PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 :		(11) International Publication Number:
H04L 12/24	**A2**	**WO 98/09403**
		(43) International Publication Date: 5 March 1998 (05.03.98)

(21) International Application Number: PCT/GB97/02160

(22) International Filing Date: 12 August 1997 (12.08.97)

(30) Priority Data:
9617907.2 28 August 1996 (28.08.96) GB

(71) Applicant *(for all designated States except US)*: BRITISH TELECOMMUNICATIONS PUBLIC LIMITED COMPANY [GB/GB]; 81 Newgate Street, London EC1A 7AJ (GB).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: COTTER, David [GB/GB]; 23 Moorfield Road, Woodbridge, Suffolk IP12 4JN (GB).

(74) Agent: WELLS, David; BT Group Legal Services, Intellectual Property Dept., Holborn Centre, 8th floor, 120 Holborn, London EC1A 7AJ (GB).

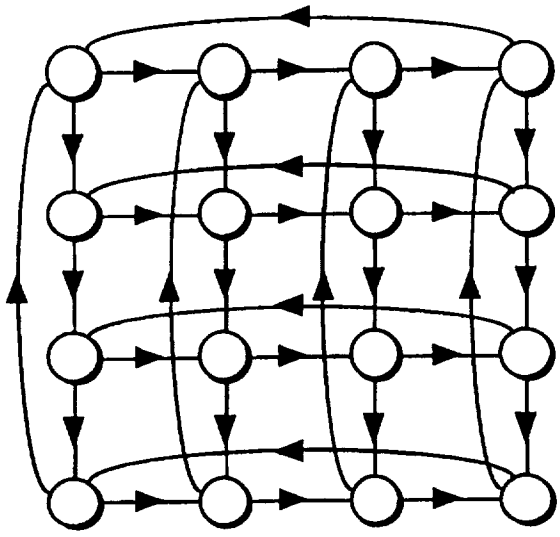
(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: COMMUNICATIONS NETWORK

(57) Abstract

A communications network, suitable, for example, for linking computer processors, is formed from a number of nodes and links. The nodes and links are configured as a multiplicity of directed trails. Each directed trail spans some only of the nodes, but in combination the directed trails span every node of the network. Packets are routed through the network by selecting the appropriate one of the directed trails which links the source node and destination node, and by outputting the packet at the source node onto the selected trail. The nodes throughout the network may switch between predetermined and prescheduled switching states, and a given trail may be selected by choosing appropriately the time slot in which the packet is put onto the network. The network may be a photonic network carrying optical packets.



FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

COMMUNICATIONS NETWORK

BACKGROUND TO THE INVENTION

The present invention relates to a communications network, and in particular to the routing of packets in a network.

There is a need for high speed packet networks for use, for example, in local area networks for interconnecting computer systems, or for use as part of the internal infrastructure of a multiprocessor computer. It has been proposed to implement such networks using ultrafast photonic technology. In implementing the network, the provision of an appropriate routing mechanism for the packets has proved to be a particularly critical design problem. There are a number of potentially conflicting requirements which have to be satisfied by any routing mechanism. In particular, it is desirable that the processing overhead at each node should be kept low: otherwise, the time required to process the packet limits the throughput of the node and hence of the network. However, it is also desirable to ensure that the routing mechanism is as efficient as possible in selecting the shortest paths for packets from a source to a given destination. If the efficiency of the routing mechanism is poor, then congestion in the network rises. This again serves to limit the throughput of the network.

Many different approaches to packet routing have been proposed and studied [1-3]. By selecting a suitable network topology the decision making associated with routing can be greatly simplified. This then meets the first of the requirements identified above, by ensuring a low processing overhead. Some network topologies allow "one-dimensional" routing. For example, in a uni-directional ring the source simply places a packet onto the ring and the packet eventually reaches its destination without requiring any routing decision by intermediate nodes. One-dimensional routing may also be used, for example, on buses or rings with bi-directional links. However, although one dimensional routing offers a number of attractive features, it suffers a serious limitation in that there is poor scaling of the relative routing efficiency and maximum throughput of the network as the number of nodes in the network is increased.

As an alternative to the use of one-dimensional network topologies, multi-dimensional networks may be used. The main advantage of using a multi-dimensional network, such as a two-dimensional torus network, is that there is a

multiplicity of paths available between any pair of nodes on the network. It is therefore possible to adopt a routing method that selects the shortest available path, and such a method in general will have a higher relative routing efficiency and a higher value for the maximum throughput. However, existing methods of routing on multi-dimensional networks have their own serious drawbacks. Because the routing methods can select from a multiplicity of paths through the network it is possible for two or more packets to attempt to occupy the same link simultaneously. The method therefore needs to be capable of resolving the contention that arises in this case. This may be done either by using buffering within the network to store one or more of the packets until the link is free, or by deflecting one or more packets away from its optimal path, the technique known as deflection routing [1,7]. Buffering can successfully maintain the packets in their correct sequence during their journey across the network. Buffering is however an unattractive solution for high-speed networks, because it introduces variable and unpredictable delay and adds to the control complexity and cost. In the case of photonic networks, there are severe technological limitations to the construction of buffers. Deflection routing does not suffer these technological limitations and is much easier to control, but deflection also causes packets to suffer variable and unpredictable delays. The packets may therefore be delivered to the destination in an incorrect sequence. A further limitation of conventional multi-dimensional networks and routing mechanisms is that it is difficult to achieve broadcasting to the nodes without using higher level transport control layers. The use of higher level transport control layers is undesirable as it introduces substantial processing delays.

The paper by Yener, B. et al., PROCEEDINGS OF THE GLOBAL TELECOMMUNICATIONS CONFERENCE (GLOBECOM), SAN FRANSISCO, NOV. 28 - DEC. 2, 1994, vol. 1 of 3, IEEE, pp 169-175, and patent US-A-5297137, discuss the use of multiple spanning trees to enhance the fault tolerance of a switch-based network. The use of two virtual rings is proposed. Each ring is embodied by an edge-disjoint spanning tree of the network and each virtual ring therefore spans every node of the network. It is therefore not possible to route a packet efficiently simply by choosing one of the virtual rings at the outset. As a packet passes

through the network, routing is carried out in a conventional fashion by making a local routing decision at intermediate nodes.

The paper by ZHENSHENG ZHANG et al, published in NETWORKING IN THE NINETIES, Bal Harbour, Apr. 7-11, 1991, vol. 3, 7 April 1991, IEEE, pp 1012-1021 discloses a network which is synchronised so that the transmissions of all the nodes begin in the start of a time slot. However, although the switching of nodes is to this extent pre-scheduled, the switching between different states is not predetermined. That is to say, it is not possible to predict in advance that a certain switch will be in a certain state at a given instant. Rather, whether or not a switch changes from one state to another in a given time slot is determined by a local routing decision made using a hot potato algorithm. The switch state therefore depends on local traffic conditions.

SUMMARY OF THE INVENTION

According to a first aspect of the present invention, there is provided a method of routing a packet in a communications network which comprises a multiplicity of nodes and links, and in which the nodes and links are configured as a multiplicity of directed trails, each directed trail linking some only of the multiplicity of nodes and the directed trails in combination spanning every node of the network, the method comprising:

a) selecting a directed trail T from the multiplicity of directed trails in dependence upon the destination of a packet, the selected trail including the source node and destination node of the packet; and

b) outputting the packet at the source node onto the selected one of the multiplicity of directed trails.

The present invention provides a fundamentally new approach to packet routing in multi-dimensional communications networks. It potentially offers all the advantage of one-dimensional routing, whilst overcoming the crucial disadvantage of poor scalability of routing efficiency and of the maximum throughput. It also completely avoids the need for contention resolution within the network.

The invention uses a method termed by the inventor "directed trail routing". This takes advantage of the fact that a network having an appropriate topology, as further described below, can be divided into a set of distinct trails, such that no one single trail spans all of the network, but there is always one trail

which leads from a given source node to a given destination node. Routing can then be carried out simply by selecting the appropriate trail linking a source node to the desired destination node. Once on the trail, the packet can be routed in a quasi-one-dimensional fashion. As in one-dimensional routing, and by contrast with the prior art approaches in the papers by Yener et al. and Zhang et al., the source node selects the entire trail from the source to the destination before sending the packet.

Preferred implementations of the invention offer the advantages of very simple processing and routing nodes which may be based on simple header-word recognition. Message delivery time is dominated by the speed of light delay. No contention is required within the network, there is no need for buffering within the network and the network is free of deadlock or livelock. Nodes in the network can be named in an arbitrary fashion. Networks embodying the present invention offer efficient routing and good throughput and these advantages are maintained as the method is scaled for larger networks. Using the invention, there is zero delay variation and packets can be delivered in the correct sequence. The invention can support both connectionless (datagram) and connection-orientated modes and in connection-orientated mode provides guaranteed bandwidth. The method also makes physical-layer broadcasting practical.

Preferably the method includes reading the destination address at each node traversed by the packet.

In preferred implementations of the invention, the only processing required at the node is that needed to determine whether or not the packet address matches the node address.

Preferably each intermediate node forwards a packet which is received at the intermediate node and which is addressed to another node in a direction which is predetermined and independent of any information carried by the packet.

An important advantage of the present invention is that it enables intermediate nodes to function without requiring any processing of the packet beyond that necessary to determine whether or not the packet destination address corresponds to the node address.

Preferably the packet is an optical packet carried on a photonic network.

Although the present invention is by not means limited in applicability to systems operating in the optical domain, it does offer special advantages when used with ultrafast photonic networks. It enables efficient use of the bandwidth offered by such networks, while avoiding the disadvantages of using deflection routing or the technological problems associated with the provision of optical buffers.

Preferably each directed trail in the network is a subgraph of at least one closed directed trial and consists of a directed cycle or union of a plurality of connected directed cycles from a link-disjoint directed-cycle decomposition of the network.

Preferably the step of selecting a directed trail T includes synchronising the initial despatch of the packet with prescheduled switching at an intermediate node.

It is found that a particularly effective way of routing a packet along a trail formed from a number of directed cycles is to switch the optical output of the intermediate nodes at prescheduled times e.g. with a fixed periodicity, so as to connect one cycle to another cycle. The source node then determines the trail followed by the packet by outputting the packet at a time determined in relation to the switching schedule so that, at a desired node, it is switched from one cycle to the next cycle in the trial. Preferably the switching occurs at a point of connection between cycles from a link-disjoint directed-cycle decomposition of the network. Preferably the nodes switch in synchronism throughout the network between pre-scheduled pre-determined switching states. For example, in the 4x4 torus network described below, a crossbar switch is associated with each node. All the crossbar switches are normally set to the cross state and repeatedly, at predetermined intervals, the crossbar switches are set to the bar state.

