(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
22 June 2006 (22.06.2006)

PCT

(10) International Publication Number
WO 2006/063459 A1

(51) International Patent Classification:
H04L 12/56 (2006.01)

(21) International Application Number:
PCT/CA2005/001913

(22) International Filing Date:
19 December 2005 (19.12.2005)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/636,485 17 December 2004 (17.12.2004) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/636,485 (CON)
Filed on 17 December 2004 (17.12.2004)

(71) Applicant *(for all designated States except US)*: **ONECHIP PHOTONICS INC.** [CA/CA]; 46 Antares Drive, Suite 200, Ottawa, Ontario K2E 1Z7 (CA).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **HALL, Trevor** [GB/CA]; 5532 Millview Street, Ottawa, Ontario K4M 1J3 (CA). **PAREDES, Sofia** [MX/CA]; 2107-180 Lees Avenue, Ottawa, Ontario K1S 5J6 (CA). **TAEBI, Sareh** [IR/CA]; 908-2870 Cedarwood Drive, Ottawa, Ontario K1V 8Y5 (CA).

(74) Agent: **FREEDMAN, Gordon**; Freedman & Associates, 117 Centrepointe Drive, Suite 350, Nepean, Ontario K2G 5X3 (CA).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RU, SC, SD, SE, SG, SK, SL, SM, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US (patent), UZ, VC, VN, YU, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

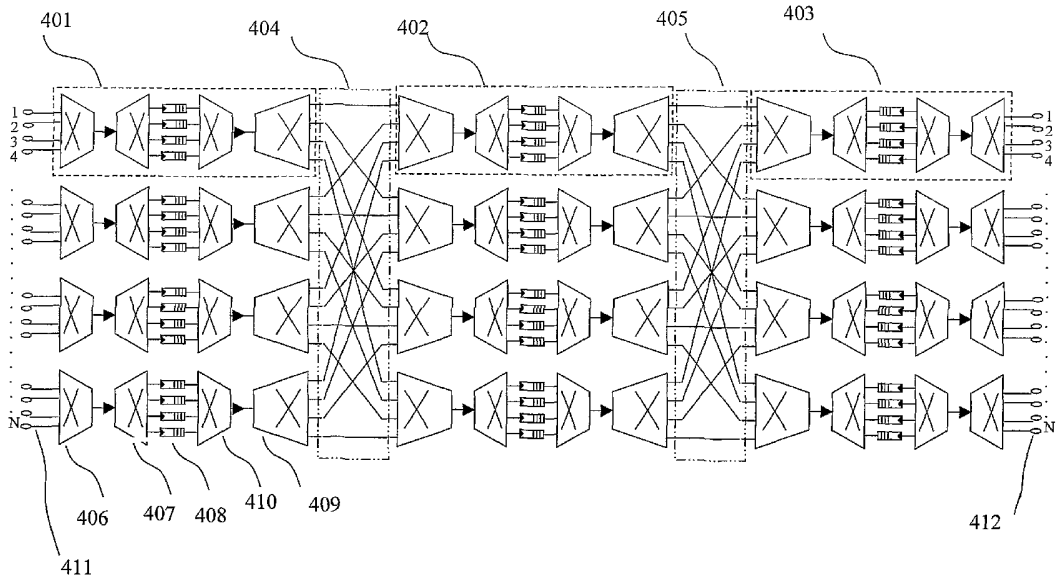
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

— *of inventorship (Rule 4.17(iv))*

[Continued on next page]

(54) **Title:** COMPACT LOAD BALANCED SWITCHING STRUCTURES FOR PACKET BASED COMMUNICATION NETWORKS



(57) **Abstract:** A switching node is disclosed for the routing of packetized data employing a multi-stage packet based routing fabric combined with a plurality of memory switches employing memory queues. The switching node allowing reduced throughput delays, dynamic provisioning of bandwidth and packet prioritization.

WO 2006/063459 A1

Published:
— *with international search report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

Compact Load Balanced Switching Structures for Packet Based Communication Networks

FIELD OF THE INVENTION

[001] The invention relates to the field of communications and more particularly to a scaleable architecture for packet based communication networking.

BACKGROUND OF THE INVENTION

[002] Telecommunications networks have evolved from the earliest networks having few users with plain old telephone service (POTS) to networks in operation today interconnecting hundreds of millions of users with a wide variety of services including for example telephony, Internet, streaming video, and MPEG music. Central to these networks is the requirement for a switching fabric allowing different users to be connected either together or to a service provider. Supporting an increase in a number of users, connections and bandwidth are networks based upon segmentation, transmission, routing, detection and reconstruction of a signal. The segmentation results in a message being divided into segments - referred to as packets, and such networks being packet switched networks.

