



(22) Date de dépôt/Filing Date: 1998/04/30

(41) Mise à la disp. pub./Open to Public Insp.: 1999/10/30

(45) Date de délivrance/Issue Date: 2003/01/21

(51) Cl.Int.⁶/Int.Cl.⁶ H04L 12/42

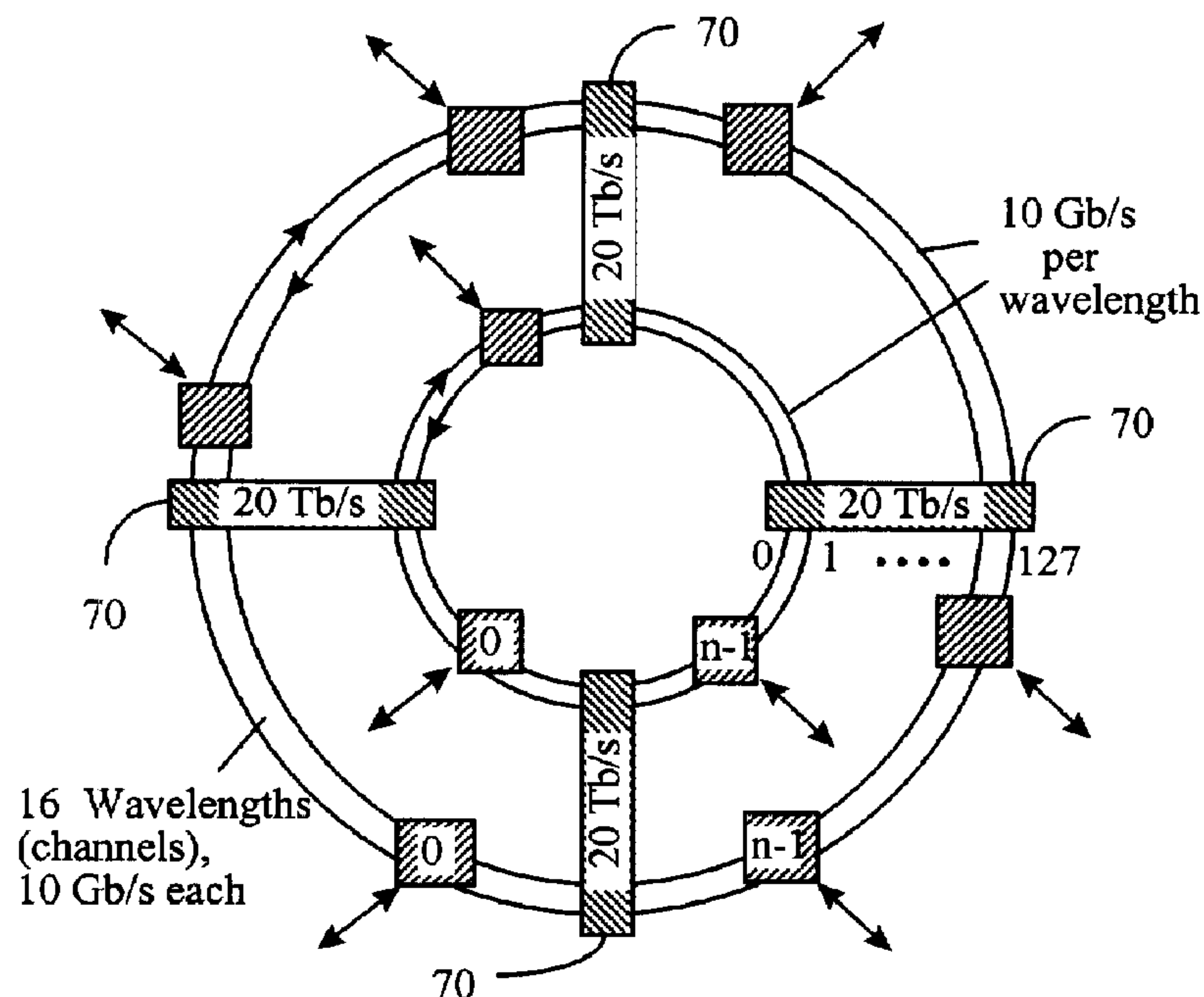
(72) Inventeurs/Inventors:
LIVERMORE, FREDERICK C., CA;
SKILLEN, RICHARD P., CA;
BESHAI, MAGED E., CA

(73) Propriétaire/Owner:
NORTEL NETWORKS LIMITED, CA

(74) Agent: GOWLING LAFLEUR HENDERSON LLP

(54) Titre : ARCHITECTURE DE RESEAU FONDEE SUR LE TRANSPORT PROGRAMMABLE

(54) Title: PROGRAMMABLE TRANSPORT AND NETWORK ARCHITECTURE



(57) Abrégé/Abstract:

This invention relates to the design of a scaleable high performance multiservice network based on programmable transport. A meshed network with dynamically-adjustable link capacities and nodes which provide data packing into "containers" for transport is proposed. A ring-based network, exchanging data containers among its nodes, is the preferred implementation, due to its flexibility, maintainability, and high reliability. With lossless rings, the quality of service is controlled solely by the origin and destination nodes, without any interference from other data streams. Flexible programmable transport greatly improves the performance, simplifies the controls, and facilitates scaleability. The concept is a departure from classical network thinking. By reducing the complexity of the network core, an economical, reliable, and manageable network with feature-rich edge nodes can be realized. An architecture with recursive ring-based structures provides a high degree of flexibility in bandwidth allocation and is compatible with current transport networks and future all-optical networks.



Title

5 PROGRAMMABLE TRANSPORT AND NETWORK ARCHITECTURE

Abstract of the Disclosure

10 This invention relates to the design of a scaleable high performance
multiservice network based on programmable transport. A meshed
network with dynamically-adjustable link capacities and nodes which
provide data packing into "containers" for transport is proposed. A ring-
based network, exchanging data containers among its nodes, is the
15 preferred implementation, due to its flexibility, maintainability, and high
reliability. With lossless rings, the quality of service is controlled solely
by the origin and destination nodes, without any interference from other
data streams. Flexible programmable transport greatly improves the
performance, simplifies the controls, and facilitates scaleability. The
20 concept is a departure from classical network thinking. By reducing the
complexity of the network core, an economical, reliable, and manageable
network with feature-rich edge nodes can be realized. An architecture
with recursive ring-based structures provides a high degree of flexibility
in bandwidth allocation and is compatible with current transport networks
25 and future all-optical networks.

PROGRAMMABLE TRANSPORT AND NETWORK ARCHITECTURE**Field of Invention**

5 This invention relates generally to efficient transfer of digital information among nodes. In particular, it is directed to a novel digital network in which digital information is exchanged among nodes in containers of any number of data units.

Background of Invention

10 To date, communication networks have been primarily used for voice services and support limited data and computer services. In recognition of the predominance of voice traffic, a circuit switched channelized architecture evolved and was optimized for telephony. Data and computer communications access and transport have been provided as an overlay on this channelized infrastructure.

15 Although many data technologies have been developed, the advent and subsequent popularity of the Internet, coupled with its ability to support multiple services, including telephony, is changing the way users communicate and do business. Today, telecommunications networks are shifting from specialized networks toward multipurpose multi-functional networks.

20 The emerging data network must be able to grow to a much-higher capacity than the capacity of today's voice and data networks. In addition to the huge-capacity requirement, the emerging networks must provide diverse and versatile services. The multiplicity of connection protocols, and the effort required for their interworking, inhibit the ability of the network to provide service diversity. The simplest network would be fully connected, allowing every
25 networking device to have a physical connection to every other networking device. However, as the network size grows, this fully-meshed structure rapidly becomes impractical. Due to the spatial variation of traffic loads, and the typically large modular sizes of transport links, a fully-meshed network normally leads to
30 underutilized transport facilities.

Currently, transport-capacity sharing is based on coarse-granularity, where point-to-point connections are defined as multiples of standard tributaries, such as OC3, OC12, etc. This renders tandem switching - at a lower granularity - necessary to establish end-to-end connections with an acceptable quality-of-
35 service (QOS) and an acceptable overall network efficiency. A low-connectivity

network with excessive tandem switching may however lead to an uneconomical network, due to the increased number of hops between origin and destination and the cost of processing in intermediate nodes. Tandem switching in the multi-protocol environment is rather complicated, and several techniques have been
5 proposed to reduce the processing at intermediate nodes by creating “short-cuts”, see for example, an article in IEEE INFOCOM, 1996, pp. 1251-1260, “Flow-Labelled IP: A Connectionless Approach to ATM” by P. Newman et al. In addition, tandem switching adds variable delay and increases latency.

The multiplicity of communications protocols and the need for coexistence
10 between new networks and legacy networks complicate the network planning function. In addition, QOS is difficult to realize in such a heterogeneous network. Currently, IP-based networks do not take the QOS into account, and one of the approaches to provide QOS assurance is to interwork - at a prohibitive cost - with an underlying switch-based network, such as ATM networks. The complexity and
15 fragility of such approaches are formidable impediments to network scalability.

Traditional transport systems can offer a meshed network by providing direct interconnections between the networking devices. However, the connections would be based on channelized time division multiplexing, where the bandwidth allocated to a node-pair is fixed and dedicated to the node-pair. When
20 the connection between two devices is inactive for some period of time, the transport bandwidth is still reserved so that it cannot be utilized by other active connections. Thus the networking device interfaces may not be as efficiently utilized as in the tandem approach which allows interconnections from many networking devices to share the interface to a given device.

25 U.S. Patent 5,293,376, issued March 8, 1994 (White), describes an upgradable telecommunication network which comprises a plurality of interconnected nodes or central offices, such as a SONET ring network. In its network, a unique controller enables a subscriber to change the central office or node in the network to which it is connected without changing the telephone
30 number of the subscriber location.

U.S. Patent No. 5,247,518, issued September 21, 1993 (Takiyasu et al), teaches a high speed ring LAN system in which SONET subframes flow in a time-divisional n-multiplexed format. The respective node devices inserted in the transmission path have one or more ports to accommodate sub-LANs or public
35 networks. Information is exchanged in units of a fixed-length container (packet)

between a received SONET subframe and an asynchronous port, whereas information is exchanged in units of a byte between the SONET subframe and a synchronous port.

U.S. Patent No. 5,406,401, issued Apr. 11, 1995 (Kremer) describes a technique of selective tributary switching in a bi-directional ring transmission system. The patent describes how selective switching in a bi-directional transmission system is used to realize selective tributary switching. A portion of bandwidth of a link that has been provisioned to be line-switched is governed by the set-up and take down procedures of full line-switching. The remaining bandwidth can be left unprotected but can be path-switched, because the line-switching functionality is combined with path-switching functionality in the same ring transmission system.

A fully-meshed network, as depicted in Figure 1, is not scaleable to cover a large number of nodes, unless the link capacities are elastic and can be modified rapidly to follow the traffic-demand variation. With fixed capacities and fluctuating traffic demand, the transport utilization drops rapidly as the number of nodes increases. An elastic network, however, would allow all the connections to share a common pool of capacity through paths whose capacities are dynamically adjustable.

The envisaged network to be described herein has a high-connectivity structure. The basic requirements of such a network are simplicity, scaleability, transport efficiency, the ability to accommodate existing legacy subnetworks, and - most importantly - high reliability. There are several candidate topologies which, primarily, fall into two main categories: cross-connection and ring sharing.

Both cross-connection-based and ring-based networks can be configured to be fully meshed or almost fully meshed. A cross-connection-based network would normally have a lower connectivity than a ring-based network and is not discussed in this disclosure. A ring structure lends itself to fine capacity partitioning with relatively simple controls. Ring sharing can be achieved in several ways; for example by using ATM nodes and SONET rings, as depicted in Figure 2. However, a flexible-transport layer as shown in Figure 3 would achieve a more economical solution. A path of controlled variable capacity for each node pair is established in the network, thus creating a flexible fully-connected network. Each path may carry multiple traffic classes, and the capacity of the path can be dynamically shared among the classes at the container-packing stage. The

network can accommodate a mixture of data, voice, and video traffic of both unicast and multicast nature. The multiclass service discipline is decided only at the originating and terminating nodes.

5 A ring structure with capacity partitioning allows many networking devices to be interconnected with efficiently utilized interfaces, without incurring the cost and service degradation of a tandem device. A domain is defined where every networking device within the domain is connected to every other networking device within it with fixed or variable capacity. All the connections within the domain share a common pool of capacity, maximizing the utilization of the device
10 interfaces. Since this new function is integrated into the transport network, the cost and service degradation associated with tandem networking devices is avoided. Various networking devices which use different protocols, such as ATM or IP, are accommodated by defining a container structure which carries digital information in its native form between them. The containers are carried on a
15 digital facility with a defined bit rate that circulates on a ring or virtual ring past every networking device in the domain.

The present invention extends the concept of the meshed networking based on a ring configuration described in the above referenced patent application. The meshed network of the invention allows all the connections to share a common
20 pool of capacity through links among nodes whose capacities are dynamically adjustable. Like the invention of the patent application referred to above, nodes provide data packaging into "containers" for transport and a ring exchanges data containers among its nodes but in the present invention, the containers may be of fixed or variable sizes. The invention calls for a service rate calculation for each
25 source-destination node pair which is carried out by a centralized or distributed controller. Such controller either monitors the traffic or receives updated capacity - allocation requests from the nodes, and assigns an appropriate data rate at which each node can transmit to each destination. With lossless rings (traffic - wise), the quality of service is controlled solely by the source and destination nodes, without
30 any interference from other data streams within the network. By reducing the complexity of the network core, an economical, reliable, and manageable network with feature-rich edge nodes can be realized.

In summary, a flexible programmable transport simplifies network management and extends the network capacity and network coverage by letting the end nodes control the QOS and end-to-end capacity allocation. Rather than forcing the network to cope with multiple protocols, node-pairs can communicate directly through adaptive end-to-end links. Thus, interoperability is replaced by protocol disengagement. The flexible capacity allocation can be realized in frame-based or container-based schemes. This technique is particularly attractive because of its inherent compatibility with the North-American fiber-network topology.

Objects of the Invention

It is an object of the invention to provide a flexible programmable transport network which realizes a fully meshed configuration, simple network management and easy capacity scalability and yet which is compatible with current transport networks and possibly with future all-optical networks.

It is a further object of the invention to provide a fully meshed network in which the end nodes in each pair control the QOS of the traffic between them according to end-to-end capacity allocations.

It is yet another object of the invention to provide a ring network in which, node-pairs can communicate directly through adaptive end-to-end links whose capacity is dynamically adjustable.

It is still an object of the invention to provide a fully meshed network in which traffic is exchanged among the nodes in containers of any desired number of data units.

It is an additional object of the invention to provide a fully meshed network in a ring configuration in which container scheduling is provided, sharing bandwidth at the topological level.

It is a further object of the invention to provide a fully meshed network in ring configuration in which centralized control and distributed control schemes can coexist.

5 **Summary of the Invention**

Briefly stated, in accordance with one aspect, the invention is directed to a communication network for transporting digital traffic among a plurality of nodes in a fully meshed logical configuration. The network comprises a dual transport ring for transporting digital traffic in opposite directions, the digital traffic
10 consisting of containers of digital signal, each container containing any number of data units and being destined to one or more nodes among the plurality of nodes. The network further includes the plurality of nodes located on the dual transport ring, each node extracting from the dual transport ring containers destined to it and inserting thereto containers destined to any of the remaining nodes and each node
15 controlling a rate at which it inserts to the dual transport ring the amount of data in the containers destined to any of the remaining nodes.

According to another aspect, the invention is directed to a node in a communications network which transports digital traffic among a plurality of nodes via a dual transport ring in a fully meshed logical configuration, the digital
20 traffic consisting of containers of any number of data units. The node comprises two ring buffers for storing traffic in the dual transport ring coming into the node and an add buffer for storing containers to be inserted into the dual transport ring. The node further includes a drop buffer for storing containers extracted at the node from the dual transport ring and a rate scheduler for scheduling insertion of
25 containers into the dual transport ring in accordance with an allowed rate approved by the communication network.