Although the use of temporal switching is preferred, the invention is not limited solely to such techniques. The invention might be implemented, for example, using wavelength-switching instead.

Preferably a timing sequence for the prescheduled switching of intermediate nodes comprises a frame divided into a plurality of time slots, and a source node outputs a packet onto the network in a selected one of the plurality of time slots within the timing frame, and the length of the link between successive

nodes in a trial is such that a packet leaving one node in a first time slot arrives at the next successive node in a second time slot. For example, the packet may be advanced or retarded by one time slot as it passes from one node to the next.

According to a second aspect of the present invention there is provided a communications network comprising:

a) a multiplicity of nodes and links which are configured as a multiplicity of directed trails each of which spans some only of the multiplicity of nodes and the multiplicity of directed trails in combination spanning every node of the network and including at least one directed trail linking each source node and each destination node in the network;

b) means for selecting a directed trail T from the multiplicity of directed trails in dependence upon the destination of a packet; the selected trail including the source node and destination node of the packet; and

c) means for outputting the packet at the source node onto the selected one of the multiplicity of directed trails.

According to a third aspect of the present invention there is provided a communications network comprising:

a) a multiplicity of nodes and links which are configured as a multiplicity of directed trails each of which spans some only of the multiplicity of nodes and the multiplicity of directed trails in combination spanning every node of the network and including at least one directed trail linking each source node and each destination node in the network;

b) a routing controller arranged to select, in dependence upon the destination of a packet, a directed trail T from the multiplicity of directed trails, which directed trail includes the source node and destination node of the packet;

c) an input for receiving packets for transmission on the network; and

d) an output connected to the said input for receiving packets and connected to the routing controller and arranged to output a packet received at the said input onto a selected one of the said multiplicity of directed trails.

According to a fourth aspect of the present invention, there is provided an optical communications network comprising:

a) means for selecting a time slot for a packet in dependence upon the desired path for the packet through the network;

b) means for outputting the packet onto the network at a source node in the time slot selected by the means for selecting;

c) means for switching periodically the routing states of a multiplicity of routing nodes;

d) at the network nodes, means for switching the packet to different outputs according to the routing state of the respective node in the time slot in which the packet arrives at the node; and

e) means for receiving the packet at a destination node.

According to a fifth aspect of the present invention there is provided a method of routing an optical packet from a source node to a destination node in an optical communications network, comprising:

a) selecting a time slot for a packet in dependence upon the desired path for the packet through the network;

b) outputting the packet onto the network at a source node in the time slot selected in step (a);

c) switching periodically the routing states of a multiplicity of routing nodes;

d) as the packet traverses the network, at the different nodes switching the packet to different outputs according to the routing state of the respective node in the time slot in which the packet arrives at the node; and

e) receiving the packet at a destination node.

The present invention also encompasses multiprocessor computer systems, local area networks, metropolitan area networks, campus networks and communications switches using networks in accordance with the preceding aspects of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

Systems embodying the present invention will now be described in further detail, by way of example only, and will be contrasted with the prior art, with reference to the accompanying drawings in which:

Figure 1 is a schematic of a first example of a communications network;

Figure 2 is a circuit diagram for one of the nodes of Figure 1;

Figure 3 is a diagram showing the partitioning of a packet time slot;

Figure 4 is a diagram showing the topology of a 4x4 torus network;

Figure 5 is a diagram showing the topology of a 4x4 Manhattan Street Network (MSN);

Figure 6 is a graph showing the efficiency of different routing schemes in nxn torus and nxn MSN networks relative to shortest-path routing in the nxn MSN versus the network dimension n;

Figure 7 is a diagram showing connected link-disjoint directed cycles;

Figures 8a and 8b are diagrams showing torus networks using crossbar switches with the switches set to the cross and bar state respectively;

Figure 9 is a timing diagram showing packet time slots arranged in a frame structure;

Figure 10 is a diagram showing the lengths of links connecting adjacent nodes;

Figure 11 shows examples of directed-trail routings using switching between horizontal and vertical cycles in the network of Figure 8;

Figure 12 shows a two-dimensional 3x3 mesh with bidirectional links;

Figure 13 shows the network of Figure 12 with the switches in the cross position;

Figure 14 shows the network of Figure 12 with the switches in the bar position;

Figure 15 shows a 2-ary 4-cube hypercube network;

Figure 16 shows a packet switch; and

Figure 17 shows a multiprocessor computer.

DESCRIPTION OF EXAMPLES

An optical communications network comprises a LAN 1 linking a number of personal computer workstations 3. Each workstation is connected to the LAN via a network interface 3a. The workstations and LAN together provide a distributed computing environment which may be used, for example, for the visualisation of complex data. Each workstation is connected to a respective node 2 of the network. Packets of data 4 are communicated between the workstations 3 via the nodes 2 and links 5. In this example, the links 5 are formed from optical fibre and transmit the packets 4 in the optical domain.

Although, for ease of illustration, only a few nodes are shown in the Figure, in practice, the network may comprise many hundreds of nodes.

In this first example, the nodes and the interconnecting fibres are configured as an nxn torus network.

The $n \times n$ torus network is a regular network with unidirectional links, and the nodes have indegree and outdegree of 2. Logically, the links form a grid on the surface of a torus, and all the links in the *n* rows or *n* columns are codirectional. An example of a 4×4 network is shown in Fig. 4. Each node contains a 2×2 'crossbar' switch or its logical equivalent. In the cross configuration the switch connects the input column to the output column, and the input row to the output row; in the bar configuration the input column is connected to the output row, and the input row is connected to the output column. Figure 8(a) shows the situation in which all switches are set to the cross position. In that case, the network consists of a set of 2*n* cycles, each of length *n*. In the notation here, the *n* horizontal cycles are denoted C_{ih} where $i = 0,1,...,n-1$, and the *n* vertical cycles are denoted C_{jv} where $j = 0,1,...,n-1$. This link-disjoint directed-cycle decomposition of the network graph is well suited to the routing method of the present invention. The $n \times n$ torus network contains n^2 distinct closed directed trails, defined as $T_{ij} = C_{ih} \cup C_{jv}$, with $i,j = 0,1,...,n-1$. The cutpoint of trail T_{ij} occurs at the intersection of the component cycles, at the node (*i,j*); in other words, each of the n^2 nodes in the network is the cutpoint of exactly one of the closed directed trails T_{ij}. This cycle-decomposition of the network is well suited to the directed-trail routing method of the present invention, because a packet can be routed from its source to its destination, both located anywhere in the network, along a directed trail consisting of a vertical or horizontal cycle or the union of one vertical and one horizontal cycle; therefore the packet must be switched between cycles a maximum of once (at the cutpoint which is the point of connection between the vertical and horizontal cycles). A different cycle-decomposition of the network is obtained when all the switches are set to the bar position, as shown in Fig. 8(b); in that case the network consists of *n* cycles, each of length 2*n*. However this cycle decomposition is less well suited to directed-trail routing because a directed trail leading between a source-destination pair may, necessarily, be the union of many cycles.

The switching operations that maintain a packet on its selected trail leading from its source to its destination can operate in an automatic fashion, without requiring the intermediate nodes to interrogate the packet destination address or to perform any intelligent route selection. The network operates in a slotted fashion with packets constrained to some maximum length, i.e. time is divided into a regular time slots which are dimensioned to contain a packet of the maximum allowable size together with a guard band. The crossbar switches in all the routing nodes in the network are arranged to operate in a regular, coherent fashion, locked to a global network clock at the time-slot rate. When the switches change configuration they do so during the guard band so as not to corrupt packets. Figure 9 is a time diagram showing the packet time slots, each of length T, arranged in frames of length n time slots. In the first $n-1$ time slots in a frame, the crossbar switches are all set in the cross position (denoted *c* in the diagram); in the final time slot of the frame the switches are all set to the bar position (denoted *b*). As shown by way of example in Figure 10, the length of each link connecting a pair of adjacent nodes in the network is selected and controlled so that the signal group time-of-flight is equal to $(qn+1-\Delta)T$, where q is any integer, and Δ is the phase difference between the clock signals at the two nodes, expressed as a fraction of the time slot period T. In other words, apart from the clock phase difference ΔT, the length of every link in the network is equal to an arbitrary integer number of frames plus one time slot. Thus a packet which exits from a node in the jth time slot of a frame will arrive at the next node in the $(j+1)$th time slot of a frame. More generally, the packet may be advanced or retarded by a fixed integer number of time slots. The packet may be advanced/retarded by any fixed integer number of slots which is not a multiple of n, if n is odd, or by a number which is odd if n is even.

Figure 11 shows how a 4×4 torus network might appear to some of the packets travelling inside it. Suppose, in Fig.11, node *A* wishes to transmit a packet to node *D*. The source node *A* will use a look-up table or some other algorithm to determine that it should transmit the packet along the outward link in the horizontal cycle C_{2h} using a vacant time slot in the third position in a frame (in this example each frame contains 4 slots). On arrival at the next node, *B*, the packet will find itself in the fourth (i.e. the last) slot in a frame, and therefore the

crossbar switch at *B* will be configured in the bar position, as shown in Fig. 11. The packet is therefore switched into the vertical cycle C_{1v} and progresses onwards through node *C* (where it is now in the first time slot of a frame, so the switch at *C* is in the cross state), eventually reaching its destination node *D*. Not shown in Fig. 11 is an alternative routing; exiting from *A* along the vertical cycle C_{0v} in the second time slot of a frame, via nodes *E* and *F*.

In the torus network, directed-trail routing using the trails $T_{ij} = C_{ih} \cup C_{jv}$ is 100% efficient; i.e. the directed-trail routing gives the shortest path between any source-destination pair. It can be shown that the average shortest-path distance in the $n \times n$ torus network is equal to $n^2/(n+1)$ hops. Since the maximum steady-state throughput is given by the indegree of the nodes divided by the average distance taken, it is equal to $2(n+1)/n^2$. Therefore the throughput scales as $O(1/\sqrt{N})$, where $N = n^2$ is the number of nodes, in contrast with one-directional routing where the throughput scales as $O(1/N)$.