[003] From a viewpoint of users, this process is transparent provided that the telecommunications network acts in a manner such that the packetization, and all other processes occur in a manner such that the user has available the services and information as required and "on demand." The users perception of this "on demand" service varies substantially depending upon the service used. For example, when downloading most information via the Internet, a small delay is acceptable for text and photographs but not for streamed video unless sufficient memory buffer exists. Amongst the most sensitive services is telephony as the human perception of delay in voice is extremely acute. The result is that network providers prioritize packets according to information content, priority information included as part of the header of a packet.

[004] The switching fabric of current telecommunications packet networks is a massive mesh of large electronic cross-connect switches interconnected generally by

very high speed optical networks exploiting dense wavelength division multiplexing to provide interconnection paths offering tens of gigabit per second transmission. Within this mesh are a limited number of optical switches which generally provide protection switching and relatively slow allocation of bandwidth to accommodate demand.

[005] But the demands from users for increased services, increased bandwidth and flexible services are causing the network operators to seek an alternative architecture. The alternative is "agile" networks which are widely distributed implementations of packet switching, as necessary to provide dynamic routing / bandwidth very close to users and with rapidly shifting patterns as they access different services. Agility to the network operators implies the ability to rapidly deploy bandwidth on demand at fine granularity. Helping them in this is the evolution of access networks which have to date been electrical at rates up to a few megabits per second but are now being replaced with optical approaches (often referred to as fiber-to-the-home or FTTH) with data rates of tens to hundreds of megabits per second to customers, and roadmaps to even gigabit rates per subscriber.

[006] As the network evolves, and services become more flexible and expansive, speeds increase such that the network provider is increasingly focused to three problems:

- Delay - the time taken to route packets across the network, where excessive delay in any single packet of a message prevents the message being completed

- Mis-Sequencing - the mis-sequencing of packets through the network causes delays at the user as until the mis-sequenced packet arrives the message cannot be completed

- Losses - the loss of packets due to blocked connections within the network causes delays as the lost packets must be retransmitted across the network.

[007] It is therefore desirable within the network to address these issues with a physical switching fabric. The invention disclosed provides such an architecture for the distributed packet switching wherein the fabric acts to balance the traffic load on

different paths and network elements within the distributed packet switch. In doing so the disclosed invention removes additionally the requirement for rapid reconfiguration of the packet switches, which has the added benefit of allowing the deployment of optical switches within the network which are slower and smaller than their electrical counterparts.

SUMMARY OF THE INVENTION

[008] In accordance with the invention there is provided a switching node in respect of routing data packets arriving at the switching node within a communications network. The switching node contains a plurality of input ports each of which receives data packets addressed to it from the broader communications network. Within the switching node are multiple memory switches which are implemented by a combination of a plurality of memory queues, for storing the packet data therein, coupled to a first switch matrix for switching of packet data for storage within a memory queue of the plurality of first memory queues, and a second switch matrix for switching of packet data retrieved from within a memory queue of the plurality of first memory queues.

[009] The multiple memory switches are then coupled to a third switching matrix, which is coupled on one side to the plurality of input ports and the plurality of memory switches on the other. The multiple memory switches are then coupled to a fourth switching matrix coupled such that on the one side are the plurality of memory switches and on the other the plurality of output ports.

[0010] At least one of the third or fourth switching matrix is implemented with a second set of multiple memory queues which are coupled between a fifth switch matrix and sixth switch matrix. In this invention the packets of data arriving at the switching node are sequenced within the memory queues and memory switches with the packets of data then being routed appropriately between the input and outputs using the multiple switching matrices.

[0011] As a result the switching node can meet all of the demands of the network provider in terms of quality of service, flexibility of provisioning to a users varied demands for services, and prioritizing packet data switching based upon predetermined

priorities of the packets and the dynamic bandwidth allocation between input and output ports. The control approach allows this to be achieved in an architecture where the loading of activities such as switching, memory queuing etc is balanced across the node.

In another embodiment of the invention the use of multiple memory queues and memory switches allows the switching node to store packet data having a lower priority in an earlier stage of the multi-stage memory queue. Additionally the matrices coupled to the memory queues may be spatial switches, time division multiplexing switches, or a combination thereof.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] Exemplary embodiments of the invention will now be described in conjunction with the following drawings, in which:

[0013] FIG. 1 illustrates a prior art approach to packet switching using a centralized shared memory switch with queues.