In accordance with another aspect, the invention is directed to a method of transporting digital traffic among a plurality of nodes located on a dual transport ring in a fully meshed logical configuration. The method comprises steps of
30 storing a plurality of data streams in a plurality of add buffers according to their destinations and acquiring an allowed data rate at which each data stream is to be inserted to the dual transport ring. The method further includes steps of extracting a properly addressed data stream from the dual transport ring, storing it in a drop buffer and inserting a plurality of data streams into the dual transport ring at
35 locations properly assigned to their destinations.

In accordance with a yet another aspect, the invention is directed to a method of scheduling digital data at a node in a telecommunications network in which the network is formed by a ring transport medium having a plurality of nodes and the digital data comprises a plurality of data streams. The method
5 comprises steps of storing the plurality of data streams respectively in a plurality of storage locations and calculating a parameter indicative of an accumulated rate of the data streams stored in each storage location. The method includes further steps of declaring each data stream eligible for transport when the parameter exceeds an allowed data rate of said data stream and transporting into the ring
10 transport medium each eligible data stream in accordance with a predetermined data rate.

Brief Description of the Drawings.

Figure 1 shows a fully meshed network.
15 Figure 2 depicts an existing network architecture.
Figure 3 is a conceptual view of novel network architecture according to the present invention.
Figure 4 shows a fully meshed network according to the present invention.
Figure 5 is the structure of a data stream according to one embodiment of the
20 invention.
Figures 6 and 7 are structures of data streams according to further embodiments of the invention.
Figure 8 shows schematically the construction of a node handling traffic in one direction according to an embodiment of the invention.
25 Figure 9 depicts a dual ring configuration of a domain.
Figure 10 shows a domain according to one embodiment of the invention having 80Tb/s capacity.
Figure 11 shows an expanded illustration of nodes and container switches.
Figure 12 shows intersecting domains.
30 Figure 13 shows a multi-domain configuration.
Figure 14 shows a node construction indicating capacity partitioning (asymmetrical node construction).
Figure 15 is a frame construction including M containers.
Figure 16 is a detailed illustration of the data structure according to one
35 embodiment of the invention.

Figure 17 shows an outgoing container allocation in successive frames.

Figure 18 is a table showing credit accumulation in time for several classes.

Figure 19 is a time scale diagram of the rate control mechanism applied to
5 a class of traffic with fixed-sized containers.

Figure 20 is a routing table at a node.

Figure 21 is a routing table at a container switch.

Figure 22 is a construction of a connection packet.

Figures 23 and 24 show mechanisms of path allocation.

Figure 25 shows a wavelength cross-section with electronic edge-switching
10 and tandeming.

Figure 26 shows a ring-based wavelength add-drop with electronic edge-switching and tandeming.

15 **Detailed Description of Preferred Embodiments of the Invention**

Figure 4 shows a basic configuration of a transport network according to the invention. Referring to the Figure, in the network 10, data is exchanged among nodes 12, e.g., five nodes, in fully meshed connectivity. The transport media in this description are rings and they are, in particular though not limited to,
20 fiber optic rings. As mentioned earlier, the domain is a digital facility with a defined bit rate consisting of a continuous series of digital data units and there can be more than one domain on the same level or different levels of hierarchy. The data units can be any arbitrary preset unit, e.g., a byte, and data is transferred in containers of any desired number of data units. Each node in the network is a
25 simple container packing and unpacking device. Each node is assigned a number of containers to carry its data to other nodes in the ring. The set of containers available to all nodes may form a frame. The frame size is arbitrary and is governed by granularity or data-delay specifications. The containers belonging to a given originating node are said to form a subframe, i.e., a frame contains a
30 subframe for each originating node.

Figure 5 illustrates one example of a signal structure 20 made up of a series of containers, each of which contains a header 22, a trailer 24 and payload data 26. The header and the tail have fixed numbers of data units which contain predetermined signal patterns to identify themselves. The header also includes an
35 indication (address) of a destination node and the length of its container (the

number of data units for example) They may include other fields for other purposes which may be desired. Other container and frame structures are possible and just two such structures are shown in Figures 6 and 7. In Figure 6, the signal stream is formed into a series of frames which are made of a variable number of containers. Each frame includes a header 30 and maybe a trailer 32 in some cases. Within a frame, containers are provided for each source node to insert any allowed number of data units destined to any of the destination nodes. In Figure 7, there are frame headers 34 and container headers 36. Each container header 36 identifies its destination node. The train of containers from a given source to a destination node is naturally concatenated at the destination node. Therefore, the container-packing process need not be aware of the data format and it need not limit the content of a container to an integer number of data transaction units, such as ATM cells, IP packets, or "circuit-switched" data blocks.

For simplicity of explanation and in aid of understanding the operation, however, either or both the container size and the frame size are fixed in some of the embodiments of the invention.

At least two control schemes are possible: one being based on processing a frame header, and the other on processing the headers of the individual containers. In the first, a node identifies the locations of its receivable payload from the information in the frame header. In the second scheme, each node reads the labels of all containers but copies only the relevant data. Some containers may be multicast containers. A multicast container is inserted once per ring but may be read by each node in the ring. The container duration is determined by the container size and the ring speed. With an OC192 (10 Gb/s) ring, for example, a 4096-bit container consumes 0.41 microseconds. A container's share of the ring capacity is the reciprocal of the number of containers per frame. Generally, a large number of containers per frame results in a better resolution at the expense of a higher delay. A frame of 1000 containers, of 4096 bits capacity each, has a period of 410 microseconds in an OC192 (10 Gb/s) ring, with each container representing 1/1000 of the ring rate; 10 Mb/s in the OC192 case. Increasing the size of the frame to 10000 containers increases the frame period to 4.1 milliseconds and each container represents a capacity unit of 1 Mb/s. A large frame period, however, is undesirable, and with appropriate container scheduling, it is possible to realize a very high resolution using relatively short frames.

Referring back to Figure 4, each node in the network is a simple container packing and unpacking device which inserts or removes a container to/from the signal stream traveling in the ring. According to one embodiment, one of the nodes within the domain generates and allocates the multiplex containers, and indicates the length and alignment of each multiplex container by generating a multiplex container label (identifier). This node is referred to as the domain controller node. Other domain nodes can assume this responsibility in the event of failure of the original domain controller. The domain controller node monitors the condition of the ring and calculates rates to be given to each node pair. The domain controller node communicates from time to time with all the nodes in a domain or domains by special controller containers informing each node topological information of the domain or domains, e.g., routing information between any node pairs, and a maximum number of data units in a container the node is allowed to transfer from one specific node to another on the ring. The allocation can be changed from time to time or from one container to the next using a simple rate controller which facilitates precise rate allocation for each node pair. The domain controller node facilitates precise rate allocation for each node pair.

Figure 8 is a functional schematic illustration of a node in one direction of a dual ring. The node of course includes the same arrangement for traffic propagating in the opposite direction. A signal stream 40 in the ring is buffered at a ring buffer 42. A node controller 44 monitors the header/tail and decides to send the container to a drop buffer 46 or to an output port 48 to the next node on the ring. The node controller 44 diverts to its drop buffer only those containers addressed to it and passes all other containers on to the next node. While the container is being diverted to the drop buffer, a container stored in an add buffer 50 is being inserted to the outgoing channel. The add buffer 50 in fact has a plurality of queues for a plurality of classes or a plurality of destinations or both. The traffic being inserted is separately stored in a separate queue and a scheduler 52 (queue controller) enables one queue at a time for inserting a container therefrom. The container being inserted is addressed to another node and the amount of data units that can be inserted is either governed by the rate allocated by the domain controller node or determined according to a distributed rate allocation scheme. The container being added is not necessarily in the same size as the container being dropped. In fact, they may be most of the time different from one

another. If the container being added is properly rate controlled but is larger than the container being diverted, the signal stream is stored in the ring buffer to create sufficient space for the container being added.

5 The rate control techniques of a shared medium can be effectively used in the present invention.

The generic problem of service rate control may be stated as follows:

10 A link of fixed capacity is shared by a number of traffic streams which are allocated separate queues. The streams may share the link according to one of many policies, such as a fixed rate per stream, or a guaranteed minimum rate per stream. In any of these policies, the maximum number of cells that a queue may transfer to the link in a single transaction (i.e., in a contiguous period) may be limited to a predetermined bound. Two rate control disciplines, are described in detail, the first being the fixed service rate per class, and the second the guaranteed minimum service rate per class. The former applies to
15 delay-sensitive traffic and the latter applies to delay-tolerant traffic.

While the techniques described above uses the ATM cell as the basic data unit, it should of course be noted that any fixed units can be used for the above rate controlling techniques.

20 Within the fixed service rate per class controls, two equivalent mechanisms are possible; controlling the sampling rate per class or controlling the sampling interval per class. While these two options of frequency and time domain controls are functionally equivalent, the hardware implementations can differ substantially. The frequency domain control schemes are suitable for a small or medium number of classes. For a large number of classes, several
25 hundreds for example, time-space mapping associated with time domain control schemes are preferable. The above referenced pending patent applications describe these mechanisms in detail. A technique of scheduling multiple streams of ATM traffic from a node calls for a link controller which scans a time-space map to
30 determine when a given stream will be sampled and its cells transmitted from the memory out over a link. The time - space map comprises time-slots whose spatial positions within the map correspond to times at which cell can be transmitted.

Therefore, the node controller 44 shown in Figure 8 may be provided with service rate controller which governs the operations of the scheduler 52 and add buffer 50. As mentioned above, the service rate controller may be of the type of either the frequency domain control or the time domain control.

5

Ring capacity

A single-channel ring constitutes a basic control domain. A basic ring interconnects a number of add-drop multiplexers which share a single unidirectional channel. Generally, the term channel is used to indicate a tributary or an entire wavelength in a single fiber. In this specification, the term refers to a wavelength. The total capacity per channel in a ring of n nodes, i.e., the maximum carried traffic by the channel, depends on the capacity R of the channel (e.g., 10Gb/s) and the spatial distribution of the traffic demand (the traffic distribution among the end nodes). Assuming a uniform distribution of the spatial traffic demand the capacity limit of a unidirectional ring is $C=2R$. Using a ring pair, comprising a clockwise ring and a counter-clockwise ring, as shown in Figure 9, routes with more than $n/2$ hops can be avoided, and the capacity C per channel increases to: $C=4nR/(n+1)$, if n is odd, or $C=4(n-1)R/n$, if n is even, under the same assumptions of spatially-balanced traffic demand. Thus, the capacity of a dual ring is almost double the capacity of two unidirectional identical rings. Using dual rings enhances the capacity per wavelength by a factor of approximately two.

To realize a ring (domain) capacity of the order of tens of Tb/s (Terabits/sec), it is necessary to use multiple fibers each supporting multiple wavelengths. Figure 10 illustrates such an arrangement. The channels of a domain must be interconnected at least at one point. Using M fibers, each supporting W wavelengths, i.e., MW channels, at a bit rate of R Gb/s per channel, yields a total capacity of more than RMW b/s. The MW channels must exchange payload data through one or more container switches. For example, with $M=128$, $W=16$, and $R=10$ Gb/s, container switches of 20Tb/s capacity each would be needed. The domain of Figure 10 has 128 fiber rings, each supporting 16 wavelengths, with 10Gb/s per wavelength. Four interconnecting container switches 70, topologically evenly spaced, are used. The domain capacity is not a linear function of the number of container switches, and it is sufficient to use four container switches per domain, each interconnecting all the channels of the domain, to realize almost the maximum attainable capacity for a given set of

single rings. In the domain of Figure 10, the maximum capacity is about 80Tb/s (only by coincidence this equals the sum of the capacities of the container switches). It is noted that using more than four 20Tb/s switches results in an insignificant increase in the overall capacity.

5 It may also be desirable to partially connect some of the rings, using intermediate-capacity container switches, in order to shorten some end-to-end paths as shown in Figure 11. The channels are arranged in dual pairs (clockwise/counter clockwise). The figure illustrates the inter-connection of all the rings by 20Tb/s switches 76, and also shows smaller container switches 78
10 interconnecting dual rings.

 The domain definition may be extended to include several parallel single channels, as in Figures 12 and 13. Extending the network capacity is realized by increasing the domain capacity, and extending the network coverage is realized by using several intersecting domains. A domain may share common nodes with one
15 or more domains. A node in a given domain sends the traffic to a neighboring domain through one or more common nodes. A domain may also function as a transit medium between two non-adjacent domains. A common node between domain X and domain Y merges the contents of all containers received from the domain X nodes and forms new containers to the domain Y nodes, and vice versa.
20 In this process, there are no surprises: the rate controllers at the network edge ensure that the exchange process leaves no residues, hence no buffering - apart from what is needed for the exchange process - is required. The common nodes, connecting domains, perform the extra task of containers merging.

 A node in Figure 9 is an add-drop multiplexer and interconnects a dual
25 ring and local subtending edge devices. Figure 14 depicts one example of the node in which the node has a total input plus output capacity of 60 Gb/s. In a symmetrical node, the input and output sides have 30 Gb/s capacity each. As shown in the Figure, the total capacity may, however, be divided unevenly (asymmetrical node) between input and output in order to accommodate multicast
30 traffic without wasting switch-core capacity, since in multicasting the data is inserted once but may be delivered to several destinations.

 As mentioned earlier, a very-high capacity container switch is needed to interconnect the channels in a WDM multi-fiber ring. Large container switches, with capacities ranging from 40 Gb/s to 20 Tb/s may be realizable using multiple
35 WDM links and a rotator-based core architecture. U. S. Patent No. 5, 745,486

issued on Apr. 28, 1998 to Beshai et al describes the rotator based core architecture. Several alternate realizations are possible. The simplest may be an all-electronic switch core. Other “mostly-optical” solutions may also become economically viable. Multi-dual-channel add-drop modules are advantageous in
 5 reducing the route lengths.

Determining the network layout, connectivity, and capacity allocation in a wide-coverage multi-domain network is a challenging task. Optimization procedures are required for planning and provisioning purposes. These procedures are often computationally-intensive, and, fortunately, they need not be performed
 10 in real time. On the other hand, fast topological algorithms are needed to respond (in milliseconds) to network-state changes. The topology design is influenced by the path-routing process, and vice-versa. This interdependence offers a good opportunity for network optimization.

15 **Control**

In order to explain the main features of the control system, the discussion is limited to a single domain of a single channel. The control may be either frame-based or container-based.

Frame-based control

20 The process is explained with the help of a numerical example. The container size, in this discussion, is assumed to be fixed at 4096 bits. To realize high-resolution end-to-end capacity allocation with independent successive frames, for example in units of 1 Kb/s, the frame length (the number of containers per frame) may be prohibitively large. However, using a remainder and carry-over
 25 process, the frame length can be selected to be only about $n(n-1)$, n being the number of nodes per ring; typically not exceeding 16. It is important to note that an end-to-end capacity of an arbitrary value may be allocated. For example, if a node-pair requires 3.6 Mb/s, it is allocated one container every 10.82 frames, in a 256-container frame, with a channel capacity, R , of 10 Gb/s. The capacity
 30 allocation for a node pair is determined as: $C=R/S$, where S is the number of time slots between successive containers for the node-pair. A single container is carried during a single time slot. Thus, in this example, $S=2770$, i.e., 10.82 frames. The frame period is $T=KL/R$, where K is the number of containers per frame and L is the number of bits per container. In an example shown in Figure
 35 15, $K=256$, $L=4096$, $R=10\text{Gb/s}$, and $T=105\text{msec}$.