As indicated earlier, with directed-trail routing, the only processing operation that a network node is required to perform on incoming packets is simple: the destination address of every incoming packet is examined, and if it corresponds to the address of the node the packet is removed from the network, otherwise it is forwarded. The process of comparing the packet address and the node address is a simple single-word matching operation, and can be performed at high speed; for example, optical recognition of 6-bit address words has been demonstrated recently at a peak rate of 100 Gbit/s [11]. Since the directed-trail routing does not use an algorithm that relies on any particular sequential numbering system for the network nodes, the nodes can be labelled in an entirely arbitrary fashion. This can simplify the tasks of planning, administrating and evolving the network.

Although the description of directed-trail routing in the torus network has assumed, until now, that the network is a complete $n \times n$ structure, it is also possible to cope with the situation where a node is missing or a node or link fails. If a node is missing, it is necessary only that the links that bypass this vacant position maintain the correct timing relationship, as shown e.g. in Fig. 10, i.e. in this case the link that bypasses a vacant node position should have a signal group

delay of $(qn+2-\Delta)T$ (an integer number of frames plus *two* time slots minus the phase correction ΔT). Unlike one-dimensional routing methods, the operation of the entire network is not jeopardised by the loss of a link or node. This is because there are two available directed-trail routes between any source-destination pair, provided the source and destination are not located in the same horizontal or vertical cycle. If one directed trail fails, the other available trail can be used instead. If the source and destination are located on the same vertical or horizontal cycle there is no alternative directed-trail routing, but the cycle can be healed by bypassing the defective node or link in the manner just described in the case of a missing node.

Neither is it strictly necessary for the torus network to be square; for example, directed-trail routing can be used in a rectangular network with m rows and n columns, where $m>n$, say. Then the frame must contain the number of time slots corresponding to the greater dimension (m in this case), and there are $m-n$ missing, or 'phantom', columns in a complete $m \times m$ structure. Packets travelling along a row in a time slot corresponding to a phantom cutpoint can remain only within the row. This reduces the network efficiency, but ensures that all the mn real nodes can be accessed.

The network can support both connectionless and connection-oriented modes of operation simultaneously. In connectionless mode, a source can transmit to a chosen destination by inserting packets in a vacant time slot corresponding to a directed trail that leads to the destination. There are no guarantees on the availability of suitable vacant time slots, and communication between the source and destination is transitory. In connection-oriented mode, prior signalling is used to reserve a regular sequence of suitable time slots allowing a connection to be established between a source-destination pair.

In one-dimensional networks, broadcasting from a given source to every other node in the network can be achieved in a straightforward fashion; a packet which is to be broadcast is copied by each node as it traverses the network. In multi-dimensional networks that do not use predetermined routing tables (such as self-routing algorithms with deflections for contention resolution), broadcasting is not efficient if physical-layer mechanisms are used alone. Unless higher transport-layer protocols are used, to broadcast a packet it is necessary to transmit an

individual copy to every node in the network; in other words, physical-layer broadcasting is of complexity O(*N*), where *N* is the number of nodes. Whilst transport-layer protocols can be highly effective in low-speed networks, for very high-speed networking the additional delays may be unacceptable. Physical-layer broadcasting in the torus network can be achieved using directed-trail routing by transmitting copies on a set of trails that span the full extent of the network (for example, to broadcast from a node in the *i*th row of an $n \times n$ torus, packet copies could be transmitted on the *n* trails $T_{ij} = C_{ih} \cup C_{jv}$, where $j = 0,1,2,...,n-1$. This physical-layer broadcasting is of complexity O($\sqrt{N}$), where $N = n^2$ is the number of nodes.

Figure 2 shows the structure of a 2-connected node suitable for insertion in the MSN shown in Figure 4. In a photonic network implementation, the heavy lines shown in Figure 2 are optical fibre paths. The delay units at the two inputs to the node provide the necessary adjustment on the lengths of the two incoming links to satisfy the requirement described earlier, that on each link connecting a pair of nodes, the signal group time-of-flight along the link should be equal to $(qn + 1 - \Delta)T$, where *q* is any integer, and Δ is the phase difference between the clock signals at the two nodes, expressed as a fraction of the time slot period *T*. By providing two delay lines (one on each input) with independent compensation for phase differences, it is also possible to ensure that the packets on the two incoming links are correctly synchronised relative to each other and to the time-slot clock at the node. Each delay unit could consist of a combination of: i) a length of fibre cut to a suitable length to provide coarse timing adjustment; ii) a step-adjustable delay line consisting of a chain of 2x2 space switches and fibre delays (such as described in ref [12]) to provide timing adjustment to within a few hundred picoseconds; and iii) a free-space adjustable optical delay line (such as optical delay line type ODL-300-15-SMF manufactured by Santec Corporation) to provide fine adjustment to within a few tens of picoseconds. It may be necessary also to compensate for slow drifts in the optical path length of the incoming links. These drifts may be caused by environmental factor acting on the fibre - for example, movement causing stretch, or temperature variations. This continuous environmental compensation can be achieved by detecting a variation in the relative timing of incoming packet arrivals and the time-slot clock at the node, and

providing an electrical feedback control signal to the step-adjustable delay line and the free-space adjustable delay line units.

The header-processing units performs the following tasks: i) detects the presence or absence of a packet in a time slot; ii) detects the time of arrival of a packet; and iii) determines whether or not an incoming packet is addressed to the node. For tasks i) and ii) it is sufficient to use a ~1 GHz-bandwidth photodetector to detect a fraction of the packet signal. The presence of a signal from this photodetector during the time slot indicates the presence of a packet. The phase relationship between the time-slot clock and the component of this photodetector signal which is at the time-slot rate can be detected using an electronic phase detection circuit, and a voltage proportional to this phase difference provides the control signal necessary for the feedback control circuit mentioned above. For task iii) it is necessary to compare the address in the packet header with the address of the node. For an ultrafast photonic implementation this can be performed using the method of ultrafast binary word recognition described in the present applicant's international patent application PCT/GB94/00397, with further technical details disclosed in WO 95/33324. The contents of these earlier applications are incorporated herein by reference. An experimental demonstration of this technique is described in ref [11]. As described in the above-cited applications, address words for packets are selected from the subsets of binary words for which the following condition is true for any two words A, B in the subset:

$A \otimes B = 0$ only if $A = B$,

and $A \otimes B = 1$ otherwise,

where $A \otimes B$ is the Boolean operation

$$\sum_{i=1}^{n} a_i . \bar{b}_i .$$

Word recognition is then carried out using a simple AND operation between an address word from a packet and the complement of the node address. A suitable AND gate is a semiconductor optical amplifier supporting four-wave mixing (FWM).

This method of word recognition provides a binary output signal indicating whether or not the header destination address matches the node address.

The basic space-switching operation is performed by the three crossbar switches. The reason for 3 switches rather than only 1 is that this provides the signal paths needed to connect to and from a local host computer system. Suitable space switches capable of operation in a time of 1 ns or less are lithium niobate devices such as type Y-35-8772-02 supplied by GEC Advanced Components.

Suitable partioning of the packet time slots is shown in Figure 3. In this example it has been assumed that the time slot clock is at at a rate of 155 MHz (6.45 ns period). This is a standard clock used currently in SDH networks and can be distributed over wide (national) geographical regions with timing jitter of less than 500 ps. It is also assumed in the example shown in Figure 3 that the packet consists of 53 bytes at 100 Gbit/s (4.24 ns duration), that a suitable switch band for operation of the electro-optic space switches is 1 ns, and in addition there are two time guard bands each of size 0.6 ns. Within the node, the position of the current time slot in the frame can be tracked by an electronic modulo *n* counter (for a frame *n* time slots long) which counts the time-slot clock pulses. During the initial start-up phase of the network, and subsequently when time slots are available, one node in the network (designated a master node) can broadcast packets in one fixed position in the frame (such as the first position), so that the counters in other nodes can be reset to the correct phase in synchronism with the master node.

The space switches in the node are activated by the electronic switch controller unit shown in Figure 2 which acts on the basis of the following information: i) whether or not the position of the time slot in the frame corresponds to a 'cross' or 'bar' configuration in the directed-trail routing cycle (1 bit); ii) whether or not an incoming packet occupies the current time slot (1 bit per input port); iii) whether or not the destination address for an incoming packet matches the address of the node (1 bit per input port); iv) whether or not a packet that is waiting in the host's output buffer wishes to access an output port in the current time slot (1 bit per output port). On the basis of this information, (total 7 bits) the electronic switch controller unit sends electrical drive signals to the space switches in correct synchronism with the time guard bands between packets, and in this way performs the following tasks: i) routes incoming packets to the host or to one

of the output ports; ii) routes packets from the host to one of the output ports if the required time slot is vacant. An example of the logic required to perform these tasks is as follows:

```
if not (current time slot is last position in frame )
        then
        S1 := cross;
if ( (incoming row time slot is occupied) and not (incoming
column time slot is occupied) and not ( incoming row packet is
addressed to host ) and ( a host packet is waiting to exit from
the row port in the current time slot ) and ( a host packet is
waiting to exit from the column port in the current time slot )
) { comment - destination address of vacant incoming column time
slot is not defined } then
begin
        S2 := cross;
        S3 := bar;
end;
```

The routing logic, of which this is an example, is sufficiently simple that it can be executed using hard wiring together with a fast 8-bit decoder chip, without the need for arithmetic, registers or look-up tables. It is purely a logical combination circuit, and therefore the decision time depends only on gate delays. The switch controller unit can therefore operate at high speed, suitable for routing packets in multi-Gbit/s networks.

The routing method described above offers high efficiency and low latency, and these advantages are maintained as the number of nodes in the network is scaled upwards. Figure 6 shows the efficiency of various routing schemes in nxn torus and MSN networks, relative to shortest-path routing in the nxn MSN, as a function of the network dimension n. The various curves are:A, hamiltonian-cycle routing in nxn torus network;B, directed-trail routing in nxn torus network; C, directed-trail routing in nxn MSN; D, dead-reckoning routing in nxn MSN; E,routing in nxn MSN using a routing rule that provides a shortest-path route (Maxemchuk's "first rule"[5]). It can be seen that methods B and C implementing the present invention maintain a relatively high efficiency of 60% or greater as the

network size is scaled upwards. The efficiency is markedly greater than hamiltonian-cycle routing which is a one-dimensional routing technique applicable to a torus network. As noted in the introduction above, the present invention may be applied, for example, in a packet switch. Figure 16 shows an NxN buffered packet switch, in which the interconnect 160 is a network embodying the invention. A further application is illustrated in Figure 17 which shows a multiprocessor computer in which processors, memory, file stores and I/O ports are connected to respective nodes of a network employing directed-trail routing.