[0014] FIG. 2 illustrates a prior art packet switch using a three-stage Close-like network.

[0015] FIG. 3A illustrates a first embodiment of the invention wherein the load-balanced switch is implemented with an input queued crossbar switch with route and select switches.

[0016] FIG. 3B illustrates a second embodiment of the invention wherein the load-balanced switch is implemented with an output queued crossbar switch and has the routing controller segmented with a control segment per switching stage.

[0017] FIG. 4A illustrates a third embodiment of the invention wherein the load-balanced switch is implemented in a manner mimicking a three stage Clos fabric where the external links operate at the same speed as the internal links.

[0018] FIG. 4B illustrates a fourth embodiment of the invention wherein the load-balanced switch is implemented in a manner mimicking a three stage Clos fabric but wherein the switch matrices and shuffle networks are reduced functionality.

[0019] FIG. 5 illustrates a fifth embodiment of the invention wherein the load balanced switch is implemented in a manner mimicking a three stage Clos fabric where the external links operate at the twice the speed of the internal links.

DETAILED DESCRIPTION OF EMBODIMENTS OF THE INVENTION

[0020] Referring to FIG. 1a, shown is a prior art approach to a packet switch using a single stage of memory queues. A plurality of input ports 101 are connected to physical links within a communications network (not shown). These input ports 101 are coupled to an input multiplexer 102, which multiplexes the plurality of input packet data streams to a single data stream. The single data stream is then transported to a 1:N distribution switch 103, which is coupled to N parallel memory queues 104, each memory queue 104 allowing packets of data to be stored until retrieved.

[0021] The N parallel memory queues 104 are in turn connected to an N:1 concentrator switch 105 that reads from the memory queues 104. The output data stream of the concentrator switch 105 is then connected to a demultiplexing switch 106 which in turn connects to a plurality of output ports 107.

[0022] A packet of data arriving at input port 101a of the switching fabric, being one of the plurality of input ports 101 is multiplexed by the multiplexing switch 102 to the common communications path prior to propagating within the distribution switch 103. The packet of data from input port 101a then propagates to one of the memory queues 104. The packet is then stored prior to being retrieved by the concentrator switch 105 and then being routed by the demultiplexer switch 106 to the appropriate output port 107b, being one of the plurality of output ports 107.

[0023] Now referring to FIG. 2 shown is a prior art implementation of a packet switch based upon a three-stage Clos architecture. Here a packet of data arrives at one of the N input ports 201 of one of the plurality of first stage routing switches 202. Assuming that there are R such first stage routing matrices 202, each having M output

ports, the data received is time-stamped, its header read and an identifier of the target output port communicated to the packet switch controller 210. This determines the routing through the switching node specifically, and causes the packet of data to be routed to the appropriate output port of the first stage routing matrix 201 for transport to the subsequent section of the packet switching node. When transported, the packet of data propagates through a first perfect shuffle network 203 comprising RxM paths, wherein it addresses one of the M second stage switching matrices 204, which are NxN crosspoint switches.

[0024] The packet switch controller 210 routes the packet of data within the second stage switching matrix 204 for transport to the third stage switch matrix 206. From the appropriate output port of the second stage switch matrix 204, it is routed via a second perfect shuffle network 205 to the specified third stage switching matrix 206. Within the third stage switching matrix 206, the packet is routed directly to an output port 207 of the switching node and transported via the wider communications network.

[0025] Referring to FIG. 3A, an exemplary first embodiment of the invention is shown in the form of a compact load balanced crossbar packet switch with queued input ports. Here a packet of data is incident at one of the input ports 301 of the packet switching node. The header of the packet is read and communicated to the packet switching node controller 315 which defines the appropriate routing of the packet through the node. The packet switching controller 315 communicates routing data to the first stage switch matrix 303 comprising a first NxN crossbar switch with memory queues. This is implemented using 1:N distribution switches 302, a perfect shuffle 313, a plurality of memory queues 316 and N:1 concentrator switches 304. The packet of data exits the first stage switching matrix 303 on a link connecting a second stage switch matrix 305 determined by the packet switching node controller 315.

[0026] The second stage switch matrix 305 is constructed from 1:M distribution switches 306, M memory queues 307, and M:1 concentrator switches 308. The packet of data is routed by the distribution switch 306 to one of the memory queues 307 wherein it is stored pending extraction under the control of the packet switching node controller 315. When required for transport to the third switching stage 309 of the

switching node, the data is extracted from one of the plurality of memory queues 307 and fed forward using the concentrator switch 308.