The nodes may be of different capacities. Assuming equal-capacity nodes, each node is assigned the same number of containers which may be distributed unevenly according to destination. The containers from a given node will naturally be separated by containers originating from other nodes. Figure 16 shows the destinations of containers emanating from node 0; the shaded areas in the figure refer to the time slots granted to originating node 0, i.e., the subframe of node 0. The container allocation may differ between consecutive frames for two possible reasons: change in traffic demand, and/or the allocation of a non-integer number of containers per path per frame, since a non-integer allocation necessitates fraction carry-over between successive frames.

It is noted that the unused time slots in the control frame can be granted to "best-effort" traffic streams, which make no capacity reservation, probably according to a weighted-priority discipline. A frame header identifies each container according to destination. Unassigned containers can be seized without reservation. Several techniques which attempt to maximize the transport efficiency, with predefined fairness, can be used.

Figure 17 depicts the change in allocation in a subframe, over a period of four successive frames, to accurately represent non-integer allocations. The containers in this example have a fixed size. Though the allocated containers for a node-pair appear to be contiguous in subframe of Figure 17, they are in fact well separated in the time domain. The number of containers from node 0 to node 2 during the above four successive frames are 2, 1, 2, and 1, with a mean value of 1.5 containers per frame. Similarly, the allocation for node-pair 0-7 are 2, 3, 2, and 2, with a mean value of 2.25 containers per frame. An arbitrary non-integer number of containers may be allocated for any node-pair.

It is worthwhile to mention that, for low-speed applications, containers of very-small sizes may be used and even byte-interleaving may be appropriate.

Container-based control

In a container-based system, the added load at each node is rate regulated with an allocated service-rate to each destination. Therefore referring back to Figure 14, rate control module 90 is provided on the add path of the node. The access capacity is shared by a number of traffic streams, identified by destination, which are allocated separate queues at ring access. A stream is identified by its origin and destination and/or class of service. A stream is allocated a direct end-to-end path the capacity of which may vary with time. The streams may share the

ring capacity according to one of many policies, such as strict priority, a fixed-rate per stream, or a guaranteed minimum rate per stream. The capacity allocation per stream may be updated individually as the need arises. Figure 18 shows the division of the stream capacity among four traffic classes according to the normalized capacity allocations 0.29, 0.21, 0.30, and 0.20. The first stream is allocated 0.29 of the link capacity, the second is allocated 0.21 of the channel capacity, and so on. In this example, the containers are assumed to be of a fixed size. In Figure 19, Δ represents the normalized capacity allocation for the stream under consideration, δ is the duration of the data unit, and β is the container size in data units. During each δ , a given stream is credited by its normalized capacity allocation Δ (in the above example, Δ is 0.29 for the first stream). An accumulator for each stream keeps track of an accumulated credits. A stream which accumulates a credit of β or more, becomes eligible to transmit a container. The remainder, if the credit exceeds β , is retained by the class. If the credit Δ is expressed as a fraction of the channel capacity, the value of β is unity. An entry in Figure 18 indicates the credit available to the corresponding class, and a class is ready for access when its credit is unity or more. A shaded entry corresponds to the time-slot at which the class is entitled to network access. A selector selects a stream among a plurality of eligible streams based on a preset criterion, e.g., round robin. The above description explains the mechanism of rate control.

Path-allocation for a dual-ring

In the simplest path-allocation form, each stream is allocated a single end-to-end path which may traverse more than one ring. It may be desirable, however, to have two or more paths per stream for reliability as well as even distribution of the traffic intensity across the network. The end-to-end routes may be separated between clockwise and counter-clockwise rings in such a way as to balance the traffic loads on individual inter-nodal links and optimize the overall throughput. Path selection for multiple rings is more elaborate. It should aim at reducing the number of hops traversed from origin to destination, with a view to global optimality.

Container scheduling

For a lossless network, it is of paramount importance that the data rate per node-pair be time regulated. A domain scheduler performs this function using

frequently-updated traffic and topology information. It is important to remember that, due mainly to propagation delay, the traffic input to the scheduling process may differ from the actual traffic waiting at each node at the scheduling time. To compensate for the time-delay, real-time traffic characterization and projection
 5 may be applied at each node.

In a single domain, a multicast container is transmitted once and received by several nodes, i.e., the number of received containers exceeds the number of transmitted containers. The scheduling function should attempt to exploit this asymmetry in order to maximize the overall throughput. In a multi-domain
 10 network, a multicast container propagates from one domain to neighboring domains through common nodes.

Referring to Figures 20 and 21, mechanisms for paths allocation and container scheduling according to one embodiment of the invention will be described. The domain controller uses topology information to compute and
 15 deliver a routing table for each node and a routing table for each container switch. The nodal routing table is shown in Figure 20 and has N entries, N being the number of nodes. The container-switch routing table is shown in figure 21 and also has N entries. The nodes in the domain are numbered sequentially starting from zero to N-1. A node connects a clockwise channel, a counterclockwise
 20 channel, and the add-drop links. A container switch may interconnect several dual channels. These routing tables at nodes and container switches are generated and/or updated using connection packets sent by a domain controller node.

Referring to Figure 20, each entry in the nodal routing table has two fields. The first field 110 is 1-bit wide and is used to specify the direction of the route.
 25 Arbitrarily, a value of 0 is selected to indicate the counter-clockwise direction while a value of 1 indicates the clockwise direction. The second field 112 is 4-bits wide and is used to store the number of hops towards the destination along the selected channel.

In Figure 21, each entry in a container switch has two fields. The first field
 30 210 is 12-bits wide and is used to store the outgoing channel number for a given destination. The second field 212 is 4-bits wide and is used to specify the number of hops towards the destination along the channel specified in the first field.

When an originating node X determines that it needs a specific capacity C to destination node Y, it prepares a connection packet shown in Figure 22. The
 35 connection packet has a fixed header of three fields, and a variable-length route-

trace field. The first field in the packet header stores the number of hops towards the destination or the next container switch. The second field stores the destination number. The third field stores the required capacity C .

The originating node reads entry number Y (the node number of the destination) from its routing table. The first bit of the read data determines whether the connection packet is to be routed in the clockwise or counterclockwise direction. The number of hops, H , stored in the 4-bit field of the read data is entered in the header of the connection packet. As the connection packet moves in the specified direction, each node it visits decreases H by 1. The node that brings the value of H down to zero compares the destination number Y in the packet header with its own node number. If they are equal, the path allocation is complete. Otherwise, the connection packet moves one more hop. The next point must be a container switch. The identity of the container switch need not be known to the connection process. The container switch thus encountered reads entry number Y from its routing table. The first field in this routing-table entry is the outgoing channel number. If the field is 12-bits wide, the maximum number of channels is 4096 (2048 channels in each direction). The second field in the routing table entry is the number of hops that the connection packet should scan around the specified channel. This number is written in the first field in the connection-packet header (the H field), replacing the previous value of zero. Again, each node along the route decreases H by 1, and if the node which brings H to zero is the destination, the route is determined. Otherwise, the next hop must lead to another container switch which will implement a similar procedure. The visited nodes are identified in the route-trace field of the connection packet.

Of course, with each hop, the node compares the free capacity of the channel with the required capacity C . Each node decreases the value of the channels free capacity by the value C . Failure to accommodate the connection in any hop terminates the search and releases the reserved capacity, i.e., increases the free capacity in the encountered hops by the value C , according to standard procedures. The capacity release is done with the help of the route-trace of the connection packet.

Figures 23 and 24 illustrate a few example path allocation processes. In the Figure 23, nodes are designated 320 and container switches 340, 342, 360 and

362. Channels are also shown as 0000 to 0823 in both figures. For some paths, the connections are set up by following processes:

From node to node	Connection path process
A to B	1.2
A to C	1.2/0000.3
E to F	1.1/0820.2/0820.2
F to E	0.1/0823.2/0823.2
G to H	0.1/0823.2/0823.2
J to K	0.0/0820.2
H to G	1.1/0822.2/0000.2

5 From node A to node B, the connection packet entry is (1.2), signifying the clockwise propagation for 2 hops. From node E to node F, the entry is (1.1). This indicates a clockwise propagation of 1 hop from node E. After one hop, the number of the encountered node does not match the destination node number and therefore this results in one more hop to container switch 340 which consults its
10 own routing table and routes the packet onto channel 0820. The switch sends its own packet with a hop count of 2 onto outgoing channel 0820. After 2 hops, the hop count is zero and the destination is not reached. This again results in another container switch 342 which sends its connection packet with a hop count of 2 onto channel 0820, meaning two more hops on the same outgoing channel. In the
15 above example, the J-K connection, node J reads (0.0) from its routing table. The first value of 0 indicates a counterclockwise direction. The second value of 0 indicates that node J is connected to a container switch.

 Finally, the connection rate for a stream must be acceptable to the drop link of the receiving node.

20 As mentioned frequently above, the objective is to control the QOS at the edge since this avoids the complexity of having to deal with it at intermediate nodes. When end-to-end bandwidth demand can not be accommodated for all node-pairs, distinction based on some class or other criteria is required for acceptance of new capacity demands. It is noted that QOS may have several
25 interpretations. The network architecture of the invention allows the QOS to be defined on a node-pair basis.

As described earlier, the capacity of a domain can be extended using multi-wavelength multi-fiber interconnected parallel rings. The network's reach is realized by multiple intersecting domains. Figure 12 illustrates the case of a central domain interconnecting four partially inter-connected side domains.

5 A network of wide coverage, extending to hundreds of nodes, can be arranged in a multi-level structure as illustrated in Figure 13. A four-level structure with 10 nodes per domain realizes a network of about 10,000 nodes, each of which may accommodate several subtending nodes. The use of lateral rings interconnecting side domains (not shown in Figures 12 and 13) reduces the
10 amount of traffic in the central domain and may also reduce the path length. It is noted that the network rings need not be of equal capacity. When two or more paths are available for a node pair, a given connection must follow the same path in order to maintain container-order.

Unprecedented traffic growth is providing a huge demand for fiber
15 facilities. Not only are the backbone networks outstripping their original design capacities, but the routes will need fiber cable replacements to allow full buildup of high-density WDM. In addition, as the routes grow, there will be further pressures to create better diversity to improve restoration. Currently, WDM is primarily used to increase point-to-point transport capacity. The abundance of
20 transport bandwidth due to WDM may justify a highly-meshed topology at wavelength granularity. However, due to the spatial traffic variation and with wavelength switching, a wide-coverage network may still require tandem switching and capacity-sharing controls, which are now realized electronically. Two dimensional space-WDM switching nodes may be used to realize, at a
25 wavelength granularity, either a partially-interconnected network or a fully-meshed network. Figure 25 depicts an optical cross-connect configuration, and Figure 26 depicts a ring configuration, both using space-WDM switching nodes. In either configuration, traffic consolidation and reinsertion at the edges must be done by electronic means to realize arbitrary granularity, i.e., the edges perform
30 the add/drop as well as tandem switching. The path capacity may then be defined in multiples of Kb/s rather than multiples of 2.5 Gb/s (OC48), for example. It is possible with such configurations, using electronic edge switches, to construct a fully-meshed network with fine granularity and realize the benefits of edge control. The control mechanism would differ from that of the ring architecture of
35 Figure 10.

What we claim as our invention is:

1. A communications network for transporting digital traffic among a plurality of nodes in a fully meshed logical configuration, comprising:
 - 5 a dual transport ring for transporting the digital traffic in opposite directions, the digital traffic consisting of containers of digital signal, each container containing any number of data units and being destined to one or more nodes among the plurality of nodes;
the plurality of nodes located on the dual transport ring, each node
10 extracting from the dual transport ring containers destined to it and inserting thereto containers destined to any of the remaining nodes; and
each node controlling a rate at which it inserts to the dual transport ring the amount of data in the containers destined to any of the remaining nodes.
- 15 2. The communications network for transporting digital traffic among a plurality of nodes according to claim 1, further comprising:
 - a domain controller node for collecting topological information about nodes on the dual ring;
generating a routing table built using the topological information; and
20 distributing the routing table to the nodes.
3. The communications network for transporting digital traffic among a plurality of nodes according to claim 2 further comprising a plurality of dual transport rings;
25 wherein:
 - some nodes are located on some of the dual transport rings; and
at least one intersecting node intersecting some of the dual transport rings for transferring containers among the dual transport rings.
- 30 4. The communications network for transporting digital traffic among a plurality of nodes according to claim 3 further comprising:
 - two or more groups of dual transport rings, each having a domain controller; and
a plurality of intersecting nodes which intersect two or more groups of dual
35 transport rings:

wherein each intersecting node packs and unpacks containers from one group to another according to the structures of data streams of each group.

5. The communications network for transporting digital traffic among a plurality of nodes according to claim 2 wherein one of the nodes on the dual transport ring performs the functions of the domain controller.

6. The communications network for transporting digital traffic among a plurality of nodes according to claim 2 wherein each container has a header which identifies the destination node to which the container is destined.

7. The communications network for transporting digital traffic among a plurality of nodes according to claim 6 wherein all the containers are uniform in size.

8. The communications network for transporting digital traffic among a plurality of nodes according to claim 2 wherein the domain controller further comprises a network management device for monitoring the status of the dual transport ring and calculating, based on the monitored status, an allowed data rate of each node, up to which rate each node can insert containers destined to any of the remaining nodes,.

9. The communications network for transporting digital traffic among a plurality of nodes according to claim 1 wherein the dual transport ring comprises fiber optics.

10. The communications network for transporting digital traffic among a plurality of nodes according to claim 9 wherein the dual transport ring comprises fiber optics capable of supporting multiple wavelengths.

11. In a communications network for transporting digital traffic among a plurality of nodes via a dual transport ring in a fully meshed logical configuration, the digital traffic consisting of containers of any number of data units, a node comprising:

two ring buffers for storing traffic in the dual transport ring coming into the node;

an add buffer for storing containers to be inserted into the dual transport ring;

5 a drop buffer for storing containers extracted at the node from the dual transport ring; and

a rate scheduler for scheduling insertion of containers into the dual transport ring in accordance with an allowed data rate approved by the communication network.

10

12. The node in a communications network, according to claim 11, wherein the rate scheduler comprises:

an accumulator for summing a credit for the containers to be inserted and declaring them eligible for insertion when the accumulated credit exceeds a
15 predetermined data rate.

13. The node in a communications network, according to claim 12, wherein the containers to be inserted at a node are classified to one or more classes of containers and the rate scheduler comprises;

20 a plurality of accumulators, each accumulator for summing a credit for a class of containers and declaring the class of containers eligible for insertion; and
a selector for choosing one eligible class at a time for insertion to the dual transport ring.

25 14. The node in a communications network, according to claim 13, wherein the classes are defined by categories of traffic descriptors or destination of containers or both.

15. The node in a communications network, according to claim 11 wherein the
30 allowed rate is defined by a normalized service rate allocation.

16. The node in a communications network, according to claim 11, wherein the add buffer and the drop buffer have a maximum rate of insertion and extraction respectively which are different from one another.

35

17. A method of transporting digital traffic among a plurality of nodes located on a dual transport ring in a fully meshed logical configuration, each of the nodes comprising steps of:

- 5 storing a plurality of data streams in a plurality of add buffers according to their destinations;
acquiring an allowed data rate at which each data stream is to be inserted to the dual transport ring;
extracting a properly addressed data stream from the dual transport ring
10 and storing it in a drop buffer; and
inserting a plurality of data streams into the dual transport ring according to data rates assigned to their destinations.

18. The method of transporting digital traffic among a plurality of nodes
15 located on a dual transport ring according to claim 17, wherein the data streams are a series of containers of data units and are categorized according to their destinations and classes and the step of storing a plurality of data streams comprises a step of:

- 20 storing a plurality containers in their respective add buffers according to their destinations and classes.

19. The method of transporting digital traffic among a plurality of nodes located on a dual transport ring according to claim 18, comprising further step of:
selecting one data stream at a time among eligible data streams for
25 insertion in accordance with a predetermined data rate.