Although the first example described above uses an nxn torus topology, directed trail routing can be used with a wide variety of different network topologies, including topologies which are irregular in form. In the sections which follow, a formal general description of directed trail routing is given and examples of further alternative network topologies are described.

Directed-trail routing — a general description

This section gives a general, mathematical description of the directed-trail routing method. The following terms that appear in this section, and elsewhere in the description and claims, are those commonly used and defined in graph theory (see, for example, [8]), and have the meanings normally accorded them in that field: graph, digraph, subgraph, strongly-connected, directed, points, arc, path, trail, to span, indegree, outdegree, eulerian, hamiltonian, cycle, link-disjoint, cycle-decomposition, cutpoint.

The communications network is represented by a strongly-connected digraph [8], denoted *G*, in which each point uniquely represents a network node, each arc represents a one-directional link between a pair of distinct nodes, and each node is assigned a unique address. In such a digraph, a trail is defined as an alternating sequence of nodes and links, beginning and ending with nodes, in which each link is a directed link from the immediately-preceding node to the immediately-following node, and all the links (but not necessarily all the nodes) are distinct. For every node n_i in *G*, we define a set S_i of distinct trails such that no single trail spans *G*, but there is at least one trail in S_i that leads from n_i to any other node in *G*. To effect the transmission of an information packet from its source n_i to a destination node n_j in *G* ($n_i \neq n_j$), the source merely attaches the destination address to the packet and despatches the packet on a trail *T*, selected

from S_i, which contains n_j. For the complete transmission of a packet from its source to its destination, the sole routing decision is that made by the source when *T* is selected. The routing information needed by the source may be stored in a look-up table. The packet itself carries no routing information, apart from its destination address, and intermediate nodes between the source and destination perform no routing function in reaction to the packet. The only processing operation that network nodes are required to perform on incoming packets is simple: at every node n_k in *G*, the destination address of every incoming packet is examined; if the packet destination address corresponds to the address of n_k the packet is removed from the network, otherwise it is forwarded. Mechanisms ensure that a packet, once despatched by its source onto a chosen trail, will be forwarded along that trail until removed at the packet destination. Crucially, the operation of these mechanisms at all nodes is entirely independent from the presence or absence of any packet or packet content.

A basis for engineering such routing mechanisms in certain network topologies is now described. It is necessary that the digraph *G* representing the network is eulerian (i.e., at every point in *G* the in-degree and out-degree are equal). In fact, as described later, most practical communications networks are eulerian. According to a known theorem in graph theory [9], an eulerian digraph is the union of a set of link-disjoint directed cycles (a directed cycle being a closed alternating sequence of nodes and links, in which each link is a directed link from the immediately-preceding node to the immediately-following node, and all the links and nodes are distinct). This set is a link-disjoint directed-cycle decomposition of *G*, denoted *D*(*G*). We now introduce a further theorem:

Theorem: If two or more arc-disjoint directed cycles are connected, their union is a closed trail.

Proof: Two arc-disjoint directed cycles are connected if they share a common point *x*, and their union is a closed trail in which *x* is a cutpoint (as is self-evident in Fig. 7 in the case of the two cycles *CKLAC* and *CDHJC*, for example, which share a common point *C*, and whose union is the closed trail *CKLACDHJC*). If this closed trail shares a common point *y* with a further arc-disjoint directed cycle (where *x* and *y* may be the same or differ), then their union is a further closed trail in which *y* is a cutpoint. (For example, in Fig. 7, the closed trail *CKLACDHJC*

shares a common point *D* with the cycle *DEFBD*, and their union is the closed trail *CKLACDEFBDHJC* .) By extrapolation, the theorem is proved for arbitrary numbers of connected cycles.

It follows that any open trail in *G* (such as a trail belonging to the set S_i, for any n_i in *G*) is a subgraph of at least one closed trail consisting of a cycle, or the union of several connected cycles, from *D*(*G*). (For example, in Fig. 7, the shortest open trail from *A* to *B*, namely *ACDEFB*, is a subgraph of the closed trail *CKLACDEFBDHJC* which is the union of the connected cycles *CKLAC*, *CDHJC* and *DEFBD*.) Cutpoints in the closed trails occur at every connecting point of the component cycles. Rearrangement of the internal connections between inward and outward links at the various cutpoints cause different trails to be selected. (For example, in Figure 7, the internal connections within node *C* can cause each of the inward links *AC*, *PC* and *JC* to be directed onwards to a selected one of the links *CK*, *CM* and *CD*.) To enable automatic trail-routing of packets, switching between the trails in *G* may be arranged to occur at regular, pre-determined times. A general routing mechanism for forwarding a packet along a selected trail *T* then consists in ensuring that the initial dispatch of the packet from its source is synchronised with prescheduled switching at the cutpoints. This prescheduled switching is arranged to occur in a synchronised, temporally coherent fashion throughout the network. This method of packet routing is illustrated by the examples described in the preceding and following sections.

As already stated, for this routing mechanism to work, the network must have an eulerian topography. It turns out that this is true for a very wide range of communications networks. Many networks are symmetric in the graphical sense (a digraph is symmetric if, for every link *uv* within it, there is also a link *vu* [10]), and are therefore eulerian. In a symmetric network each pair of symmetric links connecting two adjacent nodes forms a cycle, and according to the theorem stated above, the union of an arbitrary number of such cycles, provided they are connected, forms a closed trail. An example of a regular symmetric network is an *n*-dimensional mesh with bidirectional pairs of links. The links in all router networks and wide-area telecommunications networks occur in symmetric pairs, and so these networks could in principle support directed-trail routing, although the network topologies are usually highly irregular and therefore this method of routing

is likely to be inefficient. However, many of the topologies of greatest interest for multi-processor interconnection are regular eulerian networks, either symmetric (such as the *n*-dimensional meshes with bidirectional links) or oriented (i.e. having no symmetric pair of directed links, such as the *k*-ary *n*-cubes), and are therefore well suited to directed-trail routing. Examples of the *k*-ary *n*-cubes are the torus networks and hypercubes.

The routing mechanism is illustrated below by various examples: the *k*-ary 2-cube ($k \times k$ torus network); the reverse *k*-ary 2-cube ($k \times k$ reverse-torus network, also known as the Manhattan-Street network); a regular 2-dimensional $n \times n$ mesh with bidirectional links; and the 2-ary *n*-cube (hypercube). The torus and reverse-torus networks are considered in some detail.

Directed trail routing offers many of the advantages of one-dimensional routing while overcoming the limited efficiency of such techniques. One-dimensional routing may be carried out in one-dimensional networks—e.g. single-directional or bi-directional busses or rings—or using hamiltonian cycles. In both directed-trail routing and one-dimensional routing, the source node selects the entire trail from the source to the destination before sending the packet. Intermediate nodes need perform no routing function in reaction to a packet, and this greatly simplifies the design of the nodes. This also eliminates contention within the transmission path; this removes the necessity for contention resolution using internal network buffers or deflection-routing strategies, which introduce delay variations and can undesirably rearrange a packet sequence. Because neither complex processing of packet-header data nor buffering is required at routing nodes, the networks are also well suited to all-optical implementations, which allow higher transmission bandwidth and lower delay. However, the disadvantage in the case of one-dimensional routing is that the source must make a choice between at most two available paths, each of which span all the network nodes. This results in a relatively high average source-destination distance of $O(N)$, where N is the number of nodes in G. Since the theoretical maximum throughput in the steady state is inversely proportional to the average source-destination distance, the throughput is relatively poor. On the other hand, in the case of directed-trail routing in a well-chosen network topology, there can be a multiplicity of directed trails leading from the source, each of which spans only a subset of the nodes in

G, and the source is required to select a trail which contains the required destination node for the packet. This results in a lower average source-destination distance, and consequently the throughput is significantly improved relative to one-dimensional routing. As in the case of one-dimensional routing, networks based on directed-trail routing can support both connection-oriented and connectionless modes of communication. In connection-oriented mode, quality-of-service guarantees on bandwidth and delay can be provided. Also, broadcasting is simplified.

Manhattan Street network (MSN)

The $n \times n$ MSN is a regular network with unidirectional links, and the nodes have indegree and outdegree of 2. Logically, the links form a grid on the surface of a torus. The MSN differs from the torus network in that the links in adjacent rows or columns travel in opposite directions, and the MSN is defined only in the case that the numbers of rows and columns are even. An example of a 4×4 network is shown in Fig. 5. The routing scheme for the MSN using directed trails $T_{ij} = C_{ih} \cup C_{jv}$ is closely similar to the torus network; the main difference is that alternate horizontal or vertical cycles have opposite orientation.

Unlike in the torus network, the relative routing efficiency of directed-trail routing in the MSN (compared to shortest-path routing) is less than 100%. However, the average shortest-path distance in the MSN is less than in the torus network of equal size (approaching a factor of 2 shorter for large networks). The shortest directed-trail distances between source-destination pairs in the MSN depends on the relative orientations of the inward and outward links at the nodes, and the formulae are set out in Table 1. Using these formulae, the relative routing efficiency (relative to shortest-path routing) can be calculated, and the results are presented in Fig. 6. These show that directed-trail routing in the MSN, although less efficient than a shortest-path algorithm (such as Maxemchuk's 'first rule') or the dead-reckoning method, is still good. In particular, the routing efficiency for directed-trail routing in large MSNs is around 0.65.

Source node orientation	Destination node orientation	Shortest directed-trail distances (the lesser if two expressions are given)
down and right	down and right	(dr-sr) mod n + (dc-sc) mod n
	down and left	(dr-sr) mod n + (dc-sc) mod n n - (dc-sc) mod n + (dr-sr) mod n
	up and left	(dc-sc) mod n + n - (dr-sr) mod n n - (dc-sc) mod n + (dr-sr) mod n
	up and right	(dc-sc) mod n + n - (dr-sr) mod n (dc-sc) mod n + (dr-sr) mod n
down and left	down and right	(dc-sc) mod n + (dr-sr) mod n n - (dc-sc) mod n + (dr-sr) mod n
	down and left	(dc-sc) mod n + (dr-sr) mod n
	up and left	(dc-sc) mod n + n - (dr-sr) mod n (dc-sc) mod n + (dr-sr) mod n
	up and right	(dc-sc) mod n + n - (dr-sr) mod n n - (dc-sc) mod n + (dr-sr) mod n
up and left	down and right	(dc-sc) mod n + n - (dr-sr) mod n n - (dc-sc) mod n + (dr-sr) mod n
	down and left	(dc-sc) mod n + n - (dr-sr) mod n (dc-sc) mod n + (dr-sr) mod n
	up and left	(dc-sc) mod n + (dr-sr) mod n
	up and right	(dc-sc) mod n + (dr-sr) mod n n - (dc-sc) mod n + (dr-sr) mod n
up and right	down and right	(dc-sc) mod n + n - (dr-sr) mod n (dc-sc) mod n + (dr-sr) mod n
	down and left	(dc-sc) mod n + n - (dr-sr) mod n n - (dc-sc) mod n + (dr-sr) mod n
	up and left	(dc-sc) mod n + (dr-sr) mod n n - (dc-sc) mod n + (dr-sr) mod n
	up and right	(dc-sc) mod n + (dr-sr) mod n

TABLE 1 : Shortest directed-trail distances between source-destination pairs in the $n \times n$ MSN. For the purposes of these formulae only, the rows and columns are each numbered sequentially 0,1,...,$n-1$. Even numbered rows are oriented towards the 'right'; odd numbered rows are oriented towards the 'left'. Even numbered columns are oriented in the 'down' direction; odd numbered columns are oriented in the 'up' direction. The source is located at the intersection of row *sr* and column *sc*; similarly the destination is located at the intersection of row *dr* and column *dc*. If, for example, *sr* is even and *sc* is odd, the source node orientation is called 'up and right'.