[0027] Upon arrival at the third switch stage 309, the packet of data is routed to an output port using a second NxN crossbar switch implemented again using 1:N distribution switches 310, a perfect shuffle 314 and N:1 concentrator switches 311, whereupon it is available at output port 312 for transport to the wider communications network.

[0028] Referring to FIG. 3B, the exemplary first embodiment is again shown in the form of a compact load balanced crossbar packet switch but now with queued output ports. Hence, when the packet of data is routed through the first switch matrix 3030 it passes through the 1:N distribution switches 3020, a perfect shuffle 3130, and N:1 concentrator switches 3040. It is when routed via the third switch matrix 3090 that the packet of data passes through the 1:N distribution switches 3100, the second perfect shuffle 3140, the memory queues 3160 and N:1 concentrator switches 3110.

[0029] Alternatively, the first stage switching matrix 3030 and the third stage switching matrix 3090 are implemented with different matrix design architectures which optionally include memory queues in one or the other.

[0030] Additionally the packet switching controller 3150 is shown as three control sections 3150A, 3150B and 3150C each of which interfaces to a switch stage of the switching node as well as communicating with each other to provide overall control of the node. Alternatively, two controller sections are optionally combined if the switching matrices are located making such combination beneficial.

[0031] Referring to FIG. 4A, a simplified architectural diagram of a second embodiment of the invention is shown in the form of a compact load balanced three stage Clos network wherein the Clos stages operate at a same line data rate as an input port and an output port. Here a packet of data is incident at one of N input ports 411 of a packet switching node. A header of the packet of data is read and communicated to a packet switching node controller (not shown) which defines a routing of the packet through the node. The packet switching node controller communicates the routing data to a first stage switch matrix 401 comprising a first concentrator switch 406, a first

memory switch element comprising a first distribution switch 407, a plurality of first memory queues 408 and a first concentrator switch 409.

[0032] From the output port of the first concentrator switch 409, the packet of data is routed to a second distribution switch 410 which feeds the packet of data forward to a first perfect shuffle network 404. In use, the first switching stage 401 performs a grooming of packets to sequence them and route them to a second stage switch matrix 402.

[0033] Within the second stage switch matrix 402, the packet of data is again shuffled with other arriving packets and stored within memory queues awaiting transport to a third switch stage. The second stage switch matrix 402 feeds the packet of data forward to a second perfect shuffle network 405.

[0034] After being routed through the perfect shuffle 405, the packet of data arrives at the third switch stage and enters a third stage switch 403. Here the packet of data is again sequenced with other arriving packets to create output data streams stored within memory queues awaiting transport to the communications network. The third stage switch 403 feeds the packet of data forward to an output port 412 of the switching node.

[0035] Referring to FIG. 4B, an alternate embodiment of the compact load balanced three stage Clos network, wherein the Clos stages operate at a same line data rate as an input port and an output port, but exploits switching elements with reduced complexity. As the packet switch algorithm for the packet switch node controller can be implemented such that it grooms packets of data and routes them such that are grouped according to output port it also possible to adjust the algorithm such that it handles reduced complexity within the first and second shuffle networks.

[0036] In FIG. 4B the reduced complexity of the first shuffle network between the first switch stage 4010 and second switch stage 4020 is implemented with 1:(N-1) distribution switches 4100, shuffle network 4040 and (N-1):1 concentrator switches 4130. Similarly the second shuffle network between the second switch stage 4020 and third switch stage 4030 is implemented with 1:(N-1) distribution switches 4140, shuffle network 4050 and (N-1):1 concentrator switches 4150.

Additionally the memory queues 4080 are shown as constructed from three segments in series, 4080A, 4080B and 4080C. Optionally the memory segments may be assigned to store data packets with predetermined associations, these including, but not being limited to, packets destined for adjacent output ports and assigned to a dedicated output stage memory switch, packet data for packets stored within different memory queues which is assigned to a dedicated intermediate memory sector serving those queues, packet data associated with packets with adjacent input ports and assigned to a dedicated input stage memory sector, data fields arranged so as to provide a transposed interconnection between the input and intermediate stages, and data fields arranged so as to provide a transposed interconnection between the intermediate and output stages.

[0037] Alternatively to perform similar functionality, the switching matrices 401, 402 and 403 of FIG. 4A and 4010, 4020, and 4030 of FIG. 4B is implemented with different matrix architectures and/or design, optionally including memory queues.