20. The method of transporting digital traffic among a plurality of nodes located on a dual transport ring according to claim 19, wherein the containers are uniform in size.

30

21. The method of transporting digital traffic among a plurality of nodes located on a dual transport ring according to claim 17, comprising a further step of:
building a routing table based upon topological information received by a
35 domain controller node.

22. The method of transporting digital traffic among a plurality of nodes located on a dual transport ring according to claim 21, comprising a further step of:

acquiring the allowed data rate from the domain controller node.

5

23. The method of transporting digital traffic among a plurality of nodes located on a dual transport ring according to claim 21, comprising further steps of
a source node sending a connection packet to a destination node and the
allowed data rate it desires to set up a connection between the source node and the
destination node.

10

24. The method of transporting digital traffic among a plurality of nodes located on a dual transport ring according to claim 17, comprising further steps of:
acquiring a credit for insertion of packets of each data stream;
accumulating the credit for each data stream; and
declaring the data stream eligible for transport when the credit exceeds or
equals the allowed data rate.

15

25. In a telecommunications network formed by a ring transport medium having a plurality of nodes, a method of scheduling digital data at a node, the
digital data consisting of a plurality of data streams, comprising steps of:

20

storing the plurality of data streams respectively in a plurality of storage locations;

calculating a parameter indicative of the rate of the data streams stored in
each storage location;

25

declaring each data stream eligible for transport when the parameter exceeds an allowed data rate of said data stream; and

transporting into the ring transport medium each eligible data stream in accordance with a predetermined data rate.

30

26. The method of scheduling digital data at a node according to claim 25, comprising a further step of:

communicating with a destination node for establishing a connection therewith; and

35

setting an allowed data rate to each data stream in cooperation with the destination node.

27. The method of scheduling digital data at a node according to claim 26,
5 comprising a further step of:

separating digital data into a plurality of data stream in accordance with classes or destination node of the data stream.

28. The method of scheduling digital data at a node according to claim 25,
10 comprising a further step of:

communicating with a domain controller node for traffic allocation; and
setting an allowed data rate to each data stream in cooperation with the domain controller node.

- 15 29. The method of scheduling digital data at a node according to claim 28, comprising further steps of:

monitoring a digital traffic propagating in the ring transport medium, the digital traffic containing digital data being transported from each node and;

- 20 negotiating at a node with the domain controller node for the allowed data rate for each data stream.

30. The method of scheduling digital data at a node according to claim 25, comprising a further step of:

- 25 accumulating a fraction of a total bandwidth which is allocated to each data stream; and

calculating the parameter indicative of the accumulated fraction.

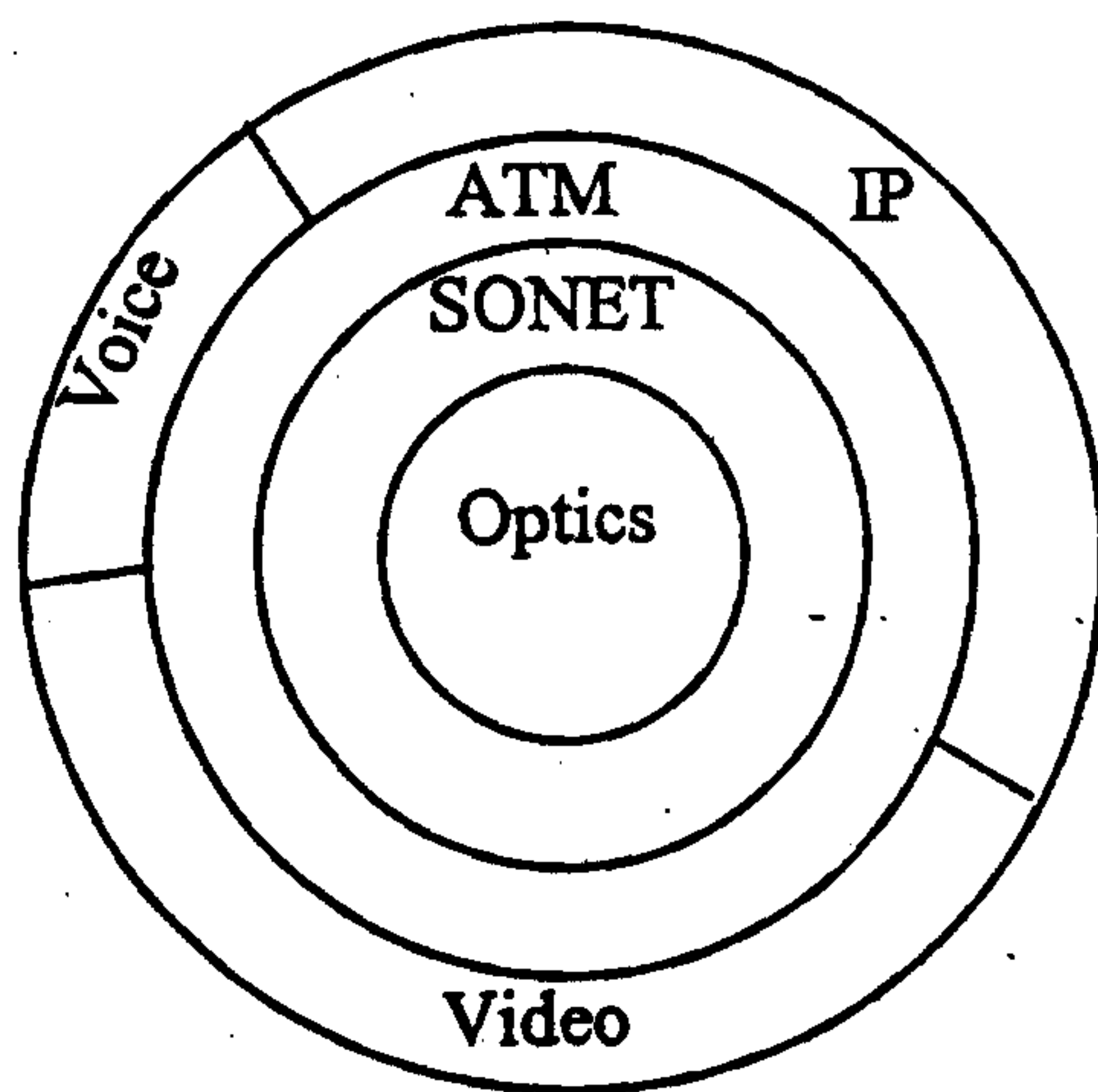
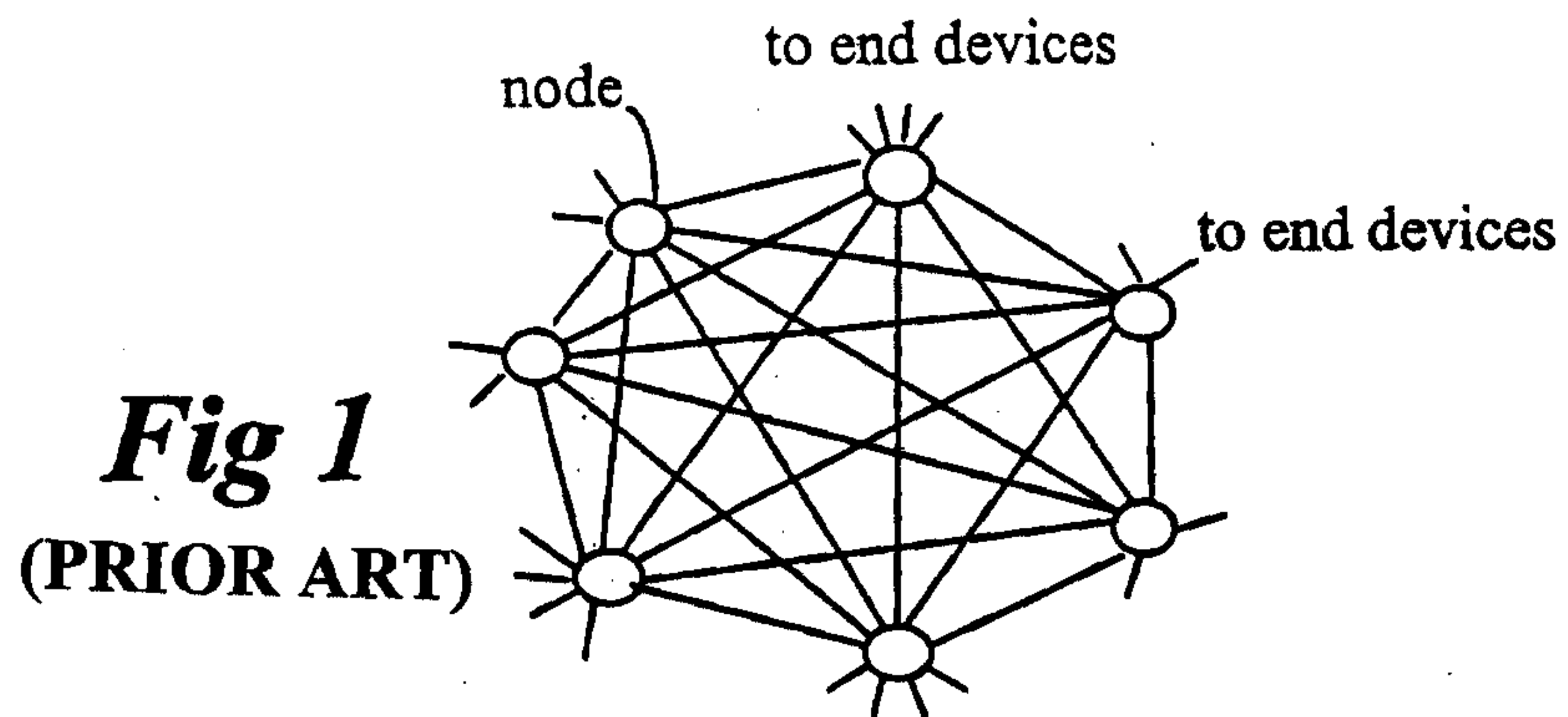


Fig 2 (PRIOR ART)