Figure 6 also allows a comparison between the efficiencies of directed-trail routing in the MSN and torus network of the same dimensions. It should be noticed that all the curves in Figure 6 indicate the efficiencies of various routing schemes relative to the efficiency of shortest-path routing in a MSN. Although (as described in the previous example) directed-trail routing in the torus network successfully finds the shortest-path routing with 100% accuracy, this routing still has a lower relative efficiency than directed-trail routing in the MSN; this is because the average shortest-path distance in the MSN is significantly less than the average shortest-path distance in the torus network (by a factor approaching 2 for large networks).

The maximum theoretical throughput in the steady state for the MSN is equal to the indegree of the nodes (2) divided by the average shortest-path distance. The throughput obtainable with directed-trail routing is equal to 2 divided by the average distance travelled, or expressed another way, the throughput obtainable is equal to the maximum theoretical throughput multipled by the routing efficiency (around 65% for large networks).

2-dimensional mesh with bidirectional links

Figure 12 shows a two-dimensional 3x3 mesh with bidirectional links. This is an example of a network which is symmetric in the graphical sense (defined earlier) because for every link there is a corresponding link with the opposite orientation, forming a symmetric pair. Each node contains a space switch that can be configured in the cross or bar position. Figure 13 shows the configuration of the network when all the switches are in the cross position. In this case, the network is composed of a set of closed trails (T_{1h}, T_{1v}, and so on), each trail

consisting of the union of cycles formed from symmetric pairs. Figure 14 shows the configuration when all the switches are in the bar position. A set of closed trails that can be used for directed-trail routing then consist of the union of one horizontal trail (such as T_{1h}) and one vertical trail (such as T_{1v}). Directed trail routing is then obtained with a timing scheme similar to that described for routing in the torus network. In this case the frames should be of length $2n$-2 time slots, for an nxn mesh with bidirectional links (e.g. in the example given of a 3x3 network, the frame length is 4 time slots), and the frame consists of $2n$-1 'cross' time slots and one 'bar' time slot. This allows directed-trail routing between any source and any destination in the network.

Hypercube

Figure 15 shows a 2-ary 4-cube, which is an example of a hypercube network. Like the torus and Manhattan-street networks, the hypercube is an example of an oriented graph (one not having symmetric pairs of links). The nodes are all 2-connected, and contain 2x2 crossbar space switches. Figure 15 also shows a directed-cycle decomposition of the network. The network is composed of 4 cycles: 1,2,3,9,10,15,16,8,1; 1,13,16,12,10,4,3,5,1; 2,7,8,6,15,14,9,11,2; and 5,6,4,7,12,11,13,14,5. A set of closed trails that can be used for directed-trail routing in this network consist of the union of any 2 out of these 4 cycles. Directed trail routing is the obtained with a timing scheme similar to that described for routing in the torus network. In this example the frames should be of length 4 time slots, and the frame consists of 3 'cross' time slots and one 'bar' time slot. This allows directed-trail routing between any source and any destination in the network. For example, supposing the node labelled 1 in Figure 15 is a source of packets. Then a packet transmitted from the source node 1 along the link towards node 2 in the first time slot of a frame will be forwarded along the closed trail 1,2,3,9,11,2,7,8,1. A packet transmitted from the source node 1 along the link towards node 13 in the first time slot of a frame will be forwarded along the closed trail 1,13,16,12,11,13,14,5,1. A packet transmitted from the source node 1 along the link towards node 2 in the second time slot of a frame will be forwarded along the closed trail 1,2,3,5,1,13,16,8,1. A packet transmitted from the source node 1 along the link towards node 13 in the second time slot of a frame will be forwarded along the closed trail 1,13,16,8,1,2,3,5,1. A packet

transmitted from the source node 1 along the link towards node 2 in the third time slot of a frame will be forwarded along the closed trail 1,2,7,8,6,15,16,8,1. A packet transmitted from the source node 1 along the link towards node 13 in the third time slot of a frame will be forwarded along the closed trail 1,13,14,5,6,4,3,5,1. A packet transmitted from the source node 1 along the link towards node 2 in the fourth time slot of a frame will be forwarded along the closed trail 1,13,16,12,10,15,16,8,1. A packet transmitted from the source node 1 along the link towards node 13 in the fourth time slot of a frame will be forwarded along the closed trail 1,2,3,9,10,4,3,5,1.

References

[1] Steenstrup, M. (ed.): *Routing in Communications Networks*, Prentice Hall, 1995

[2] Partridge, C.: *Gigabit Networking*, Addison-Wesley, 1994

[3] Baransel, C., Dobosiewicz, W. and Gburzynski, P.: 'Routing in Multihop Packet Switching Networks—Gb/s Challenge', *IEEE Network*, May/June 1995, pp. 38-61

[5] Maxemchuk, N.F.: 'Routing in the Manhattan Street Network', *IEEE Transactions on Communications*, **35**, pp. 503-512 (1987)

[6] PCT/GB 96/01823; Cotter, D. and Tatham, M.C.: 'Dead Reckoning—A Primitive and Efficient Self-Routing Protocol for Ultrafast Mesh Networks', paper submitted for publication

[7] Borgonovo, F.: 'Deflection Routing', Chapter 9 in [1]

[8] Harary, F.: *Graph Theory*, Addison-Wesley, Reading, Mass., 1969

[9] Harary F., Norman, R.Z. and Cartwright, D: *Structural Models: An Introduction to the Theory of Directed Graphs*, John Wiley, New York, 1965, p. 330

[10] *ibid.*, p. 11

[11] Cotter, D., Lucek, J.K., Shabeer, M., Smith, K., Rogers, D.C., Nesset, D. and Gunning, P.: 'Self-Routing of 100 Gbit/s Packets Using 6-Bit 'Keyword' Address Recognition', *Electronics Letters*, **31**, pp. 2201-2202 (1995)

[12] P R Prucnal et al (IEEE J Quantum Electronics, vol 29, no 2, pp. 600-612, 1993)

CLAIMS

1. A method of routing a packet in a communications network which comprises a multiplicity of nodes and links, and in which the nodes and links are configured as a multiplicity of directed trails, each directed trail linking some only of the multiplicity of nodes and the directed trails in combination spanning every node of the network, the method comprising:

a) selecting a directed trail T from the multiplicity of directed trails in dependence upon the destination of a packet, the selected trail including the source node and destination node of the packet; and

b) outputting the packet at the source node onto the selected one of the multiplicity of directed trails.

2. A method according to claim 1, including reading, at each node traversed by the packet, a destination address carried in the packet.

3. A method according to claim 2, in which each intermediate node forwards a packet which is received at the intermediate node and which is addressed to another node in a direction which is predetermined and independent of any information carried by the packet.

4. A method according to any one of the preceding claims, in which the packet is an optical packet carried on a photonic network.

5. A method according to any one of the preceding claims, in which each directed trail in the network is a subgraph of at least one closed directed trail and consists of a directed cycle or union of a plurality of connected directed cycles from a link-disjoint directed-cycle decomposition of the network.

6. A method according to any one of the preceding claims, in which the step of selecting a directed trail T includes synchronising the initial dispatch of the packet with prescheduled switching at an intermediate node.

7. A method according to any one of the preceding claims, in which switching occurs at a point of connection between cycles from a link-disjoint directed-cycle decomposition of the network.

8. A method according to any one of the preceding claims, in which the nodes switch in synchronism throughout the network between pre-scheduled pre-determined switching states.

9. A method according to claim 8, in which:

a timing sequence for the prescheduled switching of intermediate nodes comprises a frame divided into a plurality of time slots and a source node outputs a packet onto the network in a selected one of a plurality of time slots within a timing frame, and the length of the link between successive nodes in a trail is such that a packet leaving one node in a first time slot arrives at the next successive node in a second time slot.

10. A method according to claim 9, in which the length of the link is such that on travelling between the successive nodes the packet is advanced or retarded by one time slot.

11. A communications network comprising:

a) a multiplicity of nodes and links which are configured as a multiplicity of directed trails each of which spans some only of the multiplicity of nodes and the multiplicity of directed trails in combination spanning every node of the network and including at least one directed trail linking each source node and each destination node in the network;

b) means for selecting a directed trail T from the multiplicity of directed trails in dependence upon the destination of a packet, the selected trail including the source node and destination node of the packet; and

c) means for outputting the packet at the source node onto the selected one of the multiplicity of directed trails.

12. A communications network comprising:

a) a multiplicity of nodes and links which are configured as a multiplicity of directed trails each of which spans some only of the multiplicity of nodes and the multiplicity of directed trails in combination spanning every node of the network and including at least one directed trail linking each source node and each destination node in the network;

b) a routing controller arranged to select, in dependence upon the destination of a packet, a directed trail T from the multiplicity of directed trails, which directed trail includes the source node and destination node of the packet;

c) an input for receiving packets for transmission on the network; and

b) an output connected to the said input for receiving packets and connected to the routing controller and arranged to output a packet received at the said input onto a selected one of the said multiplicity of directed trails.

13. A network according to claim 11 or 12, in which each node in a directed trail is arranged to output a packet which is received at the node input from the network, and which is addressed to another node, in a direction which is predetermined and independent of any information carried by the packet.

14. A network according to any one of claims 11 to 13, in which the network is a photonic network arranged to carry optical packets.