[0038] Now referring to FIG. 5, a simplified architectural diagram of a third embodiment of the invention is shown in the form of a compact load balanced three stage Clos network wherein the Clos stages operate at half the data rate of an input port and an output port. A packet of data arrives at one of N input ports 512 of a packet switching node. A header of the packet of data is read and communicated to a packet switching node controller (not shown), which determines routing for the packet through the node. The packet switching node controller communicates routing data to a first stage switch matrix 501, which comprises a first concentrator switch 506, a first memory switch element comprising a first distribution switch 507, a plurality of memory queues 508 and a first concentrator switch 509.

[0039] From the output port of the first concentrator switch 509, the packet of data is routed to a second distribution switch 510 which feeds the packet of data forward to a first perfect shuffle network 504. In use, the first switching stage 501 performs a grooming of packets to sequence them and route them to a second stage switch matrix 502.

[0040] Within the second stage switch matrix 502 the packet of data is again shuffled with other arriving packets and stored within memory queues awaiting

transport to a third switch stage. The second stage switch matrix 502 feeds the packet of data forward to a second perfect shuffle network 505.

[0041] After being routed through the perfect shuffle 505, the packet of data arrives at the third switch stage and enters a third stage switch 503. Here the packet of data is again sequenced with other arriving packets to create output data streams stored within memory queues awaiting transport to the communications network. The third stage switch 503 feeds the packet of data forward to an output port 511 of the switching node.

[0042] Alternatively to perform similar functionality, the switching matrices 501, 502 and 503 are implemented with different matrix architectures and/or design, optionally including include memory queues.

[0043] Advantageously, in the embodiment of Fig. 5, the core switching fabric operates with a substantially lower frequency thereby facilitating implementation of this switching fabric.

[0044] As described in the embodiments of the invention with reference to figures 3 through 5 the switching matrices are depicted as spatial switches operating on timescales relatively long. However, in alternate embodiments of the invention the switch matrices may be implemented with devices which operate at high speed and can be reconfigured as required for each and every time slot associated with a packet of data. Such matrices are usually referred to as time division multiplexing switch (TDM switches).

[0045] Within the embodiments outlined the multiple stages of memory switching can further be operated synchronously or asynchronously. With an asynchronous approach to a switching node the multiple stages of the switching node can be distributed with each one of the plurality of switching stages under localised clock control. In this the shuffle networks would be transmission links rather than local interconnections.

[0046] In respect of the technology used to implement the invention the architecture is independent and can be equally photonics or electronic but may be

weighted by their specific tradeoffs. Generally photonic switches are suited to smaller switching fabrics supporting very high throughput with typically limited memory queuing, whilst electronic switches support queues which hold for long periods of time, large fabrics but tend to suffer at supporting high speed as the conventional silicon platform is firstly replaced with silicon-germanium or gallium arsenide which have fewer design options for the building blocks of the switching node.

[0047] In respect of the packet switching node controller this may be implemented optionally to include polling elements, allowing the controller to provide additional control of the spatially separated memory switches such that they can be considered in operation as a single large switch matrix.

[0048] Numerous other embodiments may be envisaged without departing from the spirit or scope of the invention.

Claims

What is claimed is:

1. A switching node comprising:
a plurality of input ports for receiving input data packets;
a plurality of output ports for providing output data from the switching node;
a plurality of memory switches each comprising:
a plurality of first memory queues coupled for storing of packet data therein,
a first switch matrix for switching of packet data for storage within a memory queue of the plurality of first memory queues, and
a second switch matrix for switching of packet data retrieved from within a memory queue of the plurality of first memory queues;
a third switching matrix coupled between the plurality of input ports and the plurality of memory switches for routing input data packets from each of the plurality of input ports to the at least one of the plurality of memory switches;
a fourth switching matrix coupled between the plurality of memory switches and the plurality of output ports for routing data packets to at least one of the plurality of output ports from each of the plurality of memory switches,
wherein at least one of the third switching matrix and the fourth switching matrix comprises:
a plurality of second memory queues coupled for storing of packet data therein,
a fifth switch matrix for switching of packet data for storage within a memory queue of the plurality of first memory queues, and
a sixth switch matrix for switching of packet data retrieved from within a memory queue of the plurality of first memory queues.

2. A switching node according to claim 1 comprising a packet switch node controller.

3. A packet switch node controller according to claim 1 comprising at least two controller sections other than located physically together.

4. A switching node according to claim 1 wherein the packet switch node controller comprises polling elements.

5. A switching node according to claim 1 wherein the polling elements, in use, limit functionality of the plurality of memory switches in combination to that of a single cross-point memory queue.