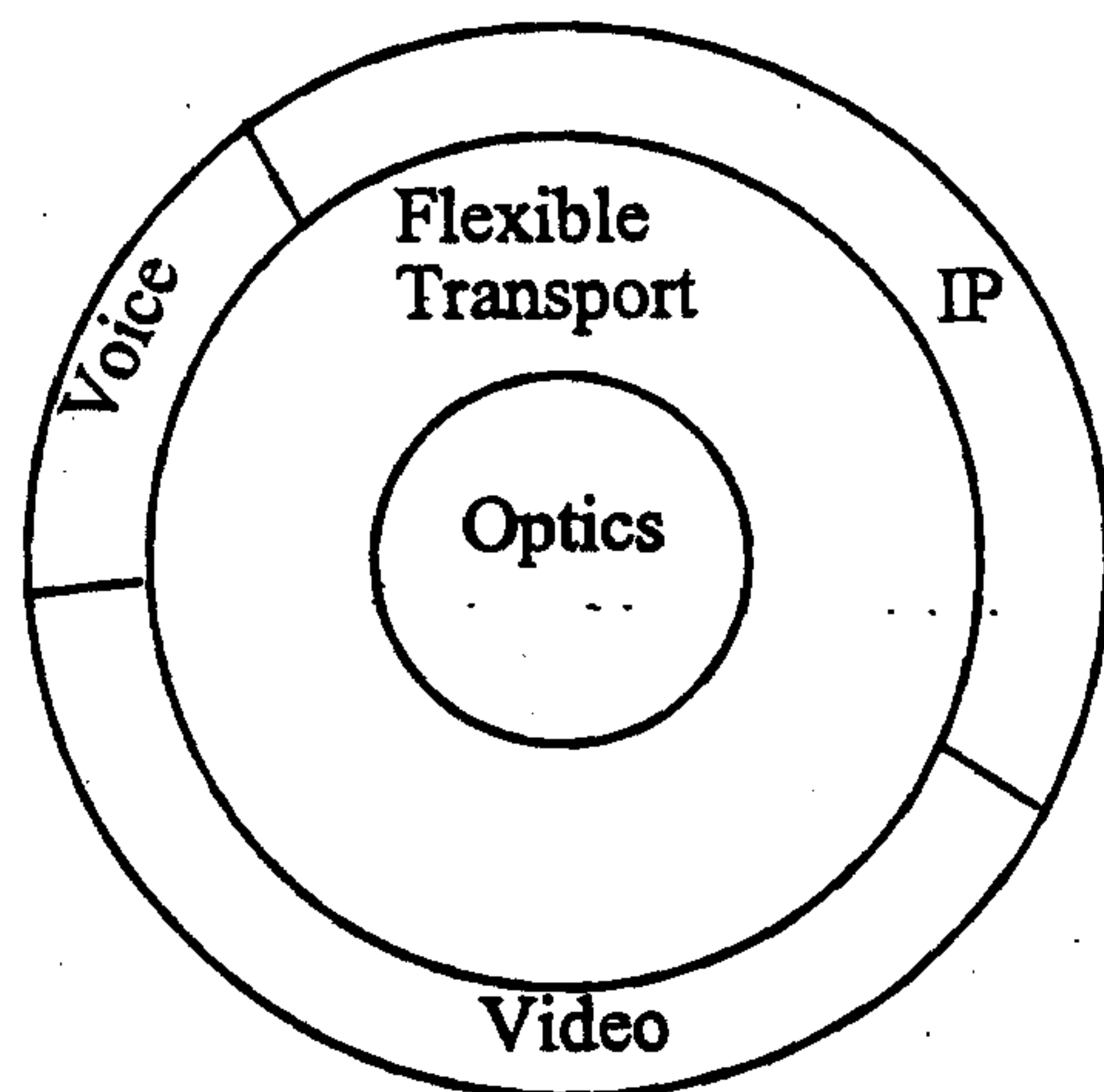
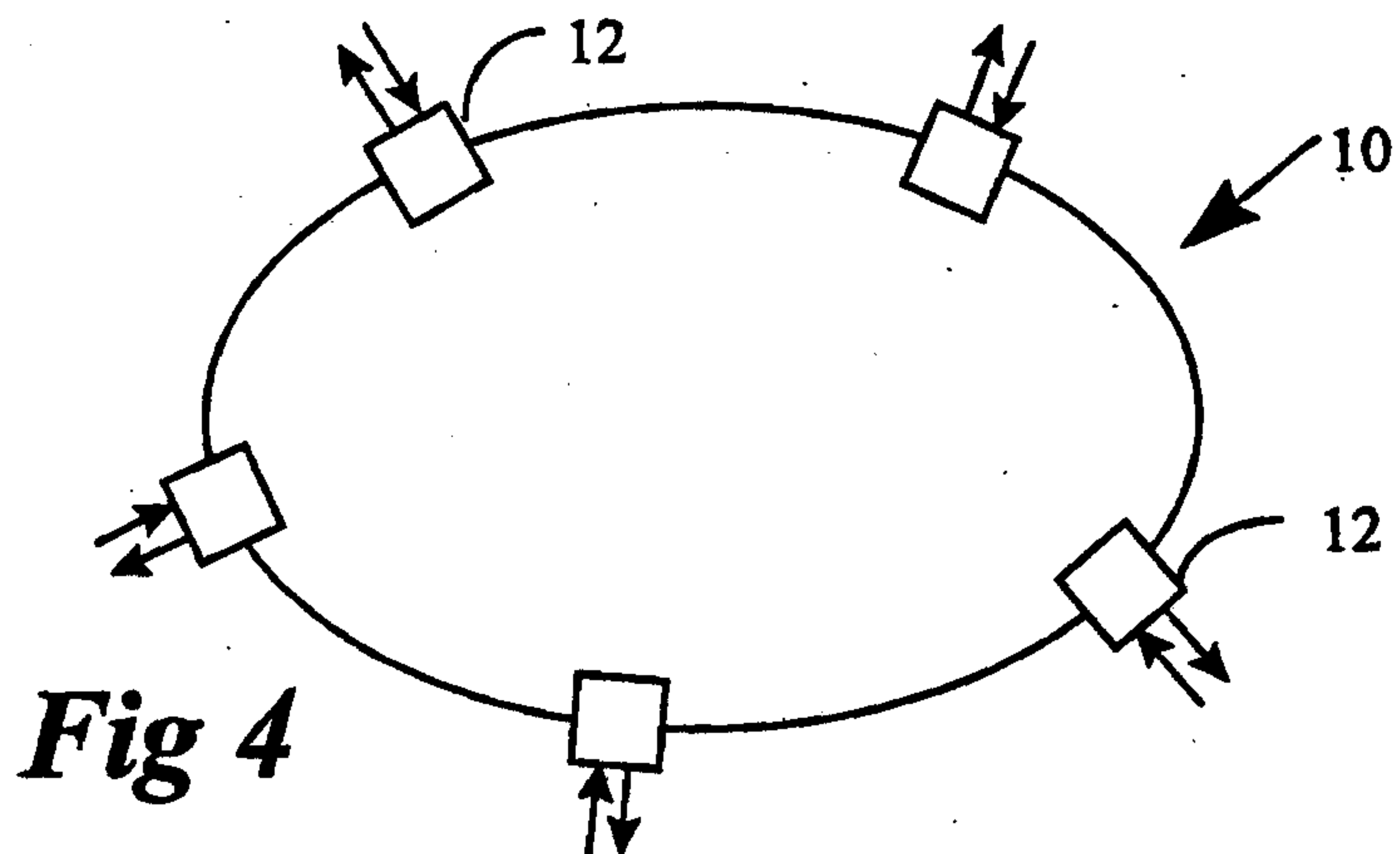
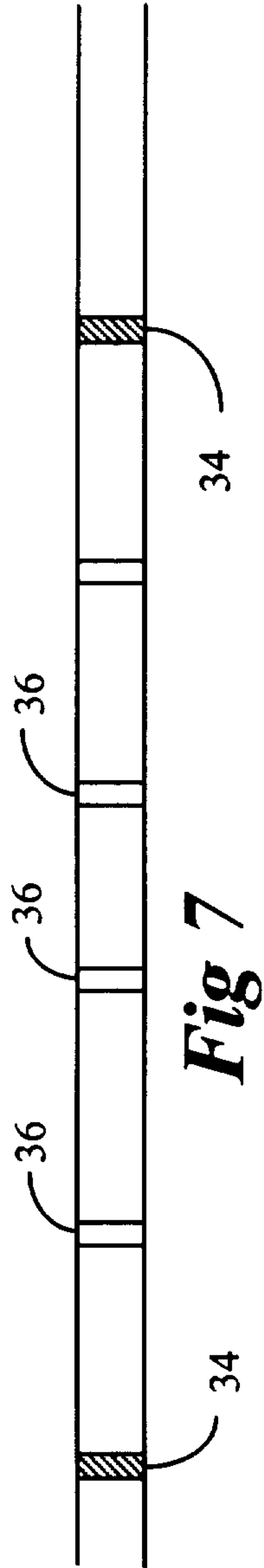
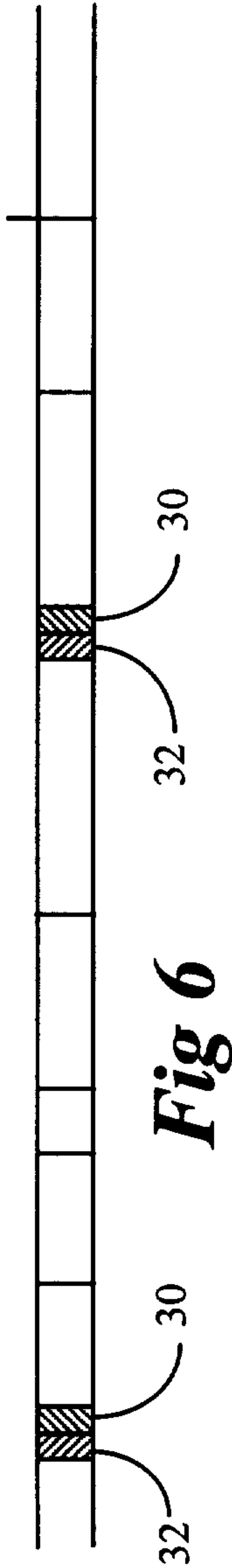
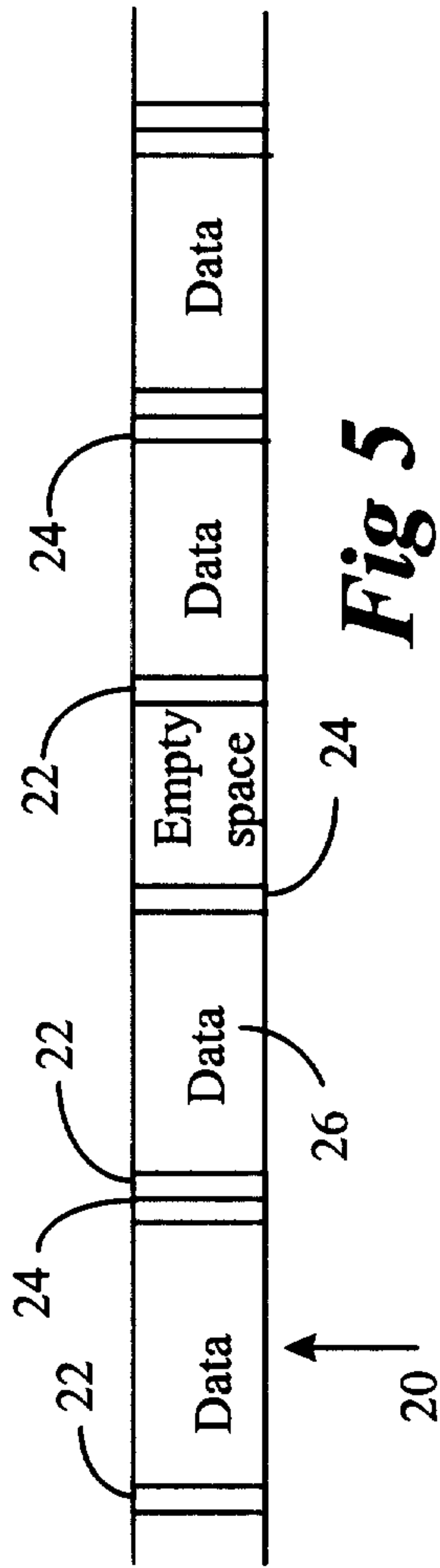