15. A network according to claim any one of claims 11 to 14, in which each directed trail in the network is a subgraph of at least one closed directed trail and consists of a directed cycle or union of a plurality of connected directed cycles from a link-disjoint directed-cycle decomposition of the network.

16. A network according any one of claims 11 to 15, in which an originating node is arranged to synchronise the initial dispatch of a packet with prescheduled switching at an intermediate node.

17. A network according to any one of claims 11 to 16, in which switching occurs at a point of connection between cycles from a link-disjoint directed-cycle decomposition of the network.

18. A network according to any one of claims 11 to 17, in which the nodes are arranged to switch in synchronism throughout the network between pre-scheduled pre-determined switching states.

19. A network according to claim 18, in which:

a timing sequence for the prescheduled switching of intermediate nodes comprises a frame divided into a plurality of time slots and

a source node is arranged to output a packet onto the network in a selected one of a plurality of time slots within a timing frame, and

the length of the directed link between successive nodes in a trail is such that a packet leaving one node in a first time slot arrives at the next successive node in a second time slot.

20. An optical communications network, comprising:

a) means for selecting a time slot for a packet in dependence upon the desired path for the packet through the network;

b) means for outputting the packet onto the network at a source node in the time slot selected by the means for selecting;

c) means for switching repeatedly with a fixed periodicity the routing states of a multiplicity of routing nodes;

d) at the network nodes, means for switching the packet to different outputs according to the routing state of the respective node in the time slot in which the packet arrives at the node; and

e) means for receiving the packet at a destination node.

21. A communications network comprising:

a) a multiplicity of nodes and links which are configured as a multiplicity of directed trails which in combination span every node of the network and which include at least one directed trail linking each source node and each destination node in the network;

b) means for selecting a directed trail T from the multiplicity of directed trails in dependence upon the destination of a packet, the selected trail including the source node and destination node of the packet; and

c) means for outputting the packet at the source node onto the selected one of the multiplicity of directed trails.

22. A method of routing an optical packet from a source node to a destination node in an optical communications network, comprising:

a) selecting a time slot for a packet in dependence upon the desired path for the packet through the network;

b) outputting the packet onto the network at a source node in the time slot selected in step (a);

c) switching repeatedly with a fixed periodicity the routing states of a multiplicity of routing nodes;

d) as the packet traverses the network, at the different nodes switching the packet to different outputs according to the routing state of the respective node in the time slot in which the packet arrives at the node; and

e) receiving the packet at a destination node.

23 A computer system comprising a plurality of processors interconnected by a network according to any one of claims 11 to 21.

24. A local area network for interconnecting a plurality of computer systems including a network according to any one of claims 11 to 21.

25. A switch for use in a communications network, the switch including an internal interconnect comprising a network according to any one of claims 11 to 21.

26. A method of routing a packet in a communications network which comprises a multiplicity of nodes and links, and in which the nodes and links are configured as a multiplicity of directed trails which in combination span every node of of the network, the method comprising:

a) selecting a directed trail T from the multiplicity of directed trails in dependence upon the destination of a packet, the selected trail including the source node and destination node of the packet; and

b) outputting the packet at the source node onto the selected one of the multiplicity of directed trails.

27. A method according to any one of claims 1 to 10 or according to claim 26, including outputting copies of a packet concurrently on a plurality of trails, thereby broadcasting the packet to a plurality of nodes.

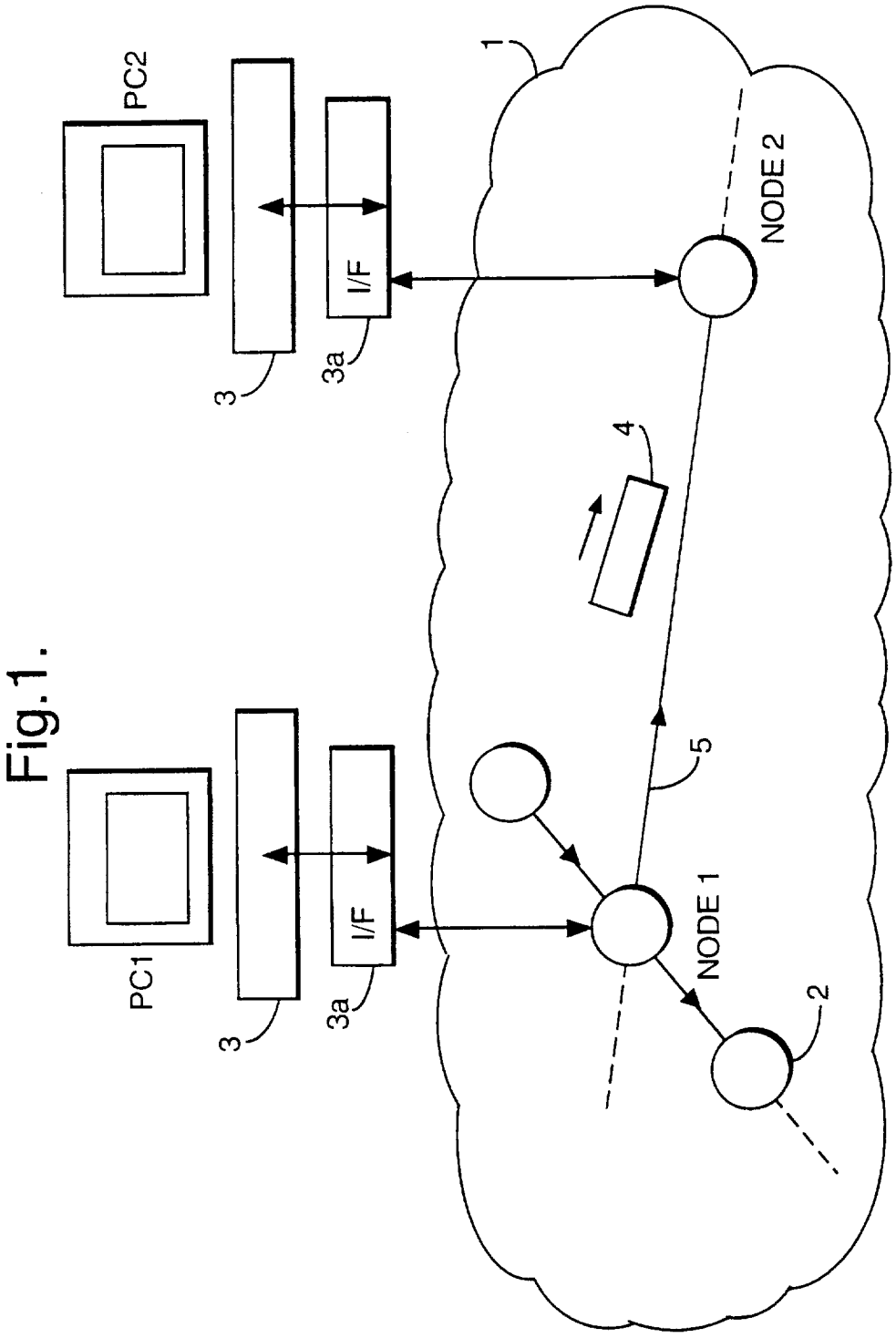


Fig.1.

Fig.2.

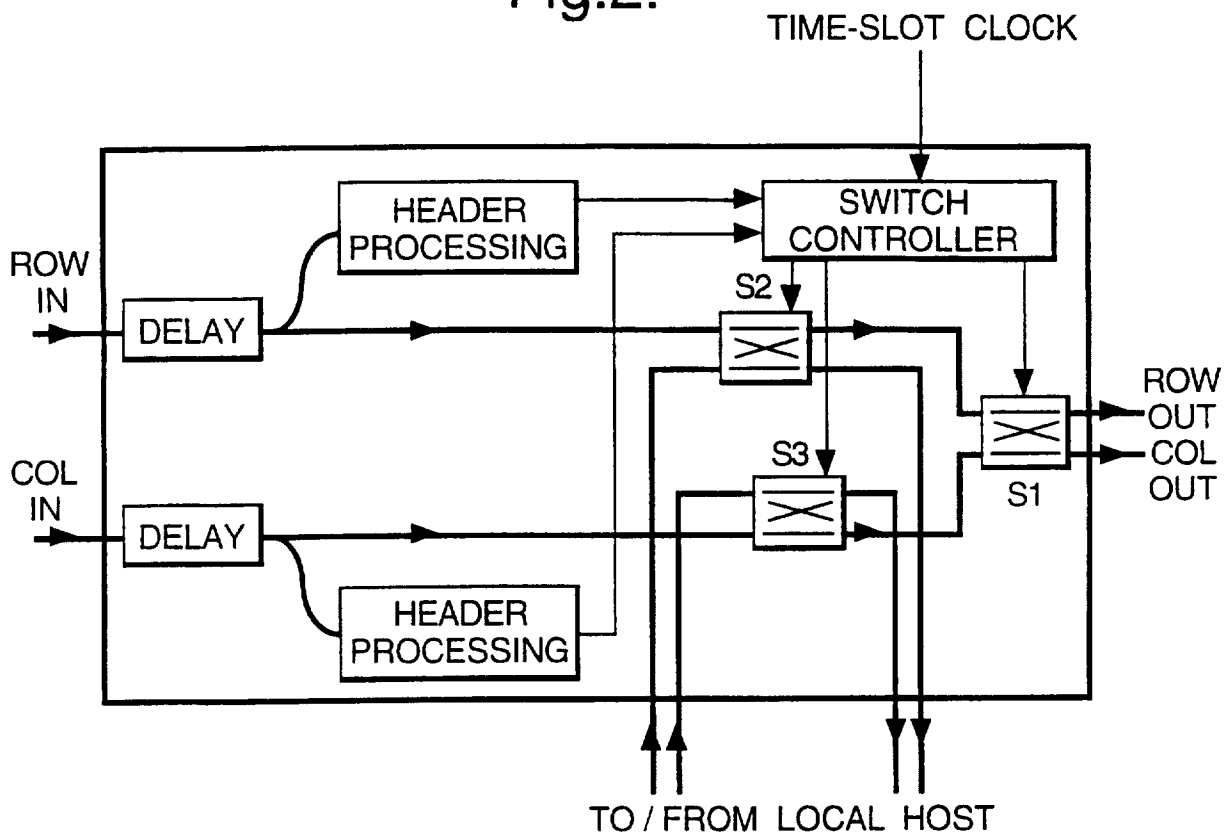

TIME-SLOT CLOCK
HEADER PROCESSING
SWITCH CONTROLLER
ROW IN
DELAY
S2
ROW OUT
COL OUT
S3
S1
COL IN
DELAY
HEADER PROCESSING
TO / FROM LOCAL HOST

Fig.3.

