadurch gekennzeichnet ist, daß jeder Knoten auch eine zweite gespeicherte Liste aufweist, deren Eintragungen direkt oder indirekt die Verbindungen zu seinen Nachbarknoten darstellen, wobei das Datenpaket dazu ausgelegt ist, zumindest einige der Zwischenzielknoten als Bahnknoten zu bestimmen und Bahn-
35 daten in den Speichereinrichtungen der Bahnknoten zu speichern, wobei die Bahn-
daten zeitweilig eine Bahn für die Datenübertragung zwischen dem Quell- oder dem Zielknoten über die Bahnknoten errichten (100), wobei jeder Knoten (N_E , N_F), der von dem Datenpaket durchlaufen wird und der nicht der Quellknoten (N_S), der Zielknoten (N_D) oder ein Bahnknoten (N_B , N_C) ist, durch Auswahl derjenigen Eintragungen aus der zweiten gespeicherten Liste im unmittelbar vorhergehenden Knoten bestimmt (60) wird, die mit dem nächsten bestimmten Bahnknoten auf dem Weg des Datenpakets verknüpft ist, und wobei die Kriterien zur Festlegung der Eintragungen für die erste und die zweite gespeicherte Liste in jedem Knoten klar umrissen und jeweils im Hinblick auf optimale Übertragung der Daten auf der zeitweiligen Bahn sowie optimale Erstellung der zeitweisen Bahn ausgerichtet sind.
- 40 2. Verteiltes Kommunikationsnetzwerk nach Anspruch 1, bei dem das Kriterium zur Bestimmung der Eintragungen in jeder ersten gespeicherten Liste darin liegt, daß die Eintragungen diejenigen sind, die die Sprunganzahl in der Bahn zwischen dem Quellknoten (N_S) und dem Zielknoten (N_D) minimieren.
- 45 3. Verteiltes Kommunikationsnetzwerk nach Anspruch 2, bei dem das Kriterium zur Bestimmung der Eintragungen in jeder der zweiten Liste darin liegt, daß die Eintragungen diejenigen sind, die die Zeit minimieren, die das Datenpaket benötigt, um zwischen aufeinanderfolgenden Bahnknoten (N_B , N_C) zu ver-
50 kehren.
- 55 4. Verteiltes Kommunikationsnetzwerk nach Anspruch 3, bei dem jeder Knoten (N) dazu ausgelegt ist, seine erste gespeicherte Liste auf Sprungzahlinformationen hin, die von einem Nachbarknoten empfangen (150) wurden und die sich auf eine Sprungzählung zwischen diesem Nachbarknoten und einem entfernt gelegenen Knoten im Netzwerk beziehen, zu aktualisieren, wobei jeder Knoten dazu ausgelegt ist, die vom Nachbarknoten kommende Information bezüglich der Sprunganzahl zu erhöhen und die erhöhte Sprunganzahl seinen anderen Nachbarknoten (falls vorhanden) mitzuteilen (200-290).