Fig 3





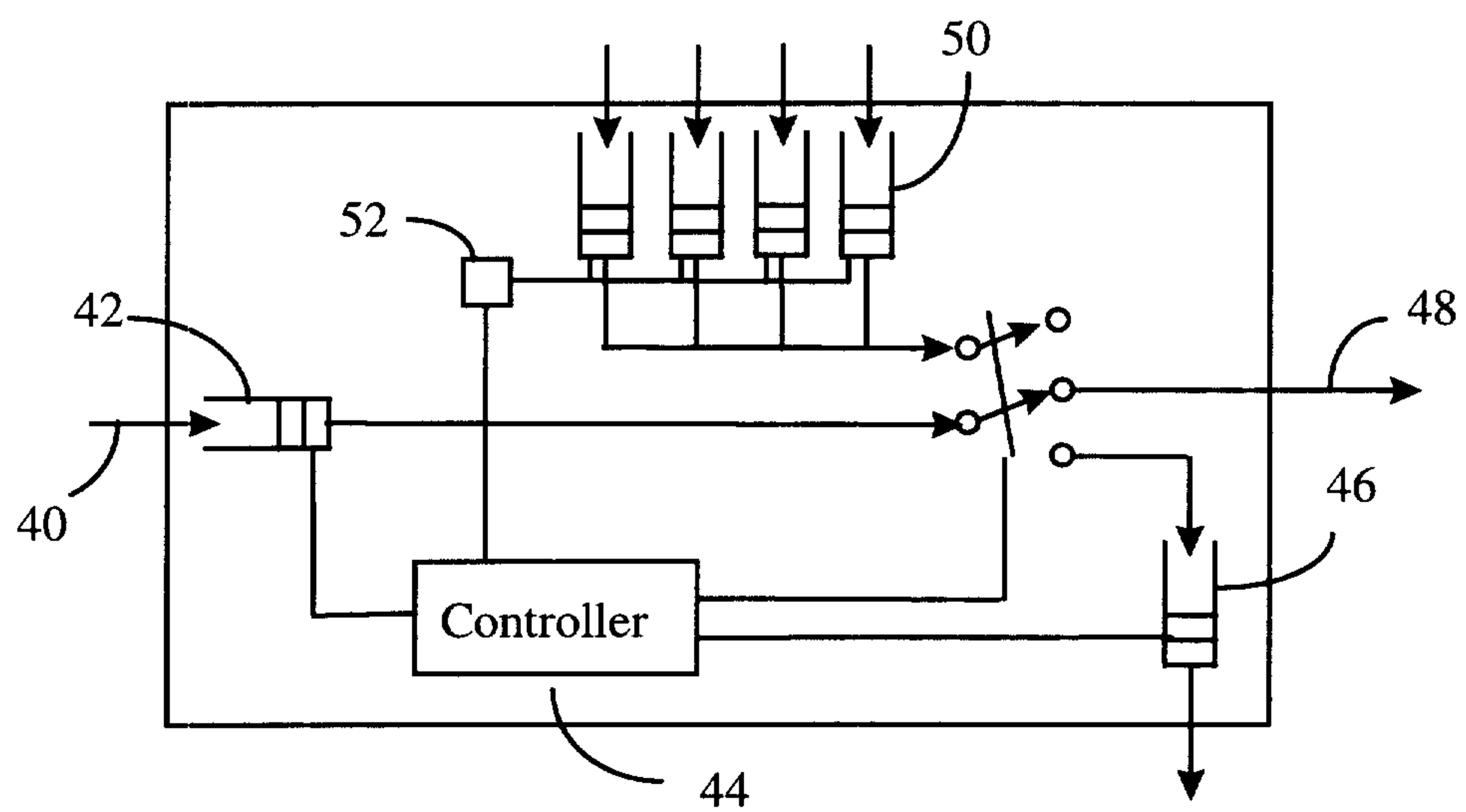


Fig 8

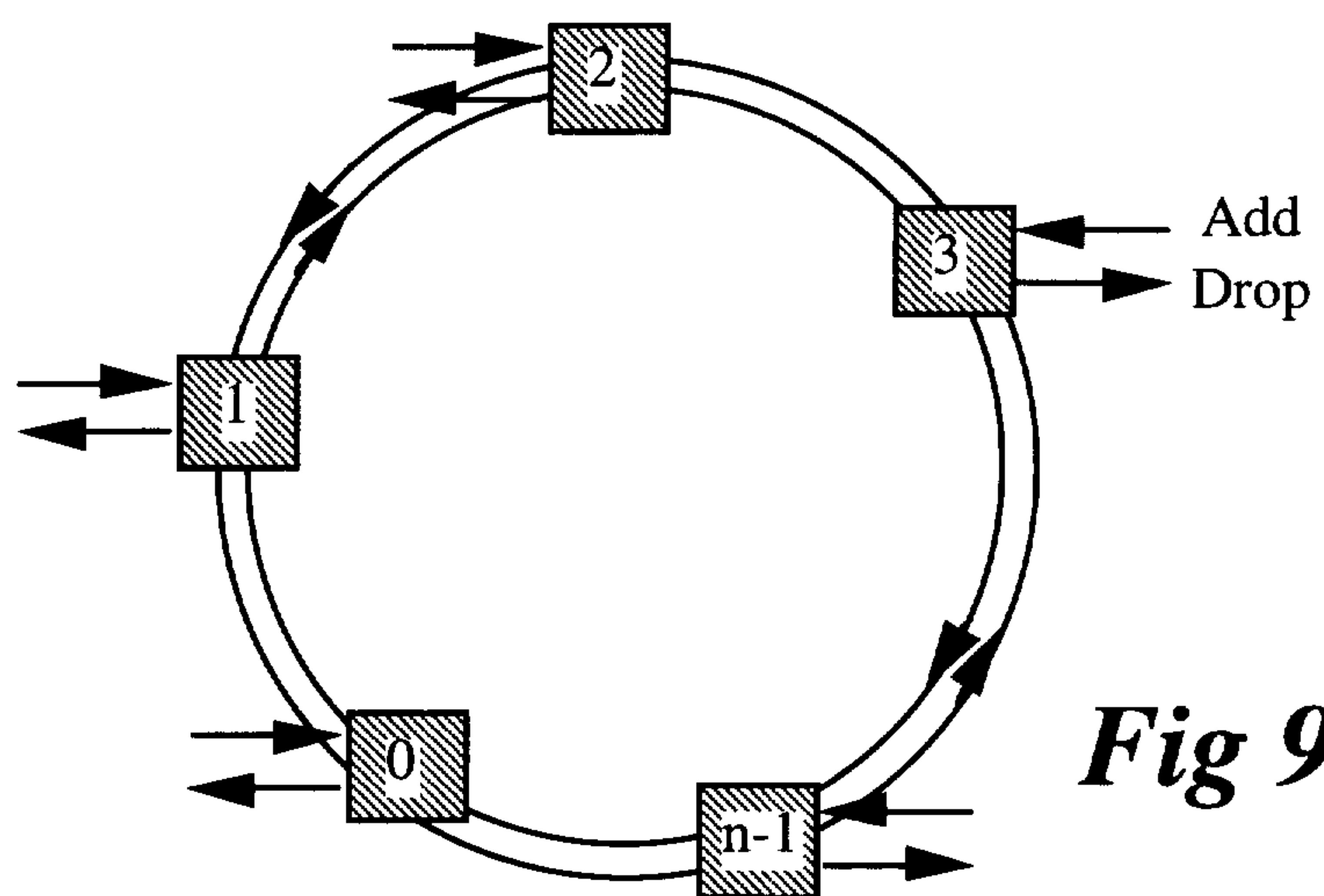


Fig 9

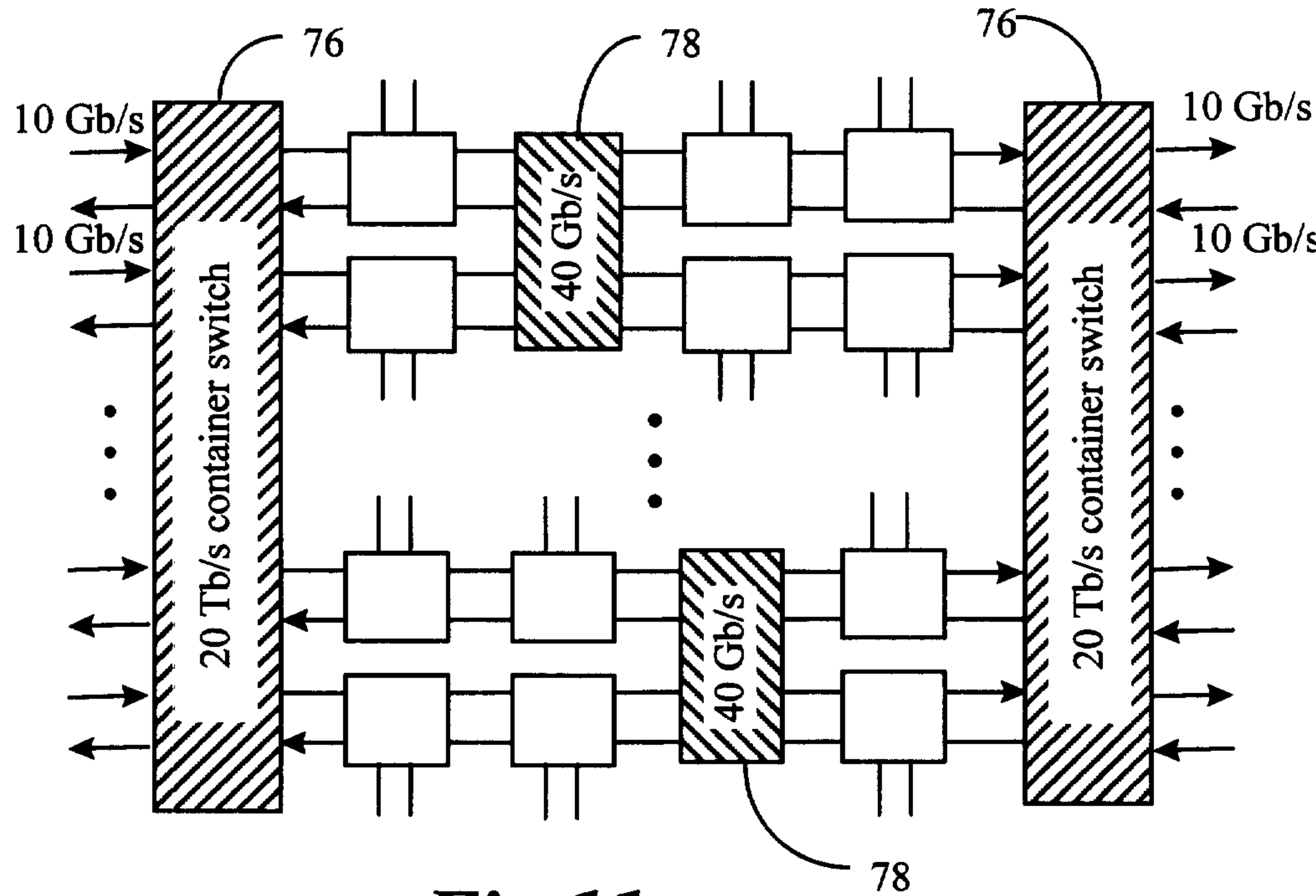
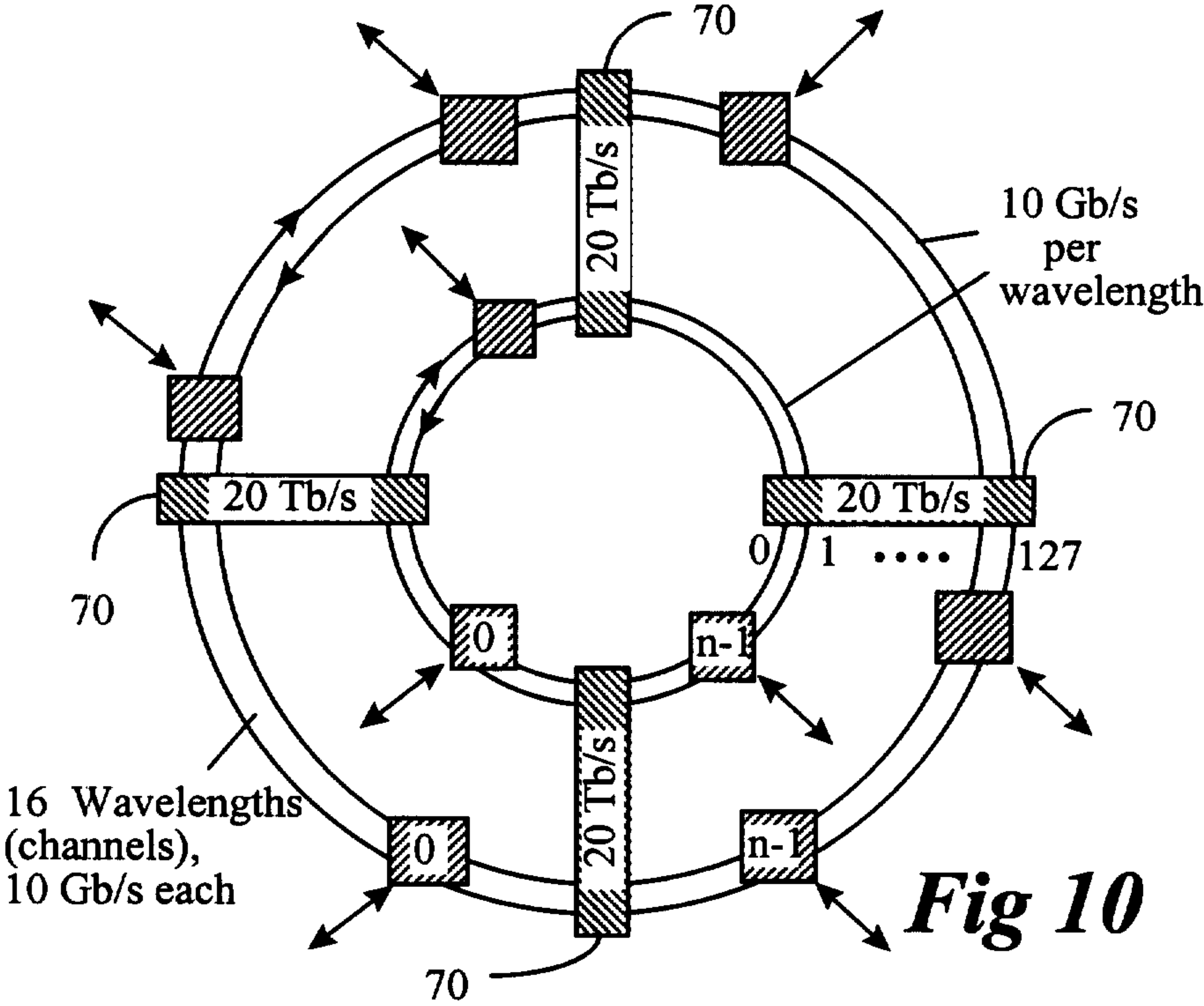