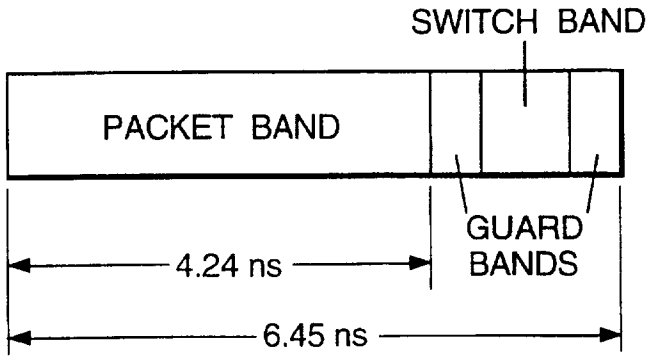

SWITCH BAND
PACKET BAND
GUARD BANDS
4.24 ns
6.45 ns

Fig.4.

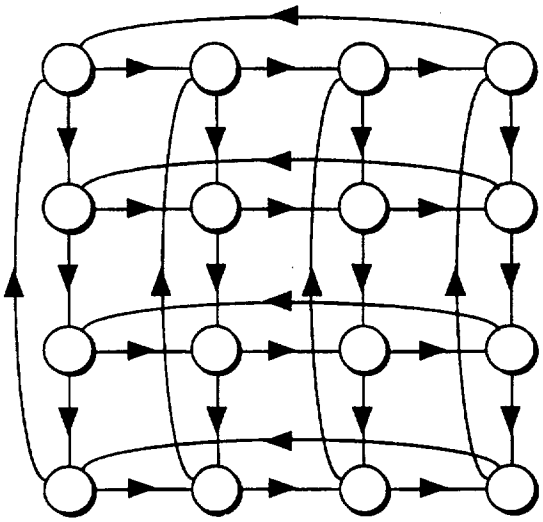


Fig.5.

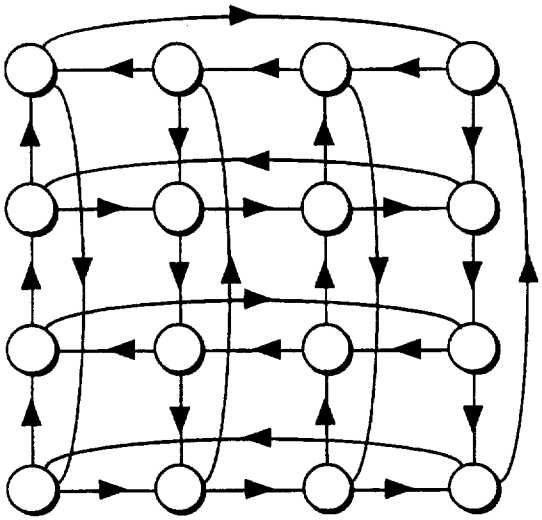


Fig.6.

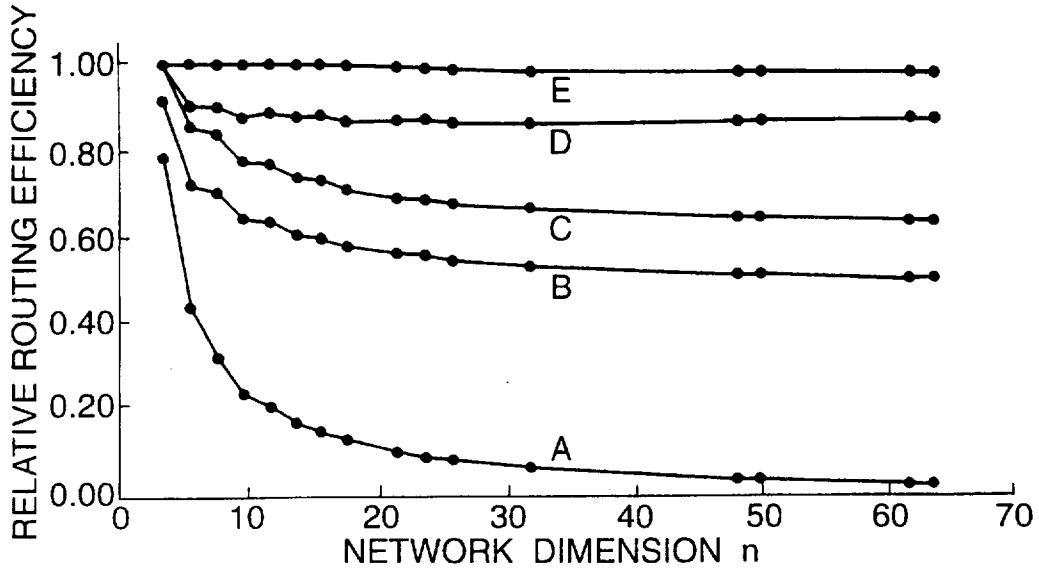

RELATIVE ROUTING EFFICIENCY
1.00
0.80
0.60
0.40
0.20
0.00
0
10
20
30
40
50
60
70
NETWORK DIMENSION n
E
D
C
B
A

Fig.7.

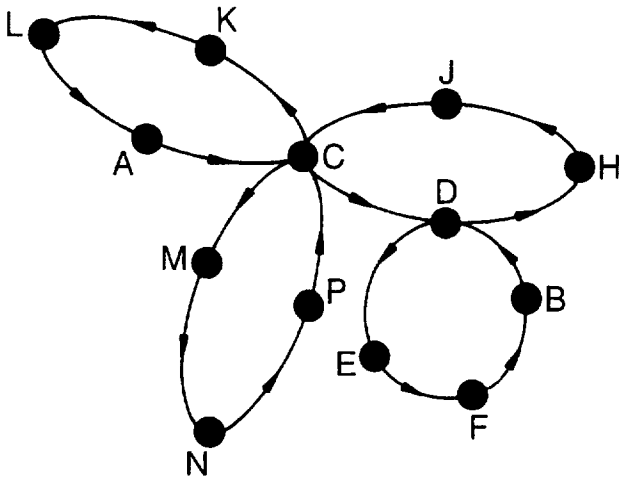

L
K
J
A
C
H
D
M
P
B
E
F
N

Fig.8(a).

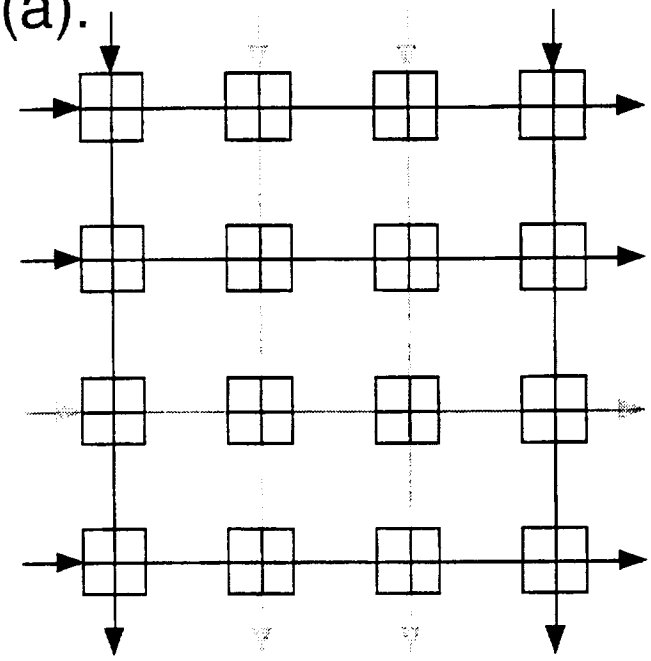


Fig.8(b).

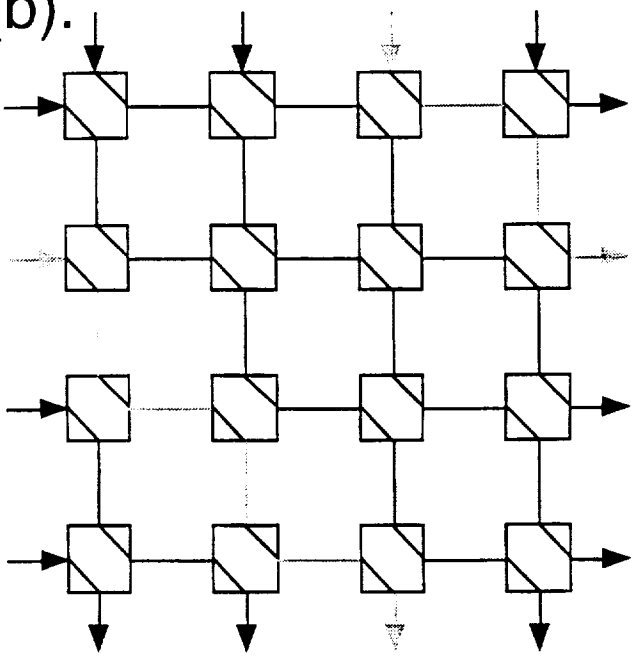


Fig.9.

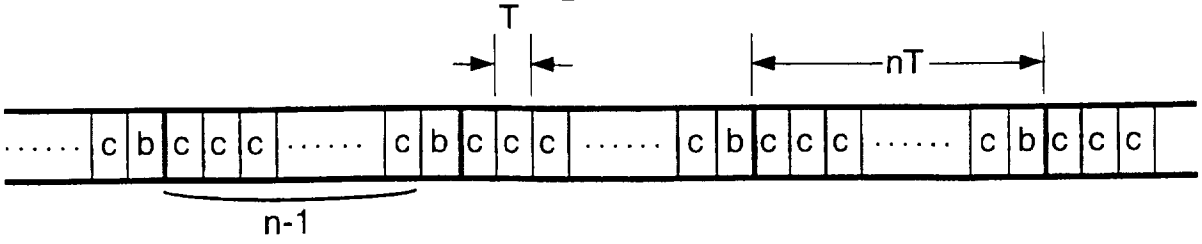


Fig.10.

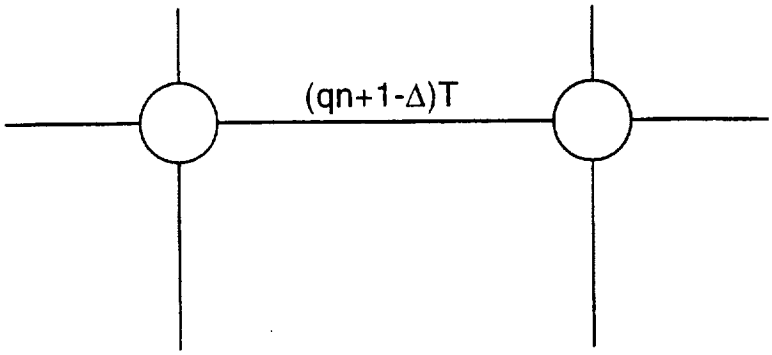


Fig.11.