5. Verteiltes Kommunikationsnetzwerk nach Anspruch 3, bei dem jeder Knoten (N) dazu ausgelegt ist, seine zweite gespeicherte Liste auf Verzögerungsinformationen hin, die von einem Nachbarknoten empfangen wurden (150) und die sich auf die Übertragungsverzögerung zwischen diesem Nachbarknoten und einen entfernt gelegenen Knoten im Netzwerk beziehen, zu aktualisieren, wobei jeder Knoten dazu ausgelegt ist, zur Übertragungsverzögerung die zusätzliche Verzögerung (d) zu addieren, die mit dem Empfang der Verzögerungsinformation über die Verbindung vom Nachbarknoten her einhergeht, die sich ergebende Gesamtverzögerungsinformation an die anderen Nachbarknoten (falls vorhanden) zu übertragen (200-290), und die zusätzliche Verzögerung durch Aussenden und Empfangen eines Zeitmeßpakets über die Verbindung (L), über die die Verzögerungsinformation empfangen wurde, zu messen.
6. Verteiltes Kommunikationsnetzwerk nach einem der Ansprüche 1 bis 3, bei dem die Knoten (N) die Aktualisierung ihrer ersten und/oder zweiten gespeicherten Liste (180) nach Maßgabe von Bedingungen im Netzwerk vornehmen, wobei jeder Knoten dazu ausgelegt ist, Änderungen (200-290) in seiner ersten und/oder zweiten gespeicherten Liste seinen Nachbarknoten mitzuteilen und seine erste und/oder zweite gespeicherte Liste nach Maßgabe von Aktualisierungsinformationen, die er von seinen Nachbarknoten empfangen hat, zu modifizieren (180).
7. Verteiltes Kommunikationsnetzwerk nach Anspruch 6, bei dem jeder Knoten (N) außerdem eine dritte und eine vierte gespeicherte Liste aufweist, deren Eintragungen direkt oder indirekt Verbindungen zu seinen Nachbarknoten darstellen, wobei jeder Zielknoten (N_D) in der gespeicherten Liste von Zielknoten mit den jeweiligen Eintragungen in der dritten und vierten gespeicherten Liste verknüpft ist, und wobei die Kriterien, auf denen die Eintragungen in der dritten und vierten gespeicherten Liste beruhen, klar umrissen und jeweils im Hinblick auf nahezu optimale Übertragungen der Daten über die zeitweilige Bahn und nahezu optimalen Aufbau der zeitweiligen Bahn ausgerichtet sind.
8. Verteiltes Kommunikationsnetzwerk nach Anspruch 7, bei dem die Eintragungen in den dritten gespeicherten Listen suboptimale Sprungzahlen zwischen Bahnknoten (N) und die Eintragungen in den vierten Listen suboptimale Zeiten, die für den Verkehr des Datenpakets zwischen aufeinanderfolgenden Zwischenzielknoten (N_B , N_C) notwendig sind, darstellen.
9. Verteiltes Kommunikationsnetzwerk nach einem der Ansprüche 4 bis 8, bei dem jeder Knoten (N) dazu ausgelegt ist, Änderungen in seiner ersten und/oder zweiten gespeicherten Liste an ausgewählte Nachbarknoten in Abhängigkeit von den aktualisierten Eintragungen in den ersten und/oder zweiten gespeicherten Listen mitzuteilen (200-290), wobei jede verbesserte Eintragung lediglich über die Verbindungen dem Nachbarknoten mitgeteilt wird, die nicht dieser Eintragung entsprechen.
10. Verteiltes Kommunikationsnetzwerk nach Anspruch 9, soweit er von Anspruch 7 oder 8 abhängt, bei dem jeder Knoten (N) dazu ausgelegt ist, Änderungen (200-290) in seiner ersten, zweiten, dritten oder vierten gespeicherten Liste an ausgewählte Nachbarknoten in Abhängigkeit von aktualisierten Eintragungen in seiner ersten, zweiten, dritten und vierten gespeicherten Liste (230, 240, 260, 280) mitzuteilen, wobei jede verbesserte Eintragung der ersten und vierten gespeicherten Liste über diejenige Verbindung (L) einem Nachbarknoten mitgeteilt (240) wird, die in der zweiten gespeicherten Liste die verknüpfte Eintragung ist, und wobei jede verbesserte Eintragung der zweiten und dritten gespeicherten Liste über diejenige Verbindung (L) einem Nachbarknoten mitgeteilt (260) wird, die die verknüpfte Eintragung in der ersten gespeicherten Liste ist.
11. Verteiltes Kommunikationsnetzwerk nach einem der Ansprüche 4 bis 10, bei dem jeder Knoten (N) dazu ausgelegt ist, an Nachbarknoten Aktualisierungsinformationen (200-290) zur Aktualisierung derer erster und zweiter gespeicherten Listen mitzuteilen (200-290), wobei die Aktualisierungsinformationen innerhalb von Weginformationspaketen mitgeteilt werden, die zumindest ein Wegelement aufweisen, das eine Knotenadresse (A, B, C, D, E, F, G, S: 1, 2, 3, 4, 5, 6) aufweist, sowie die Eintragungen für diese Knotenadresse der ersten und zweiten gespeicherten Liste, wobei in einem einzelnen Weginformationspaket maximal ein Wegelement entsprechend einer Wegadresse ist.
12. Verteiltes Kommunikationsnetzwerk nach Anspruch 11, bei dem jeder neue Knoten (N_5) dazu ausgelegt ist, seine Existenz seinen Nachbarknoten (N_4) mitzuteilen, indem er ihnen ein Identifikationsweginformationspaket schickt, das seine eigene Knotenadresse (5) und Null-Eintragungen für die erste und zweite gespeicherte Liste aufweist.