6. A switching node according to claim 1 wherein the memory switches of at least one of the first switch stage, the second stage and the third stage comprise interfaces to the polling elements.

7. A switching node according to claim 1 wherein the memory switches of the first, second, and third switch stages comprise interfaces to the polling elements.

8. A switching node according to claim 1 comprises a packet switch node controller for routing data packets destined for a predetermined output port to predetermined memory queues associated with the predetermined output port.

9 A switching node according to claim 8 wherein the predetermined memory queues comprise the plurality of memory queues coupled to the fourth switch matrix.

10. A switching node according to claim 8 wherein the packet switch node controller comprises a processor for implementing a "round-robin" sequence in respect of scheduling packets forward to the memory switches.

11. A switching node according to claim 10 wherein the "round robin" sequence is prioritized based on packet priority.

12. A switching node according to claim 1 wherein the at least one of the third switching matrix and the fourth switching matrix comprises the third switching matrix.

13. A switching node according to claim 12 wherein the at least one of the third switching matrix and the fourth switching matrix comprises the fourth switching matrix.

14. A switching node according to claim 1 wherein the at least one of the third switching matrix and the fourth switching matrix comprises the fourth switching matrix.

15. A switching node according to claim 1 wherein the third switch matrix and plurality of memory queues comprise a first packet router conserving the data packets.

16. A switching node according to claim 15 wherein the switching node is a switching node that other than loses packet data.

17. A switching node according to claim 1 wherein the first, second and third switch matrices comprise switching elements operating as time division multiplexing switches.

18. A switching node according to claim 1 comprising:

a queuing controller in communication with the fourth switch matrix for queuing of packet data within the switching node, queuing of packet data for transmission from the node performed solely by the fourth switch matrix.

19. A switching node according to claim 18 comprising:

a packet switch node controller in communication with the first switch matrix and the third switch matrix for routing of packet data within the switching node prior to queuing of the packet data for transmission.

20. A switching node according to claim 1 wherein the second and third switch matrices comprise:

switching elements comprising time division multiplexing switches.

21. A switching node according to claim 20 wherein the switching matrices for the first and fourth switch matrices comprise:

switching elements for operating at a faster switching rate than those of the second and third switch matrices.

22. A switching node according to claim 20 comprising:

a queuing controller in communication with the second and third switch matrices for provisioning of bandwidth between the first and fourth switch matrices.

23. A switching node according to claim 20 wherein the switching matrices for the first and fourth switch matrices comprise:

switching elements comprising time division multiplexing switches.

24. A switching node according to claim 23 comprising:

a queuing controller in communication with the first, second, third and fourth switch matrices so as provide memory queuing only between second and third switch matrices.

25. A switching node according to claim 23 wherein the memory switches comprise;

switching elements operating at a rate slower than that of the input ports.

26. A method comprising:

receiving packet data;

routing of the received packet data via a switch for storage thereof within a multi-stage memory queue, where packet data having a lower priority is stored in an earlier stage of the multi-stage memory queue;

queuing of packet data from within the multi-stage memory queue for transmission thereof; and,

transmitting of the queued packet data via a switch.

27. A method according to claim 26 wherein the plurality of multi-stage memory queues store packet data received at the switching node.

28. A method according to claim 26 wherein the plurality of memory queues comprise three parts connected in series.

29. A method according to claim 28 wherein the first part of the memory queues comprises packet data associated with packets destined for adjacent output ports and assigned to a dedicated output stage memory switch.

30. A method according to claim 28 wherein the second part of the memory queues comprises packet data associated with packet data for packets stored within different memory queues which is assigned to a dedicated intermediate memory switch serving those queues.

31. A method according to claim 28 wherein the third part of the memory queues comprises packet data associated with packets with adjacent input ports and assigned to a dedicated input stage memory sector.

32. A method according to claim 28 wherein the stages of the plurality of memory queues comprise data fields arranged so as to provide a transposed interconnection between the input and intermediate stages.

33. A method according to claim 28 wherein the stages of the plurality of memory queues comprise data fields arranged so as to provide a transposed interconnection between the intermediate and output stages.

34. A method according to claim 26 wherein the plurality of the multi-stage memory queues are each divided into three parts connected in series.

35. A method according to claim 26 wherein the first and third parts of each of the plurality of multi-stage memory queues point to the same memory queue switching fabric.

36. A method according to claim 26 wherein the switching node function comprises:

the means to drop packets that cannot be routed.

37. A method according to claim 26 comprising:

a queuing controller in communication with the first, second, third and fourth switch matrices so as to ensure that cells are not mis-sequenced.