Fig 11

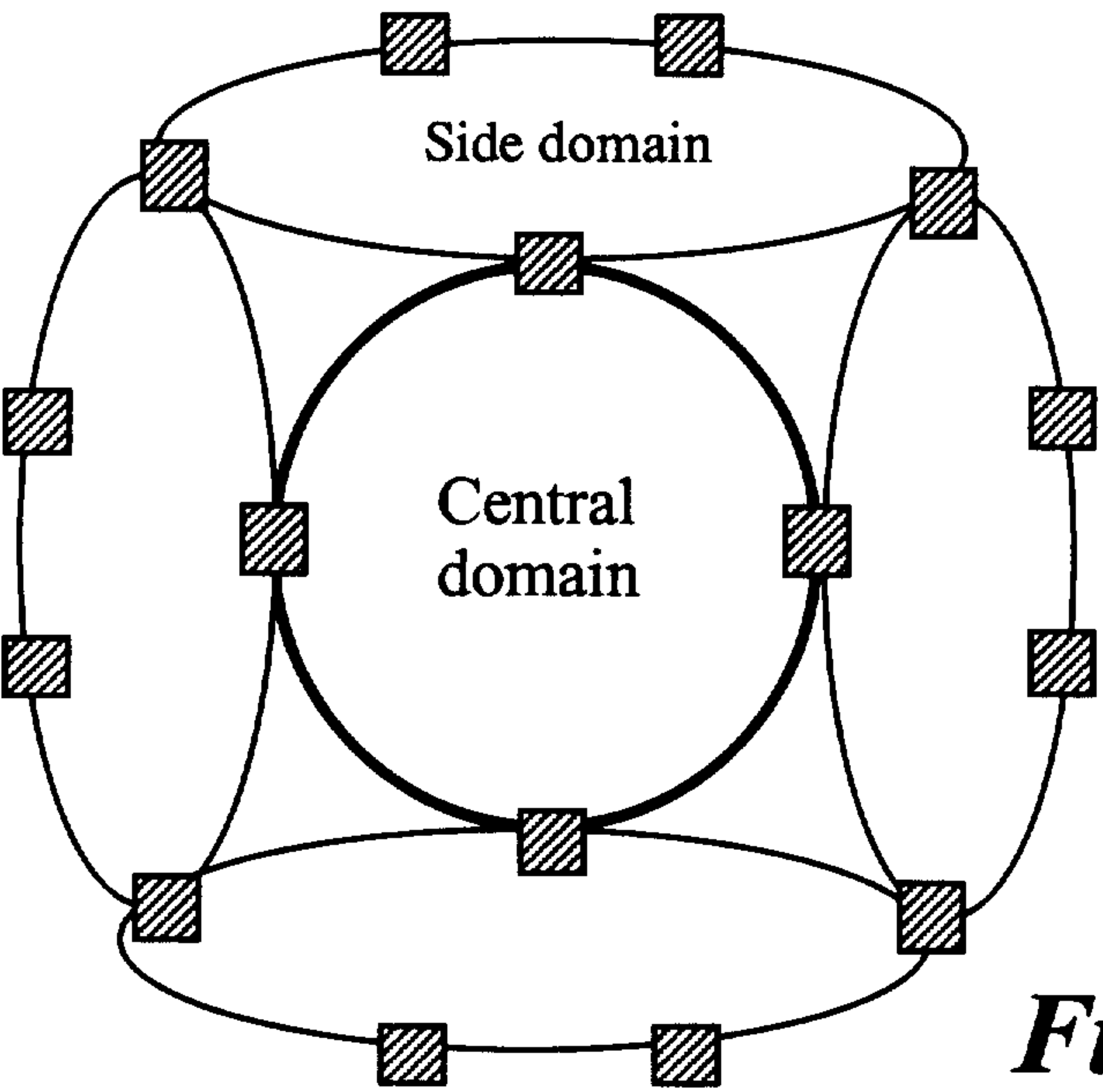


Fig 12

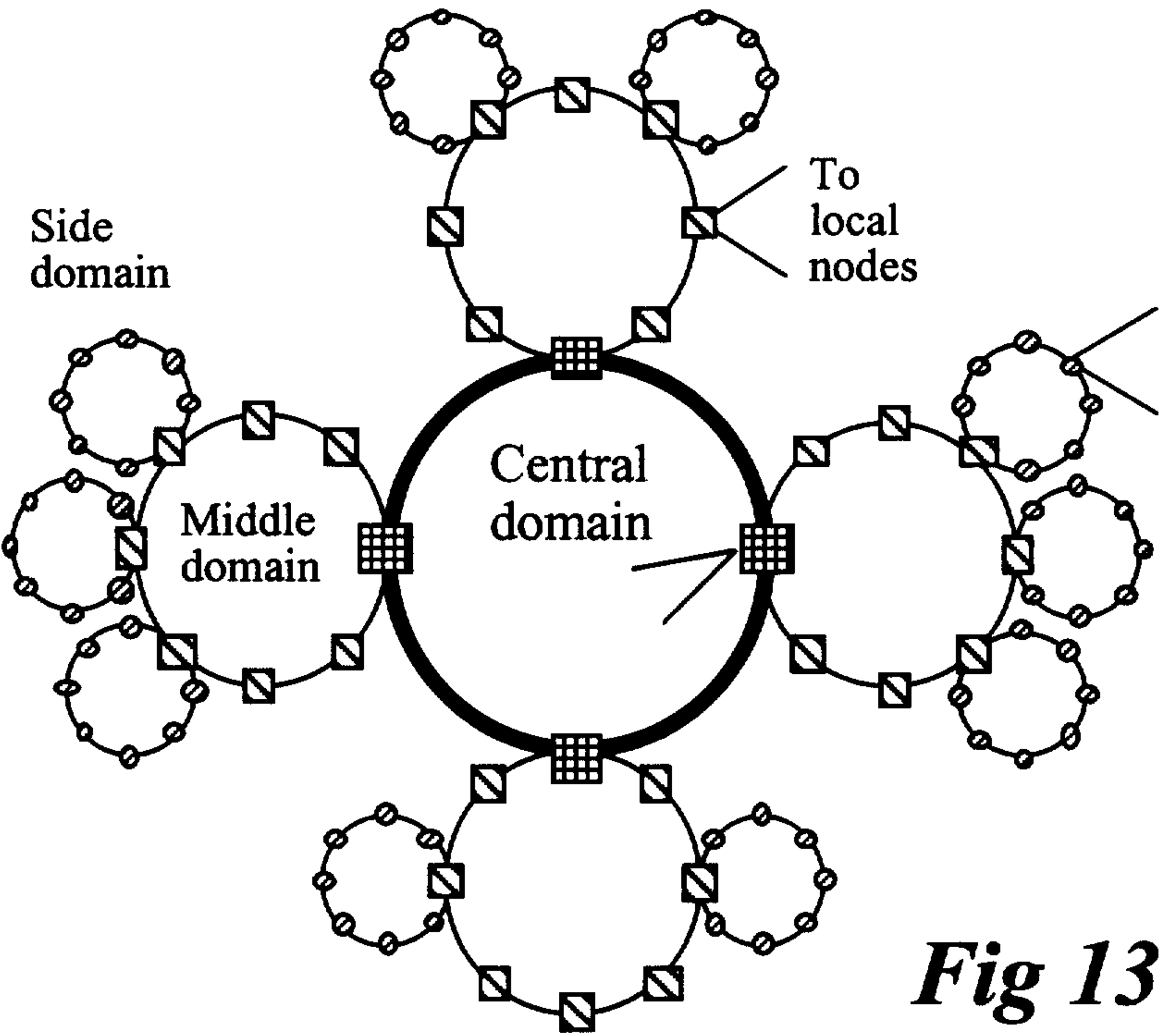
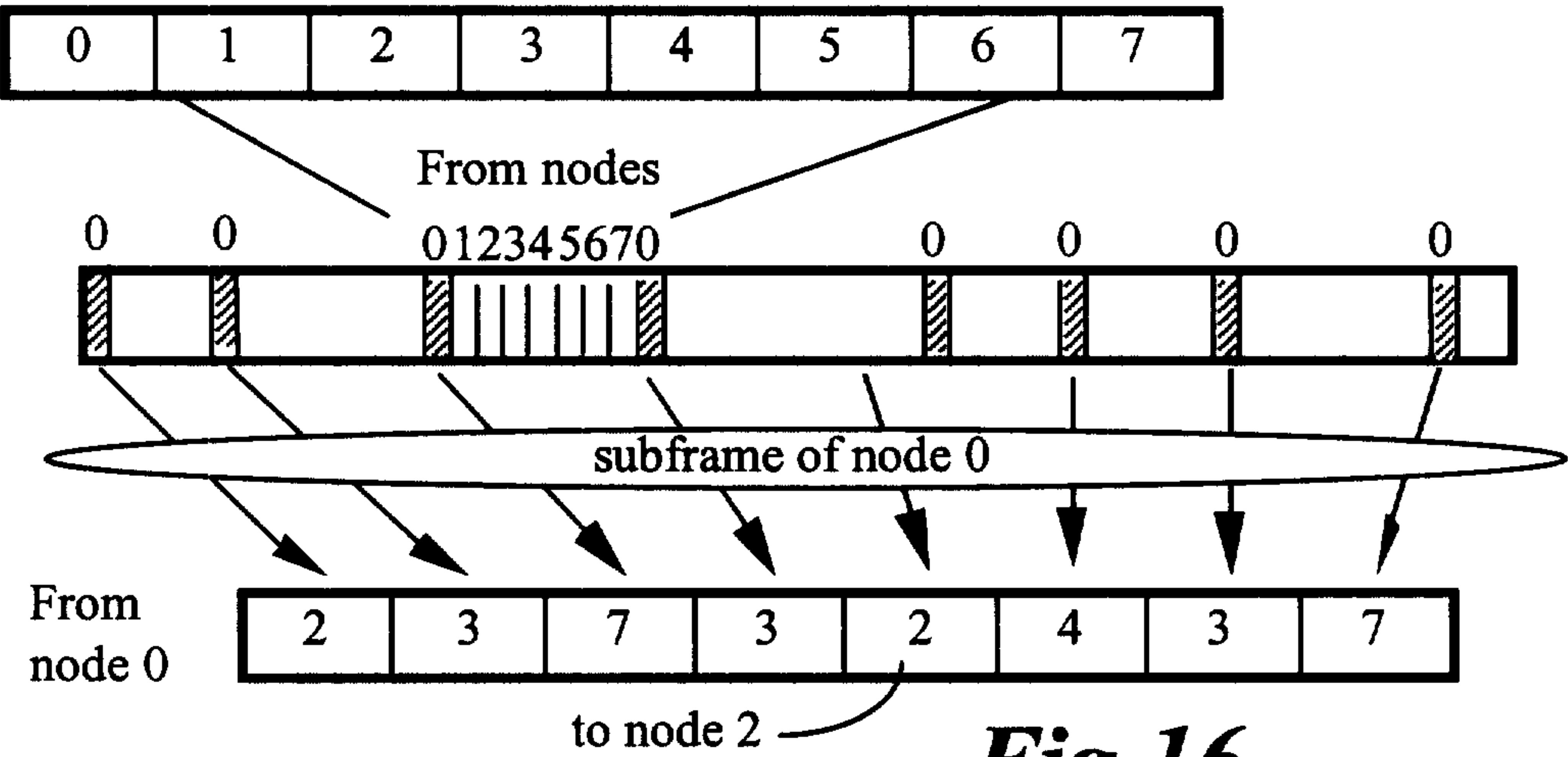
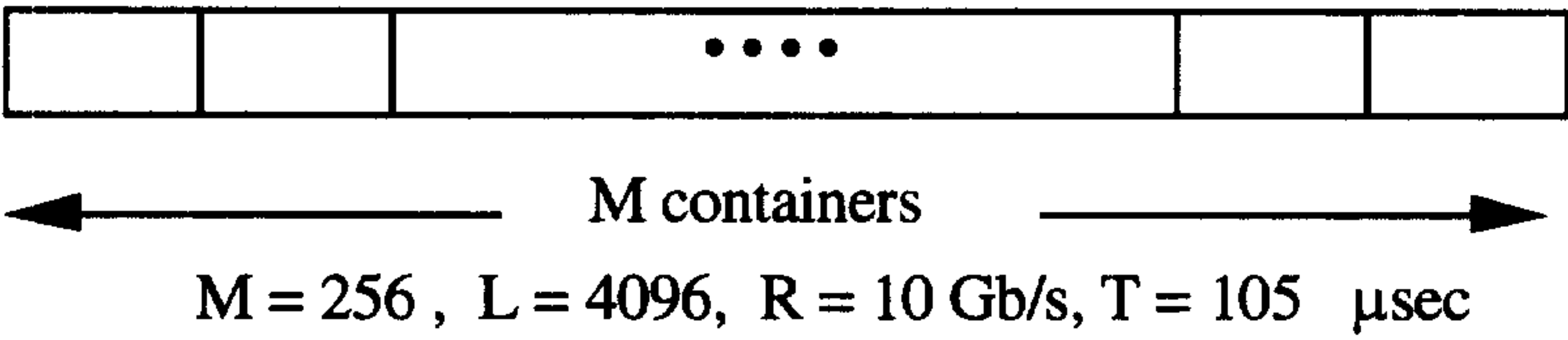
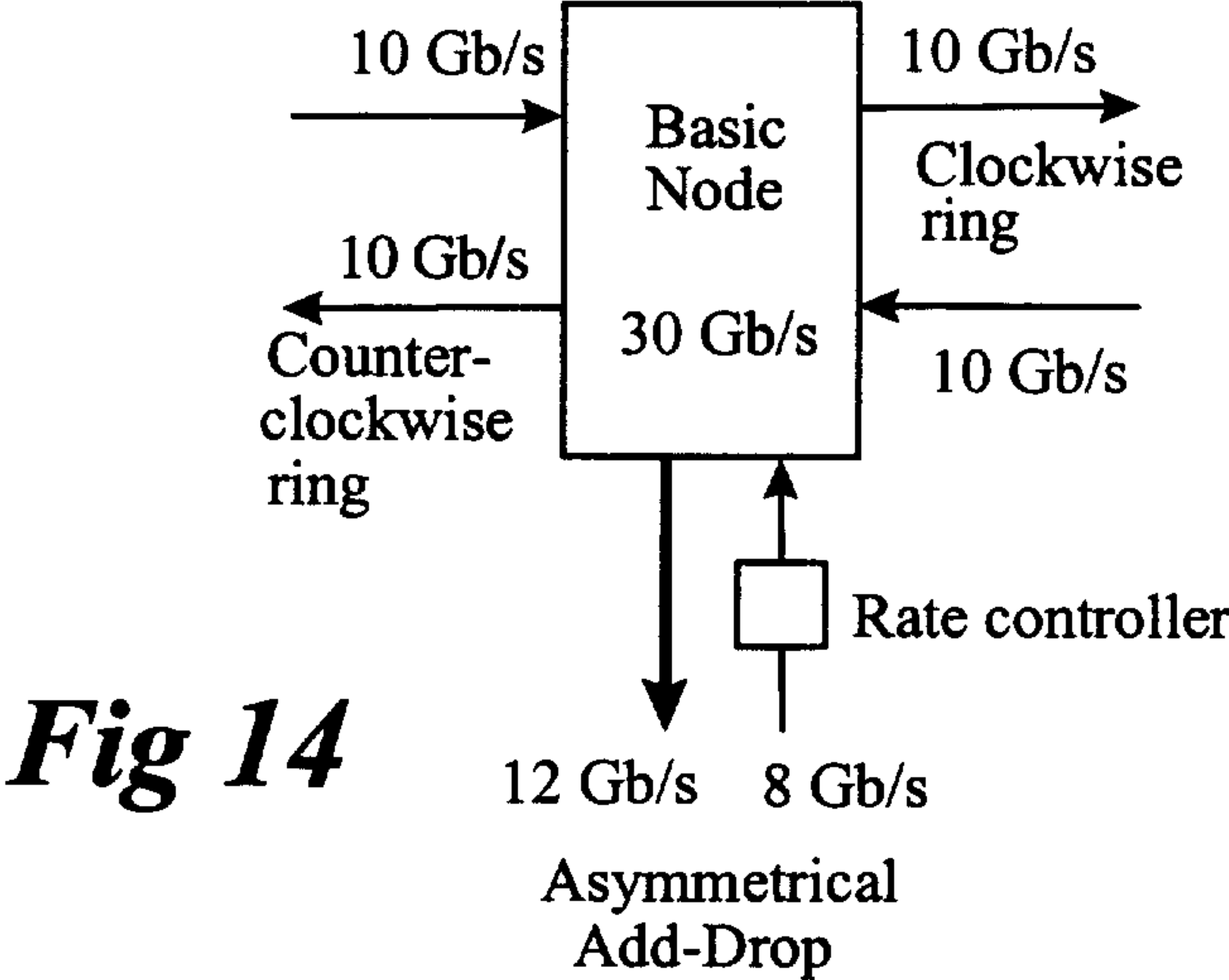


Fig 13



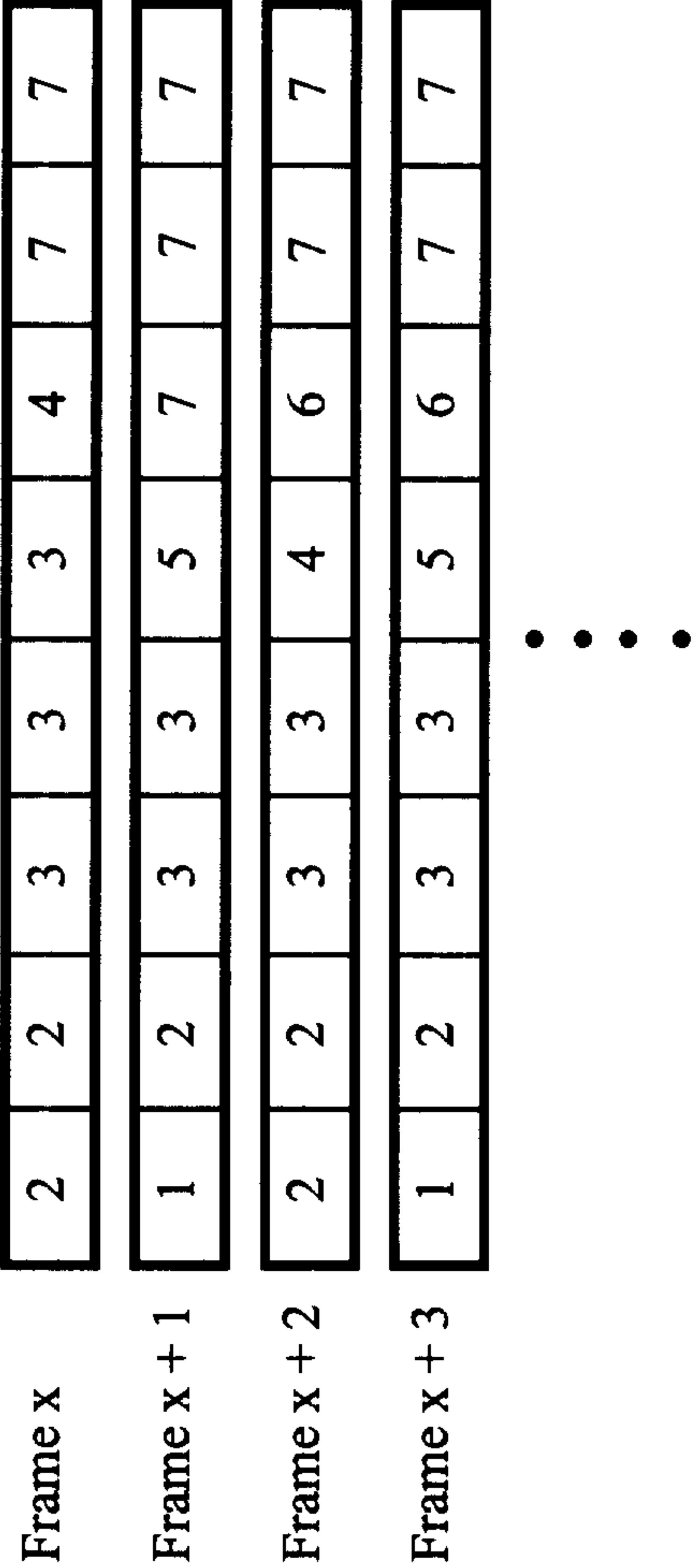


Fig 17

Outgoing-container allocation: node 0

(Successive entries are separated by n slots in the frame.)