13. Verteiltes Kommunikationsnetzwerk nach einem der vorhergehenden Ansprüche, bei dem die Bahnknoten, die in dem vom Quellknoten (N_S) ausgesendeten Datenpaket gespeichert sind, Verbindungsstrukturinformationen aufweisen, die eine Aufrufinformation enthalten, die für die durch das Datenpaket errichtete Bahn einmalig ist, und außerdem die aufwärtigen und abwärtigen Verbindungen zu den Bahnknoten, die jeweils aufwärtig und abwärtig vom gegenwärtig vom Datenpaket besuchten Knoten sind, wobei der gegenwärtig besuchten Bahnknoten dazu ausgelegt ist zu überprüfen (80, 90), daß die im Datenpaket momentan gespeicherte Aufrufidentifizierung nicht einer identischen Aufrufidentifizierung entspricht, die schon in diesem Bahnknoten gespeichert ist, wobei das Datenpaket dazu ausgelegt ist, seine Verbindungsstrukturinformation zu aktualisieren, wenn es aufeinanderfolgende Knoten auf der Bahn (N_S , N_B , N_C , N_F , N_D) durchläuft.
14. Verteiltes Kommunikationsnetzwerk nach Anspruch 13, bei dem dann, wenn der gegenwärtig besuchte Bahnknoten (N) feststellt, daß die gegenwärtig im Datenpaket gespeicherte Aufrufidentifizierung einer identischen Aufrufidentifizierung entspricht, die bereits in diesem Bahnknoten gespeichert ist, der gegenwärtig besuchte Bahnknoten dazu ausgelegt ist, die der bereits gespeicherten Aufrufidentifizierung zugeordnete gespeicherte Verbindungsstruktur zu überprüfen (120) und in dem Fall, daß die aufwärtige Verbindung der gespeicherten Verbindungsstruktur nicht zu dem zuletzt durch das Datenpaket besuchten Bahnknoten führt, dazu ausgelegt ist, seinen Status als Bahnknoten zu löschen und das Datenpaket an den zuletzt besuchten Bahnknoten zurückzuschicken (13), wobei der zuletzt besuchte Bahnknoten dazu ausgelegt ist, das zurückgeschickte Datenpaket wieder an einen anderen Zwischenzielknoten zu schicken, um zu versuchen, eine Bahn zum Zielknoten (N_G) herzustellen.
15. Verteiltes Kommunikationsnetzwerk nach einem der vorgehenden Ansprüche, bei dem jeder Knoten (N) dazu ausgelegt ist, die freie Datenübertragungskapazität jeder seiner Verbindungen zu überwachen und für den Fall, daß er ein Datenpaket empfängt, das ihn als Bahnknoten ausweist, wobei jedoch die abwärtige Verbindung in der durch das Datenpaket errichteten Bahn keine ausreichende Transportkapazität hat, dazu ausgelegt ist, ein Datenpaket zum Quellknoten zu übertragen, um diesen zu dem Versuch zu veranlassen, eine andere Bahn zum Zielknoten (N_G) aufzubauen.