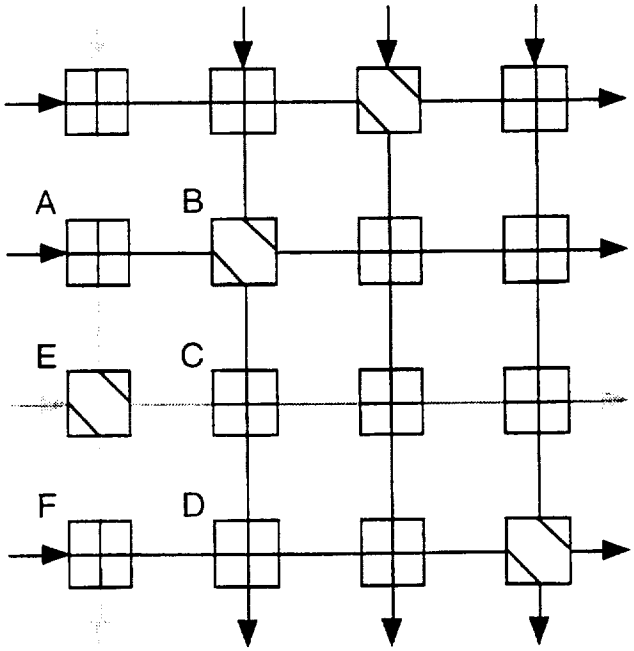


Fig.12.

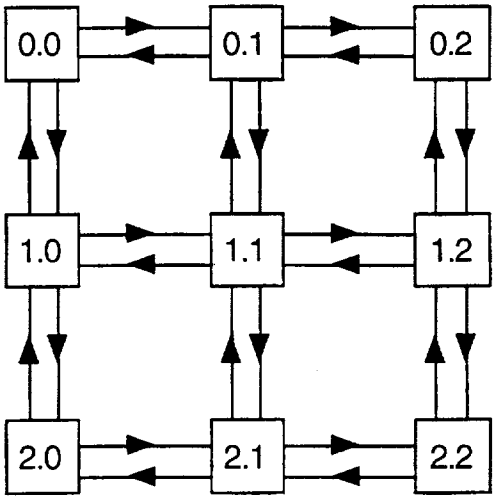


Fig.13.

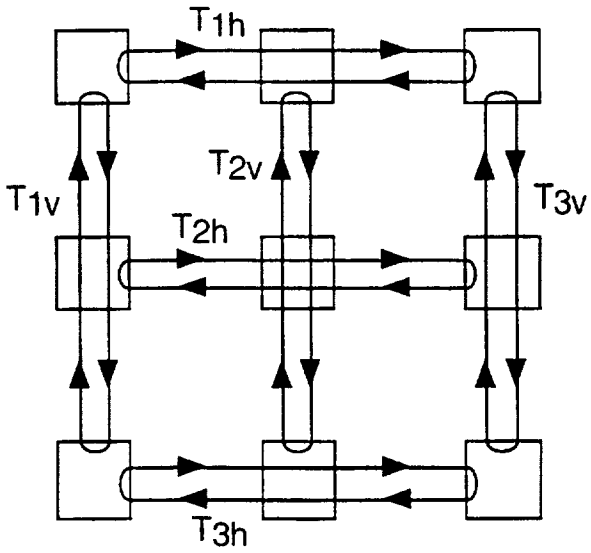


Fig.14.

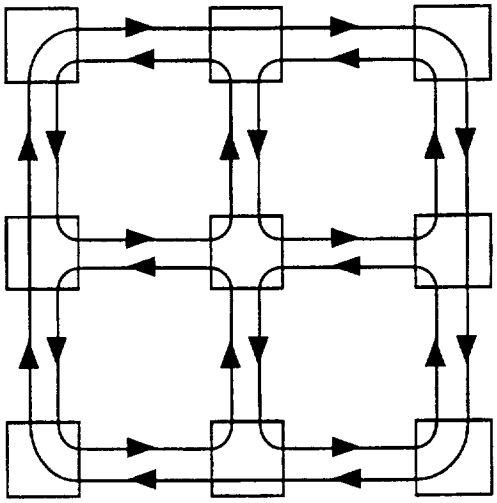


Fig.15.

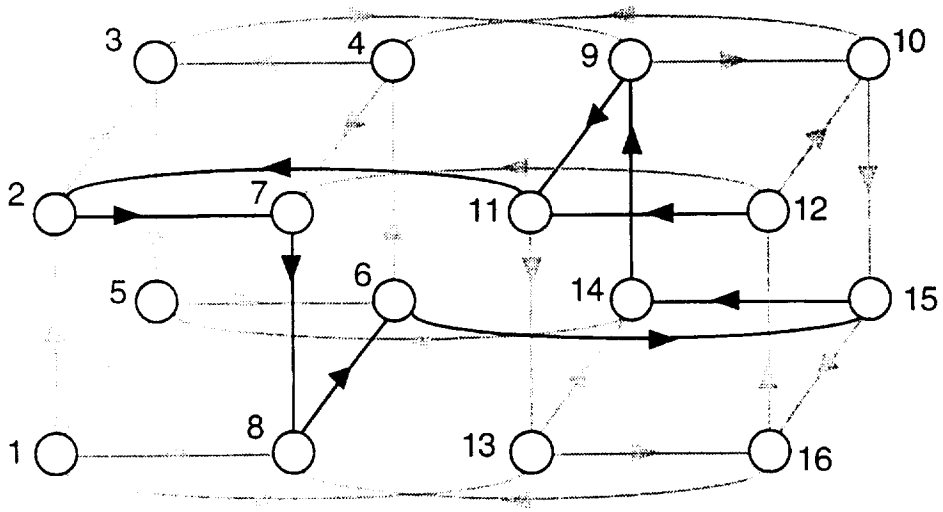


Fig.16.

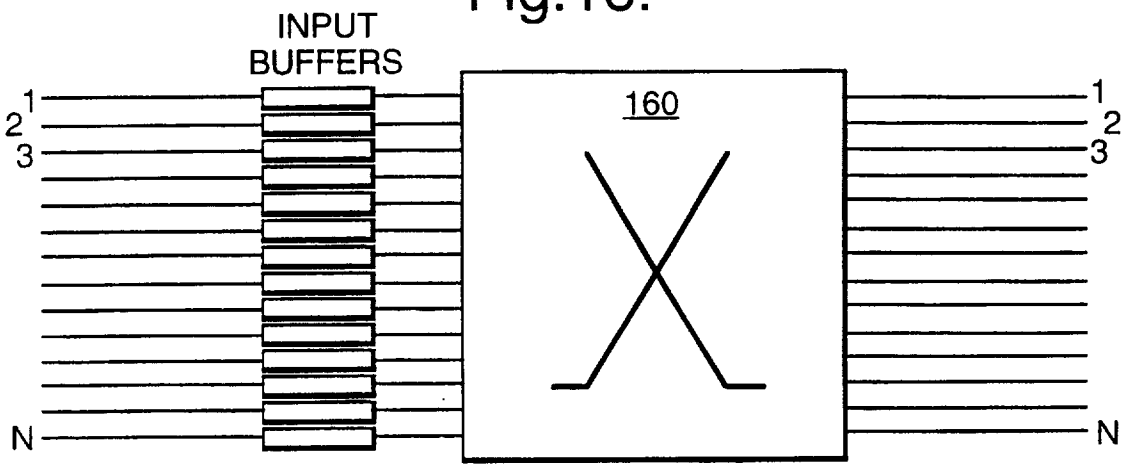
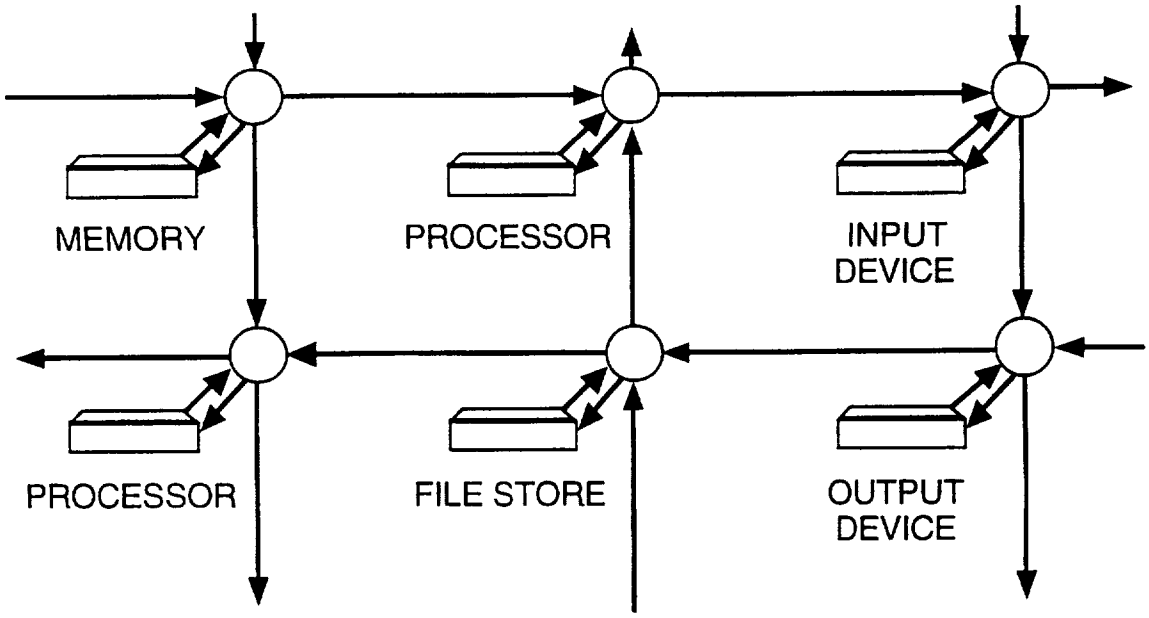

INPUT
BUFFERS
1
2
3
N
160
1
2
3
N

Fig.17.


MEMORY
PROCESSOR
INPUT
DEVICE
PROCESSOR
FILE STORE
OUTPUT
DEVICE