38. A method according to claim 26 comprising:

a packet switch node controller for controlling operation of the switching node.

39. A method according to claim 26 comprising:

a plurality of packet switch node controllers each controlling one of the plurality of switch stages.

40. A method according to claim 38 wherein the packet switch node controller comprises control for the plurality of memory queues such that they operate as a single cross-point memory queue.

41. A method according to claim 38 wherein the packet switch node controller comprises:

the means to prevent mis-sequencing of packets.

42. A method according to claim 38 wherein the packet switch node controller comprises:

the means to distribute packets across the memory switches.

43. A method according to claim 38 wherein the packet switch node controller comprises:

the means to balance the loading of packets across the plurality of memory switches

44. A method of routing packets within a switching node, comprising:

(a) initializing a first memory queue

(b) initializing a memory map corresponding to the memory queue

(c) setting the memory map pointer to its starting value

(d) detecting a packet of data, said packet of data being incident at one of a plurality of input ports of the switching node;

(e) performing an arrival process for the packet of data into the first stage of the switching node comprising;

updating the memory map to reflect the intended routing of the packet, and

addressing the packet of data to an appropriate element of the memory queue;

(f) performing a departure process for the packet of data from the first stage of the switching node; comprising;

searching the memory map for a packet of data at the head of each memory queue having the smallest time-stamp

extracting the packet of data for transport to the second stage of the switching node

(g) performing an arrival process for each memory queue of the second stage of the switching node, comprising;

identifying the packet of data by the memory source, and hence input port of the switching node and noting the intended routing of the packet

appending the packet of data to the memory queue selected

(h) performing a departure process for the second stage of the switching node, comprising;

sequentially taking each output port of the second stage of the switching node,

searching the memory map of the appropriate memory queue,

identifying the packet of data intended for the output port currently selected with the smallest time-stamp

preparing the packet of data for transport to the third stage of the switching node

(i) performing an arrival process for the third stage of the switching node, comprising;

classifying the arriving packet of data by the originating port of the switching node and the intermediate memory queue of the second stage from which the packet of data is extracted

appending the packet of data to the appropriate memory queue

(j) performing a departure process from the third stage of the switching node, comprising;

searching the memory queues for the packet of data within the memory queue associated with the currently selected output port and identifying within the memory map the packet of data with the smallest time-stamp

removing from the memory queue the selected packet of data

preparing the packet of data for transport out from the switching node

(k) incrementing the pointer index

(l) repeating steps (d) through (k) in looping manner until all pointers have been addressed

(m) repeating step (c) to reset the pointers and loop back

45. A method according to claim 44 wherein the arrival process (e) consists of:

time stamping the arrived packet of data;

identifying the output port of the switching node to which the packet of data is to be routed;

modifying the memory map in respect of the packet of data.

46. A method according to claim 44 wherein the departure process (f) instead comprises performing a "round-robin" sequence of the second stage of switching.

47. A method according to claim 44 wherein the departure process (f) of the first stage of the switching node is performed in a "round-robin" cycle for all the memory queues of the second switching stage.

48. A method according to claim 44 wherein the departure process (f) of the first stage of the switching node is performed solely on the basis of smallest time-stamp without any reference to memory queue sequence.

49. A method according to claim 44 wherein the method further comprises holding the packet of data in a final output queue if the switching node fabric is operating faster than the output port transport.

50. A method of routing packets of data within a switching node, comprising:

(a) initializing a memory map corresponding to the memory queue; said memory queue being divided into three parts connected in series, where the three parts are known as head, tail and body, such that all references for head and tail segments point to the same memory queue switching fabric;

(b) establishing a memory queue $q(i,j,k)$;

(c) an initialization of all tail-pointers $p_1(i,j)$ and head-pointers $p_3(i,j)$ to point to the same initial value.

(d) for each input sector i performing an arrival process for an arriving packet of data, comprising;

time-stamping said packet of data

establishing the classification of the packet of data by destination j

appending packet of data to tail-queue $q_1(i,j,p_1(i,j))$ and incrementing tail-pointer $p_1(i,j)$

(e) cycling through the memory queue, increment k for each timeslot;

(f) scanning over j the tail queues $q_1(i,j,k)$ in the same memory queue k and selecting the packet of data at the head of the queue with the smallest time-stamp to be transported to intermediate sector memory queue k;

(g) performing a departure process wherein the packet is routed through the intermediate sector memory queue k;

(h) scanning over i the body-queues $q_2(i,j,k)$ and select the packet at the head of the queue with the smallest time-stamp, for transport to output sector j;