16. Verteiltes Kommunikationsnetzwerk nach einem der vorhergehenden Ansprüche, bei dem jeder Knoten (N) dazu ausgelegt ist, seine Verbindungen (L) zu Nachbarknoten zu überwachen, und außerdem dazu ausgelegt ist, dann, wenn eine solche Verbindung ausfällt, seine erste und zweite gespeicherte Liste entsprechend zu aktualisieren und allen seinen verbleibenden Nachbarknoten alle Änderungen in der ersten und/oder zweiten gespeicherten Liste, die sich aus dem Verbindungsausfall ergeben, mitzuteilen.
17. Verteiltes Kommunikationsnetzwerk nach Anspruch 16, bei dem jeder Knoten (N) dazu ausgelegt ist, in seiner Speichereinrichtung (M) bezüglich jeder seiner Verbindungen (L) eine Liste von Zeigern zu speichern, die auf gespeicherte Verbindungsstrukturen zeigen, bei denen diejenige Verbindung eine aufwärtige Verbindung in der Datenübertragungsbahn ist, die durch diesen Knoten läuft, sowie eine Liste von Zeigern, die auf gespeicherte Verbindungsstrukturen zeigen, in denen diejenige Verbindung eine abwärtige Verbindung in der Datenübertragungsbahn ist, die durch diesen Knoten läuft, wobei der Knoten dazu ausgelegt ist, Ausfälle in jeder seiner Verbindungen zu erkennen, die Liste bezüglich Zeigern, die sich auf die ausgefallene Verbindung beziehen, zu durchsuchen und über alle verbleibenden Verbindungen der Verbindungsstruktur, die durch die Zeiger, die sich auf die ausgefallene Verbindungen beziehen, bezeichnet werden, Datenpakete zu schicken, wobei jede Verbindungsstruktur eine aufwärtige Verbindung aufweist, die auf den Quellknoten (N_S) zu gerichtet ist, sowie eine abwärtige Verbindung, die auf den Zielknoten (N_D) zu gerichtet ist, wobei jedes der über eine aufwärtige Verbindung geschickten Datenpakete ausgelegt ist, den zugehörigen Quellknoten dazu zu veranlassen, zu versuchen, eine andere Bahn aufzubauen, und wobei jedes Datenpaket, das über eine abwärtige Verbindung geschickt wird, dazu ausgelegt ist, den abwärtigen Teil der Bahn, die durch den Verbindungsausfall unterbrochen wurde, zu löschen.
18. Verteiltes Kommunikationsnetzwerk nach einem der Ansprüche 14 bis 17, bei dem jeder Knoten in seiner Speichereinrichtung (M) eine dritte gespeicherte Liste, wie in Anspruch 7 beschrieben, aufweist, und dann, wenn er durch ein Datenpaket dazu veranlaßt wird, es zu versuchen, eine andere Datenbahn aufzubauen, dies tut, indem diejenige Eintragung in der dritten gespeicherten Liste ausgewählt wird, die dem gewünschten Zielknoten (N_D) entspricht.
19. Verteiltes Kommunikationsnetzwerk nach Anspruch 18, bei dem jeder Knoten (N) dazu ausgelegt ist, fehl-