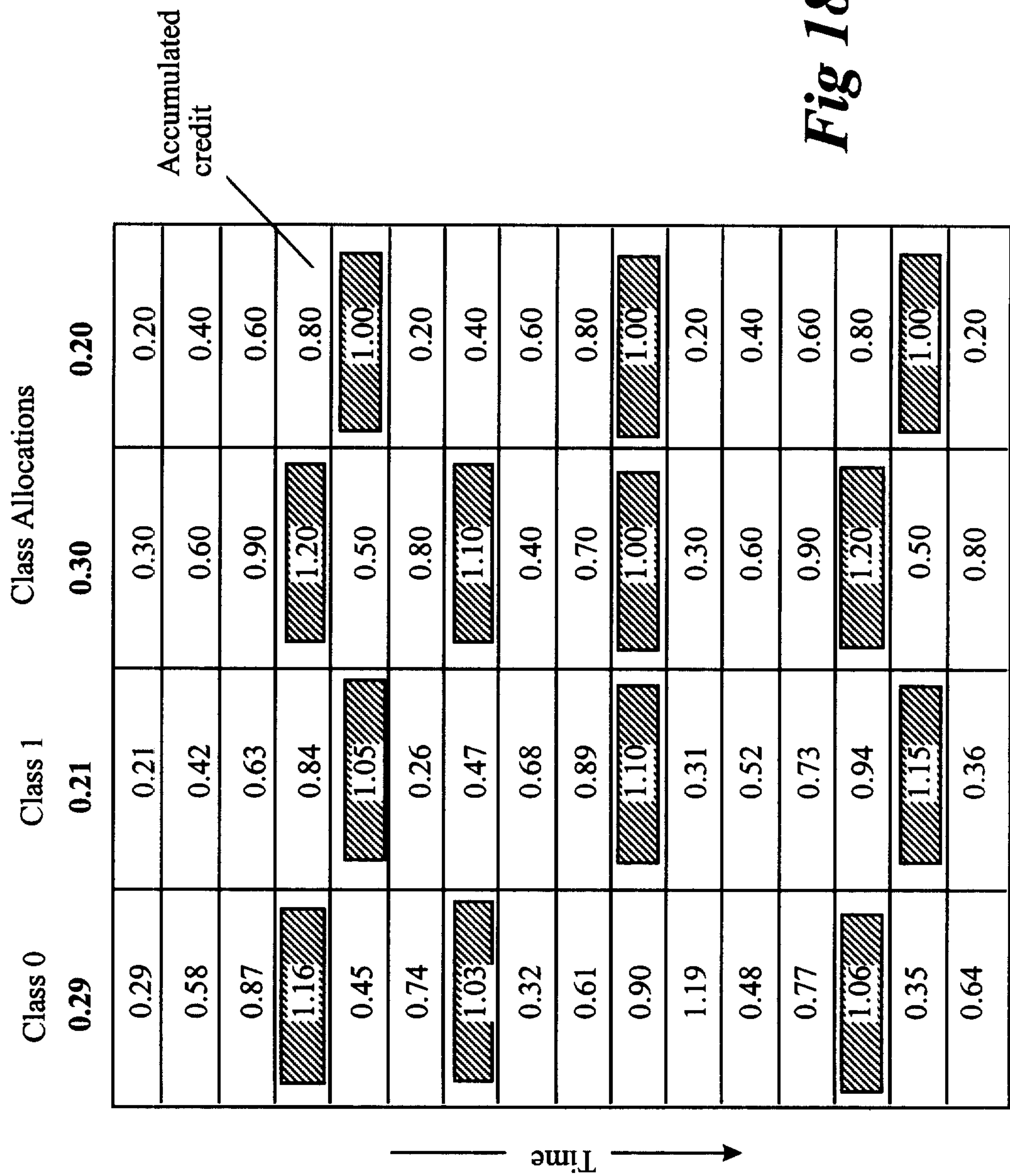


Fig 18

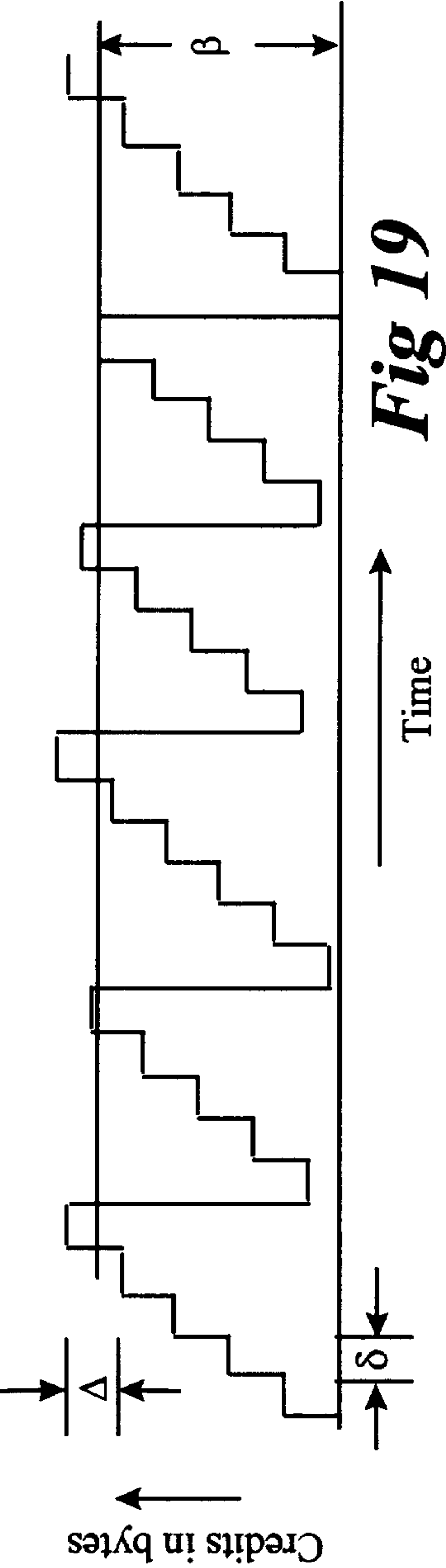
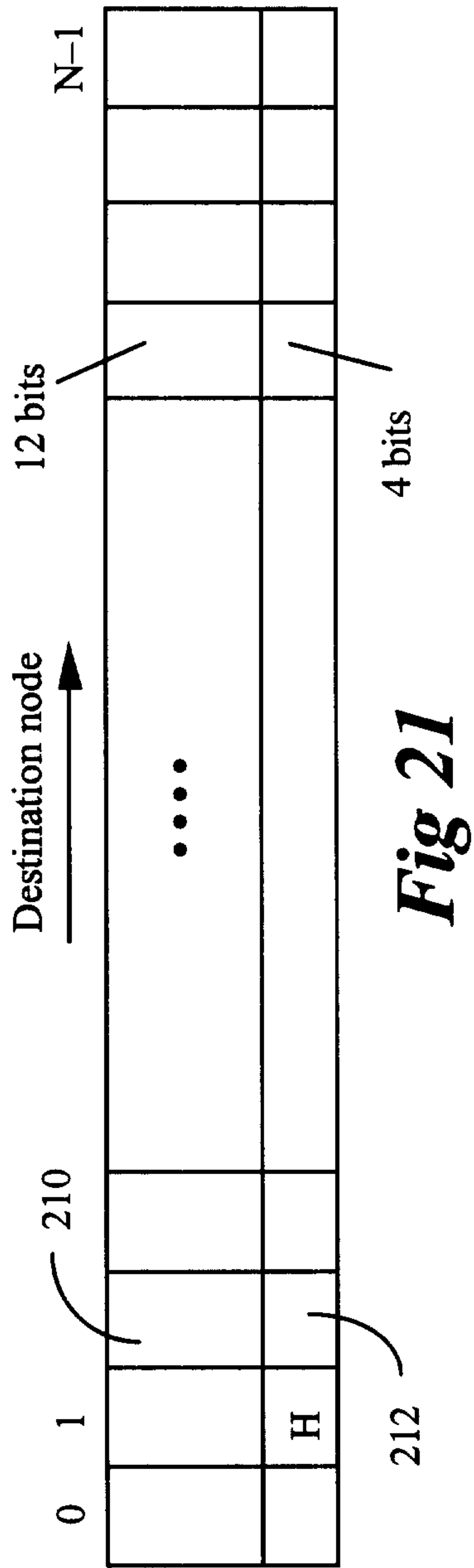
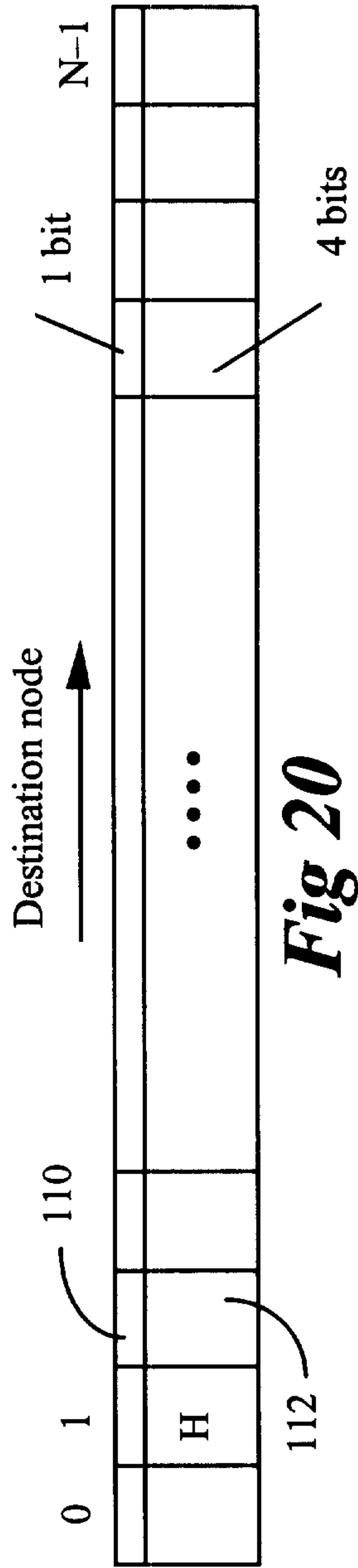


Fig 19



Number of hops	Destination Number	Capacity	Route-trace

Fig 22

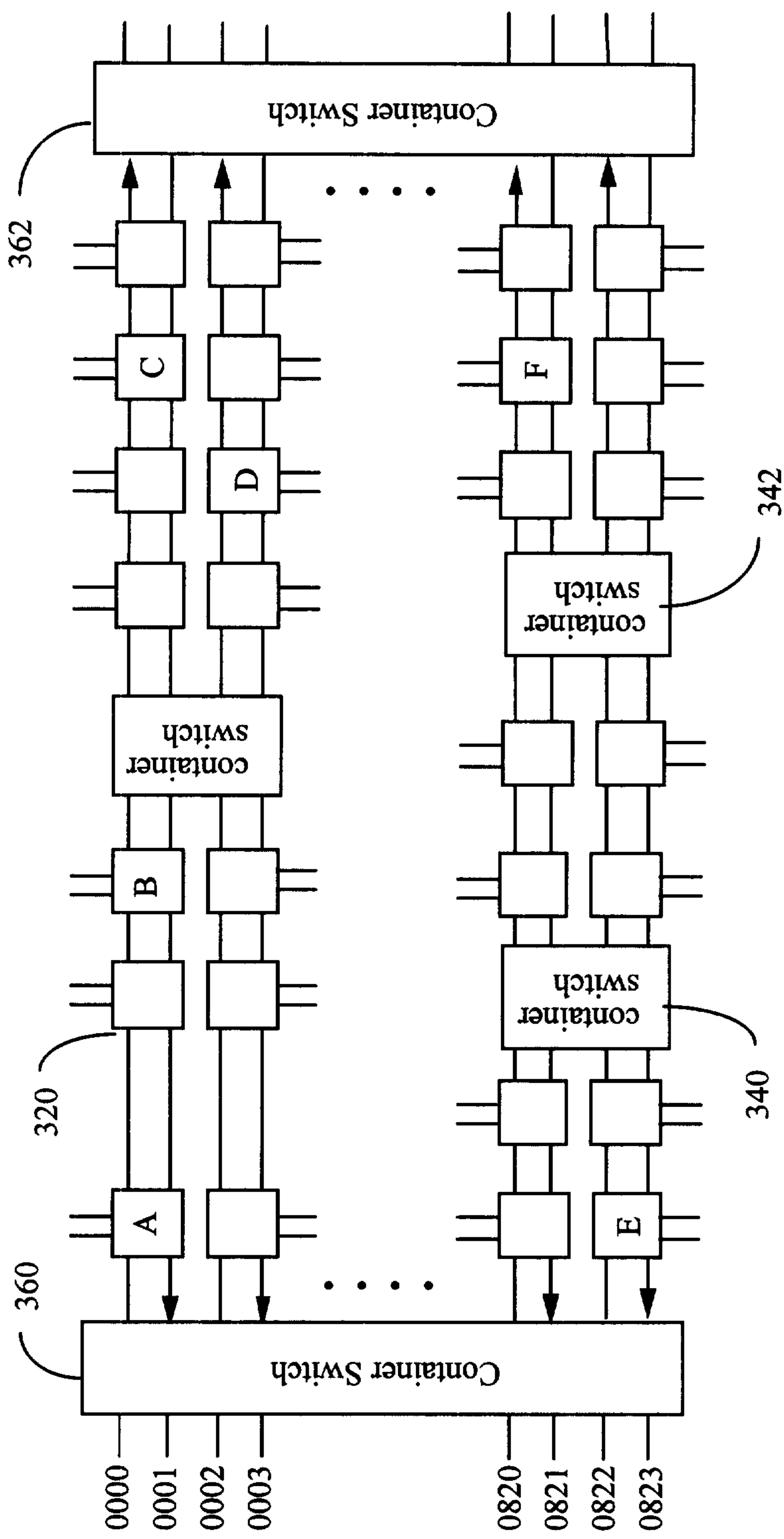


Fig 23

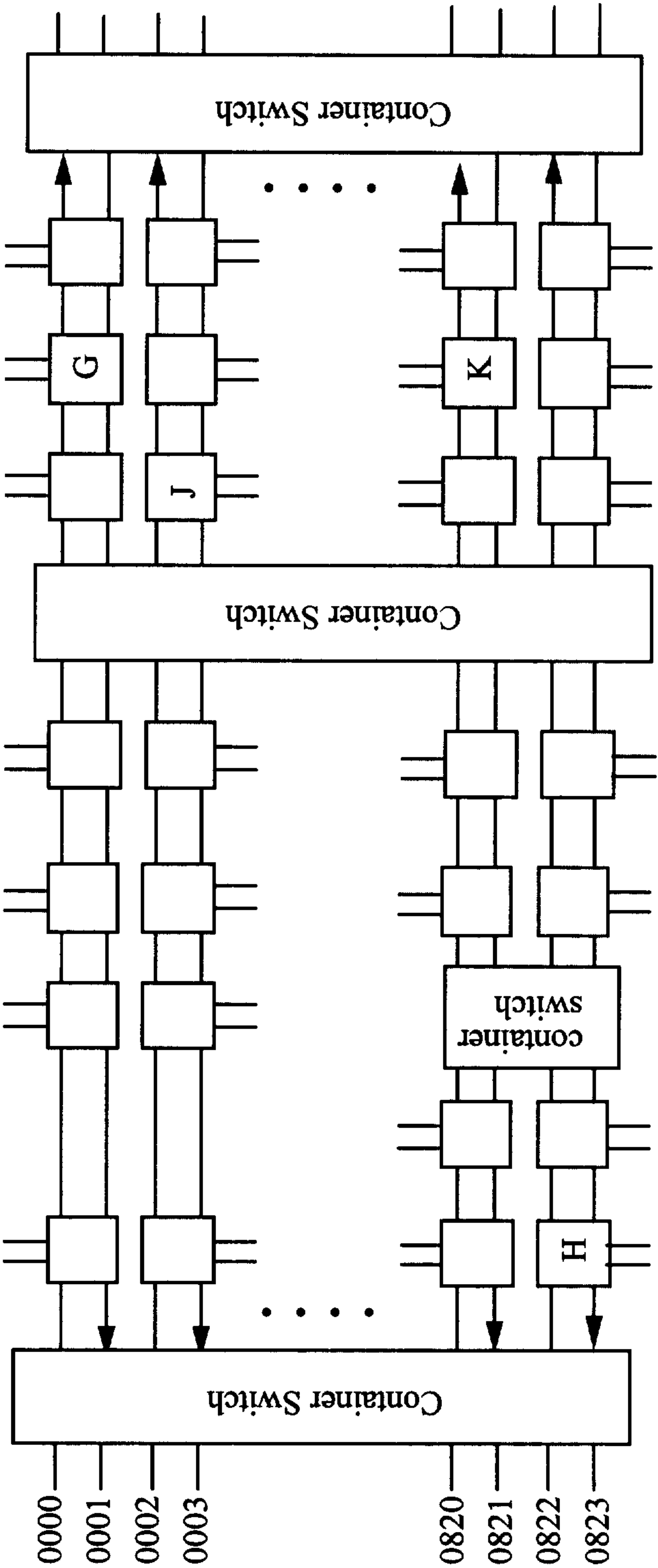


Fig 24

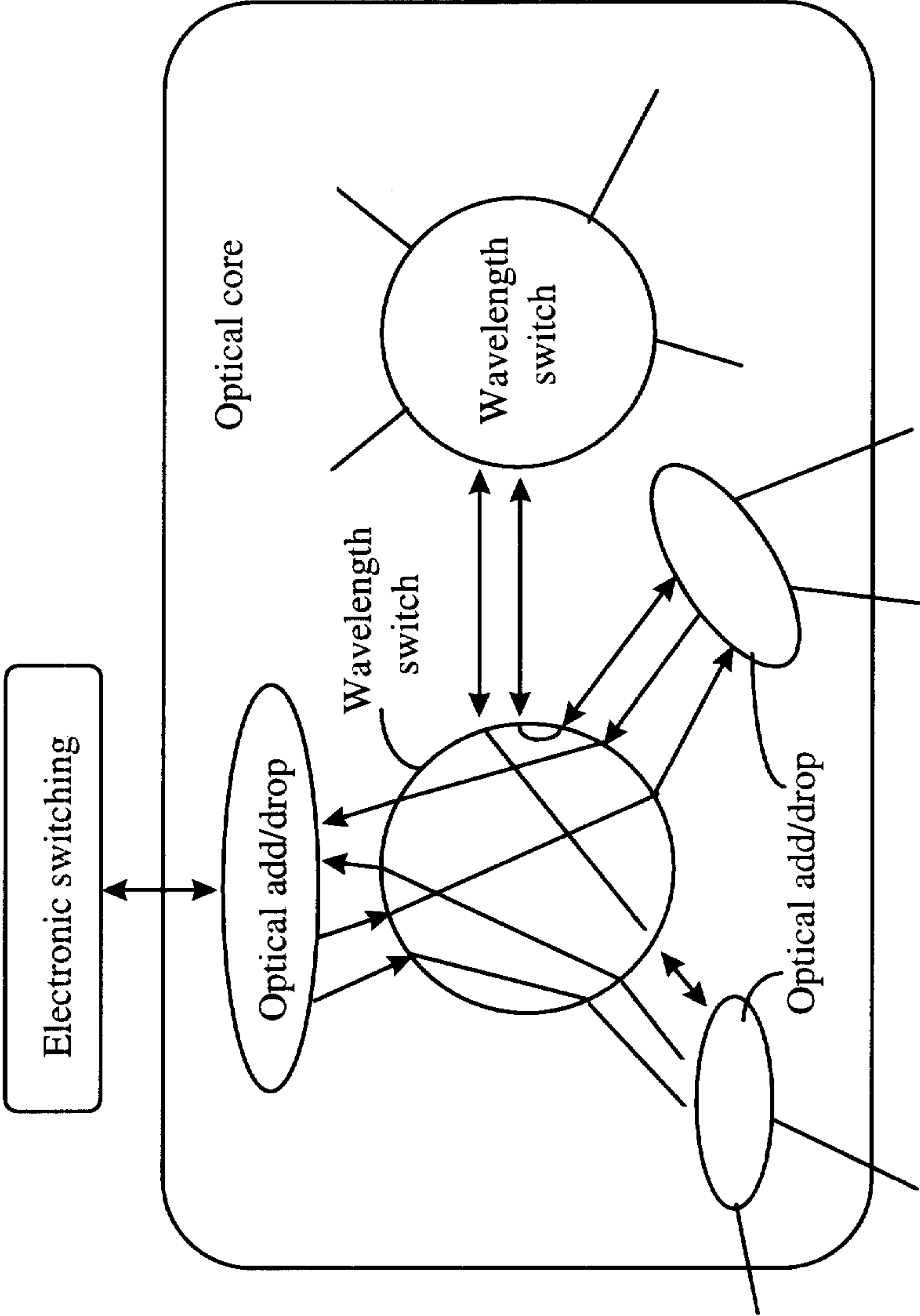


Fig 25