geschlagene Versuche zum Aufbauen einer anderen Bahn aufzuzeichnen und weitere Versuche zum Aufbauen einer Bahn vorzunehmen, in dem der Reihe nach Datenpakete über die nichtausprobierten Nachbarknoten, falls vorhanden, geschickt werden.

- 5 20. Verteiltes Kommunikationsnetzwerk nach einem der vorhergehenden Ansprüche, bei dem jeder Knoten (N) auf den Empfang eines Datenpakets hin, das ihn als Zielknoten (N_D) bezeichnet, dazu ausgelegt ist, ein Bestätigungsdatenpaket (110) zurück an den entsprechenden Quellknoten (N_S) zu schicken.
- 10 21. Verteiltes Kommunikationsnetzwerk nach Anspruch 20, bei dem jeder Quellknoten (N_S) dazu ausgelegt ist, ein weiteres Datenpaket zu schicken, um den Aufbau einer Bahn für den Fall zu versuchen, daß er innerhalb einer vorbestimmten Zeitdauer nach dem Abschicken des ersterwähnten Datenpakets kein Bestätigungsdatenpaket empfängt.
- 15 22. Verteiltes Kommunikationsnetzwerk nach einem der vorhergehenden Ansprüche, bei dem der Datentransport über die Verbindungen (L) im Zeitmultiplexverfahren vorgenommen wird.
- 20 23. Verteiltes Kommunikationsnetzwerk nach einem der Ansprüche 4 bis 22, bei dem die Knoten (N) dazu ausgelegt sind, der Übertragung von Daten zur Aktualisierung der gespeicherten Listen (200-290) Priorität einzuräumen vor der Übertragung von Datenpaketen (60), die den Aufbau und Abbau von Bahnen festlegen, und der Übertragung dieser Datenpakete Priorität einzuräumen vor der Übertragung von Daten auf den durch diesen Datenpakete erstellten Bahnen.

Revendications

- 25 1. Réseau de communications réparti comprenant une pluralité de noeuds (N) qui sont chacun capable de se comporter comme noeud origine (N_S), de sous destination (N_B , N_C) et de destination (N_D), les noeuds étant interconnectés par des liaisons (L) et chaque noeud comportant un moyen de traitement local et de mémoire (M), chaque noeud comprenant dans son moyen de mémoire une liste mémorisée de noeuds de destination en association avec une première liste mémorisée dont les entrées représentent directement ou indirectement les liaisons à ses noeuds voisins, chaque noeud de destination étant associé à une entrée respective dans ladite première liste mémorisée, le moyen de traitement étant conçu pour envoyer un datagramme entre un noeud origine et un noeud de destination par l'intermédiaire d'un ou plusieurs noeuds de sous destination successifs en réponse au noeud origine appelant un noeud de destination distant, chaque noeud de sous destination étant déterminé (50) par l'entrée sélectionnée à partir de ladite première liste mémorisée dans le noeud origine ou, comme ce peut être le cas, dans un noeud de sous destination précédent, le réseau étant caractérisé en ce que chaque noeud comporte également une seconde liste mémorisée dont les entrées représentent directement ou indirectement les liaisons à ses noeuds voisins, en ce que le datagramme est conçu pour désigner au moins certains desdits noeuds de sous destination comme noeuds de circuit et mémoriser les données du circuit dans le moyen de mémoire des noeuds de circuit, lesdites données du circuit établissant (100) un circuit temporaire pour la transmission de données entre les noeuds origine et de destination par l'intermédiaire desdits noeuds de circuit, chaque noeud (N_E , N_F) traversé par ledit datagramme qui n'est pas le noeud origine (N_S), le noeud de destination (N_D) ou un noeud de circuit (N_B , N_C) est déterminé (80) par la sélection de l'entrée de ladite seconde liste mémorisée dans le noeud immédiatement précédent, qui est associée au noeud de circuit désigné suivant dans le trajet du datagramme, et les critères pour déterminer les entrées dans lesdites première et seconde listes mémorisées dans chaque noeud sont distincts et orientés vers la transmission optimale desdites données dans ledit circuit temporaire et l'établissement optimal dudit circuit temporaire, respectivement.
- 40 2. Réseau de communications réparti selon la revendication 1, dans lequel le critère pour déterminer les entrées dans chaque dite première liste mémorisée est que les entrées soient celles qui minimisent le nombre de connexions directes dans le circuit entre les noeuds origine (N_S) et les noeuds de destination (N_D).
- 45 3. Réseau de communications réparti selon la revendication 2, dans lequel le critère pour déterminer les entrées dans chaque dite seconde liste mémorisée est que les entrées soient celles qui minimisent le temps pris par le datagramme pour progresser entre les noeuds de circuit successifs (N_B , N_C).
- 50
- 55



4. Réseau de communications réparti selon la revendication 3, dans lequel chaque noeud (N) est conçu pour mettre à jour sa première liste mémorisée en réponse à des informations de nombre de connexions directes, reçues (150) d'un noeud voisin, qui se rapportent au nombre de connexions directes entre ce noeud voisin et un noeud distant dans le réseau, chaque noeud étant conçu pour incrémenter les informations de nombre de connexions directes à partir d'un noeud voisin et pour communiquer (200 à 209) le nombre de connexions directes incrémenté à ses autres noeuds voisins (si il y en a).
5. Réseau de communications réparti selon la revendication 3, dans lequel chaque noeud (N) est conçu pour mettre à jour sa seconde liste mémorisée en réponse aux informations de retard, reçues (150) d'un noeud voisin, qui se rapportent au retard de transmission entre ce noeud voisin et un noeud distant dans le réseau, chaque noeud étant conçu pour ajouter (180) audit retard de transmission le retard supplémentaire (d) associé à la réception de l'information de retard sur la liaison provenant du noeud voisin, pour transmettre (200 à 290) les informations de retard totales obtenues à ses autres noeuds voisins (si il y en a) et pour mesurer ledit retard supplémentaire en émettant et recevant un paquet de mesure de temps sur la liaison (L) par l'intermédiaire de laquelle les informations de retard ont été reçues.
6. Réseau de communications réparti selon l'une quelconque des revendications 1 à 3, dans lequel les noeuds (N) sont sensibles à des conditions dans le réseau, pour mettre à jour leurs première et/ou seconde listes mémorisées (180), chaque noeud étant conçu pour communiquer aux noeuds voisins, des modifications (200 à 290) de ses première et/ou seconde listes mémorisées et pour modifier (180) ses première et/ou seconde listes mémorisées en réponse aux informations de mise à jour reçues à partir de ses noeuds voisins.
7. Réseau de communications réparti selon la revendication 6, dans lequel chaque noeud (N) comporte de plus des troisième et quatrième listes mémorisées dont les entrées représentent directement ou indirectement les liaisons à ses noeuds voisins, chaque noeud de destination (N_D) dans la liste mémorisée des noeuds de destination étant associé aux entrées respectives desdites troisième et quatrième listes mémorisées, les critères sur lesquels les entrées desdites troisième et quatrième listes mémorisées sont basés étant distincts et étant orientés vers une transmission presque optimale desdites données dans ledit circuit temporaire et un établissement presque optimal dudit circuit temporaire, respectivement.
8. Réseau de communications réparti selon la revendication 7, dans lequel les entrées de chaque dite troisième liste mémorisée représentent un nombre de connexions directes sous-optimal entre les noeuds de circuit (N) et les entrées de chaque dite quatrième liste mémorisée représentent le temps sous optimal pris pour que le datagramme puisse progresser entre les noeuds de sous destination successifs (N_B , N_C).
9. Réseau de communications réparti selon l'une quelconque des revendications 4 à 8, dans lequel chaque noeud (N) est conçu pour communiquer des modifications (200 à 290) de ses dites première et/ou seconde listes mémorisées à des noeuds voisins sélectionnés en fonction des entrées mises à jour dans ses dites première et/ou seconde listes mémorisées, chaque entrée modifiée étant communiquée uniquement par l'intermédiaire de ses liaisons aux noeuds voisins qui ne correspondent pas à cette entrée.
10. Réseau de communications réparti selon la revendication 9, lorsque dépendants de la revendication 7 ou de la revendication 8, dans lequel chaque noeud (N) est conçu pour communiquer des modifications (200 à 290) dans ses dites première, seconde, troisième, quatrième listes mémorisées, aux noeuds voisins sélectionnés en fonction des entrées mises à jour dans ses dites première, seconde, troisième et quatrième listes mémorisées (230, 240, 260, 280), chaque entrée modifiée des première et quatrième listes mémorisées étant communiquée (240) par l'intermédiaire de cette liaison (L) à un noeud voisin qui est l'entrée associée dans ladite seconde liste mémorisée et chaque entrée modifiée des seconde et troisième listes mémorisées étant communiquée (260) par l'intermédiaire de cette liaison (L) à un noeud voisin qui est l'entrée associée dans ladite première liste mémorisée.
11. Réseau de communications réparti selon l'une quelconque des revendications 4 à 10, dans lequel chaque noeud (N) est conçu pour communiquer des informations de mise à jour (200 à 290) aux noeuds voisins pour mettre à jour leurs dites première et seconde listes mémorisées, lesdites informations de mise à jour étant communiquées à l'intérieur de paquets d'information d'acheminement contenant au moins un élément d'acheminement qui inclut une adresse de noeud (A, B, C, D, E, F, G, S : 1, 2, 3, 4, 5, 6) et les entrées pour cette adresse de noeud dans lesdites première et seconde listes mémorisées, comportant au plus un élément d'acheminement correspondant à chaque adresse de noeud dans un seul paquet d'information



d'acheminement.

- 5 12. Réseau de communications réparti selon la revendication 11, dans lequel chaque nouveau noeud (N5) est conçu pour indiquer son existence à ses noeuds voisins (N4) en envoyant à ceux-ci un paquet d'information d'acheminement d'identification comprenant sa propre adresse de noeud (5) et des entrées de zéro pour les première et seconde listes mémorisées.
- 10 13. Réseau de communications réparti selon l'une quelconque des revendications précédentes, dans lequel lesdites données de circuit mémorisées par le datagramme qui est envoyé à partir du noeud origine (N_S) comportent des informations de structure de connexion comprenant un identificateur d'appel qui est unique pour le circuit qui est établi par le datagramme et également des liaisons amont et aval vers les noeuds du circuit qui sont respectivement en amont et en aval du noeud de circuit actuellement concerné par le datagramme, ledit noeud de circuit actuellement concerné étant conçu pour vérifier (80, 90) que l'identificateur d'appel actuellement mémorisé par le datagramme ne correspond pas à un identificateur d'appel identique déjà mémorisé dans ce noeud de circuit, le datagramme étant prévu pour mettre à jour ses informations de structure de connexion à mesure qu'il passe par les noeuds successifs dans le circuit (N_S, N_B, N_C, N_F, N_D).
- 15 14. Réseau de communications réparti selon la revendication 13, dans lequel, dans le cas où le noeud de circuit actuellement occupé (N) détermine que l'identificateur d'appel actuellement mémorisé par le datagramme correspond à un identificateur d'appel identique déjà mémorisé dans ce noeud du circuit, le noeud de circuit actuellement occupé est conçu pour vérifier (120) la structure de connexion mémorisée associée à l'identificateur d'appel déjà mémorisé et, dans le cas où la liaison amont de la structure de connexion mémorisée ne conduit pas au noeud de circuit concerné en dernier par le datagramme, est conçu pour annuler son état de noeud de circuit et pour renvoyer le datagramme au noeud de circuit concerné en dernier, (13), ledit noeud de circuit concerné en dernier étant conçu pour réexpédier le datagramme renvoyé à un noeud de sous destination de remplacement dans une tentative destinée à construire un circuit vers ledit noeud de destination (N_S).
- 20 25 15. Réseau de communications réparti selon l'une quelconque des revendications précédentes, dans lequel chaque noeud (N) est conçu pour surveiller la capacité de transport de données en réserve pour chacune de ses liaisons et, dans le cas où il reçoit un datagramme le désignant comme un noeud de circuit mais où la liaison aval du circuit qui est établi par le datagramme a une capacité de transport de données insuffisante, est conçu pour transmettre un datagramme au noeud origine afin de lui ordonner d'essayer d'établir un circuit de remplacement vers le noeud de destination (N_D).
- 30 35 16. Réseau de communications réparti selon l'une quelconque des revendications précédentes, dans lequel chaque noeud (N) est conçu pour surveiller ses liaisons (L) aux noeuds voisins et est conçu, dans le cas où une telle liaison est défailante, pour mettre à jour ses première et/ou secondes listes mémorisées de manière correspondante et à communiquer à la totalité de ses noeuds voisins restants, toutes les modifications de ses première et/ou seconde listes mémorisées qui résultent de la défaillance de la liaison.
- 40 45 17. Réseau de communications réparti selon la revendication 16, dans lequel chaque noeud (N) est conçu pour mémoriser dans son moyen de mémoire (M) eu égard à chacune de ses liaisons (L), une liste de pointeurs sur les structures de connexion mémorisées dans laquelle cette liaison est une liaison en amont d'un dit circuit de transmission de donnée passant par ce noeud, et une liste de pointeurs sur les structures de connexion mémorisées dans laquelle cette liaison est une liaison aval dudit circuit de transmission de données passant par ce noeud, le noeud étant conçu pour détecter la défaillance de l'une quelconque des ses liaisons, pour rechercher les listes des pointeurs se rapportant à toute liaison quelconque défail-
50 lante et pour envoyer des datagrammes sur chaque liaison restante des structures de connexion indiquées par les pointeurs se rapportant à la liaison défailante, chaque structure de connexion incluant une liaison amont qui est dirigée vers un noeud origine (N_S) et une liaison aval qui est dirigée vers un noeud de destination (N_D), chaque dit datagramme qui est envoyé par l'intermédiaire d'une liaison amont étant conçu pour ordonner au noeud origine associé d'essayer d'établir un circuit de remplacement et chaque dit datagramme qui est envoyé par l'intermédiaire d'une liaison aval étant conçu pour annuler la partie en
55 aval du circuit qui est interrompu par la défaillance de la liaison.
18. Réseau de communications réparti selon l'une des revendications 14 à 17, dans lequel chaque noeud comporte dans son moyen de mémoire (M) une troisième liste mémorisée comme définie dans la reven-



dication 7 et, lorsqu'un datagramme lui donne l'ordre d'essayer d'établir un circuit de données de remplacement, l'exécute en sélectionnant l'entrée de la troisième liste mémorisée qui correspond au noeud de destination désiré (N_D).

- 5 19. Réseau de communications réparti selon la revendication 18, dans lequel chaque noeud (N) est conçu pour enregistrer les essais pour établir un circuit de remplacement qui ont échoué et pour effectuer d'autres essais pour établir un circuit en envoyant en séquence des datagrammes par l'intermédiaire de ses noeuds voisins non essayés, s'il y en a.
- 10 20. Réseau de communications réparti selon l'une quelconque des revendications précédentes, dans lequel chaque noeud quelconque (N), sur réception d'un datagramme qui le désigne en tant que noeud de destination (N_D) est conçu pour renvoyer un datagramme d'acquittement (110) au noeud origine correspondant (N_S).
- 15 21. Réseau de communications réparti selon la revendication 20, dans lequel chaque noeud origine (N_S) est conçu pour envoyer un autre datagramme pour tenter d'établir un circuit dans le cas où il ne reçoit pas de datagramme d'acquittement dans un intervalle de temps prédéterminé après l'envoi du datagramme mentionné en premier.
- 20 22. Réseau de communications réparti selon l'une quelconque des revendications précédentes, dans lequel toutes les informations véhiculées par lesdites liaisons (L) sont sous une forme multiplexée par répartition dans le temps.
- 25 23. Réseau de communications réparti selon l'une des revendications 4 à 22, dans lequel les noeuds (N) sont conçus pour donner priorité à la transmission des données de mise à jour desdites listes mémorisées (200 à 290) sur la transmission des datagrammes (60) gouvernant l'établissement et la libération des circuits, et pour donner la priorité à la transmission desdits datagrammes sur la transmission des données dans les circuits établis par ces datagrammes.

30

35

40

45

50

55