(i) for each output sector j;

classifying the packet of data by source i and layer k

appending the packet of data to the head-queue $q_3(i,j,k)$

(j) performing a scan over i the head-queues $q_3(i,j,p_3(i,j))$ and select packet of dataat the head of the queue with the smallest time-stamp;

(k) dequeuing the packet of data ready for transmission

(l) incrementing the pointer

51. A method according to claim 50 wherein the $I \times J \times K$ logical links between the stages of the switching node can be shared using memory queues which preserves arriving packets of data; such that

sharing occurs on $I \times K$ and $K \times J$ physical links; and

each link is operating at a factor κ/K more slowly than the external incoming and external outgoing line rates respectively where κ is the speed-up, where κ is a positive factor.

52. A storage medium having stored therein data for when executed results in the routing of data packets within a switching node, by steps comprising:

(a) initializing a memory queue; wherein the memory queue stores packet data received at the switching node;

(b) initializing a memory map corresponding to the memory queue; said memory queue being divided into three parts connected in series, where the three parts are known as head, tail and body, such that all references for head and tail segments point to the same memory queue switching fabric;

(c) setting the memory map pointer to its starting value

(d) detecting a packet of data, said packet of data being incident at one of a plurality of input ports of the switching node;

(e) performing an arrival process for the packet of data into the first stage of the switching node comprising;

updating the memory map to reflect the intended routing of the packet, and

addressing the packet of data to an appropriate element of the memory queue;

(f) performing a departure process for the packet of data from the first stage of the switching node; comprising;

scanning through the memory map and identifying the packet of data at the head of each memory queue having the smallest time-stamp

extracting the packet of data for transport to the second stage of the switching node

(g) performing an arrival process for each memory queue of the second stage of the switching node, comprising;

identifying the packet of data by the memory source, and hence input port of the switching node and noting the intended routing of the packet

appending the packet of data to the memory queue selected

(h) performing a departure process for the second stage of the switching node, comprising;

sequentially taking each output port of the second stage of the switching node,

scanning the memory map of the appropriate memory queue,

identifying the packet of data intended for the output port currently selected with the smallest time-stamp

preparing the packet of data for transport to the third stage of the switching node

(i) performing an arrival process for the third stage of the switching node, comprising;

classifying the arriving packet of data by the originating port of the switching node and the intermediate memory queue of the second stage from which the packet of data is extracted

appending the packet of data to the appropriate memory queue

(j) performing a departure process from the third stage of the switching node, comprising;

scanning over the memory queues for the packet of data within the memory queue associated with the currently selected output port and identifying within the memory map the packet of data with the smallest time-stamp

removing from the memory queue the selected packet of data

preparing the packet of data for transport out from the switching node

(k) incrementing the pointer index

(l) repeating steps (d) through (k) in looping manner until all pointers have been addressed

(m) repeating step (c) to reset the pointers and loop back

53. A storage medium having stored therein data for when executed results in the routing of data packets within a switching node, by steps comprising:

(a) initializing a memory map corresponding to the memory queue; said memory queue being divided into three parts connected in series, where the three parts are

known as head, tail and body, such that all references for head and tail segments point to the same memory queue switching fabric;

(b) establishing a memory queue $q(i,j,k)$;

(c) an initialization of all tail-pointers $p_1(i,j)$ and head-pointers $p_3(i,j)$ to point to the same layer,

(d) for each input sector i performing an arrival process for an arriving packet of data, comprising;

time-stamping said packet of data

establishing the classification of the packet of data by destination j

appending packet of data to tail-queue $q_1(i,j,p_1(i,j))$ and incrementing tail-pointer $p_1(i,j)$

(e) cycling through the memory queue, increment k for each timeslot;

(f) scanning over j the tail queues $q_1(i,j,k)$ in the same memory queue k and selecting the packet at the head of the queue with the smallest time-stamp to be transported to intermediate sector memory queue k;

(g) performing a departure process by a "round-robin" cycle through all destinations , one destination j each timeslot,

(h) scanning over i the body-queues $q_2(i,j,k)$ and select the packet at the head of the queue with the smallest time-stamp, for transport to output sector j;

(i) for every output sector j;

classifying the packet of data by source i and layer k

appending the packet of data to the head-queue $q_3(i,j,k)$

(j) performing a scan over i the head-queues $q_3(i,j,p_3(i,j))$ and selecting the packet of data at the head of the queue with the smallest time-stamp;

(k) dequeuing the packet of data ready for transmission

(l) incrementing the pointer.

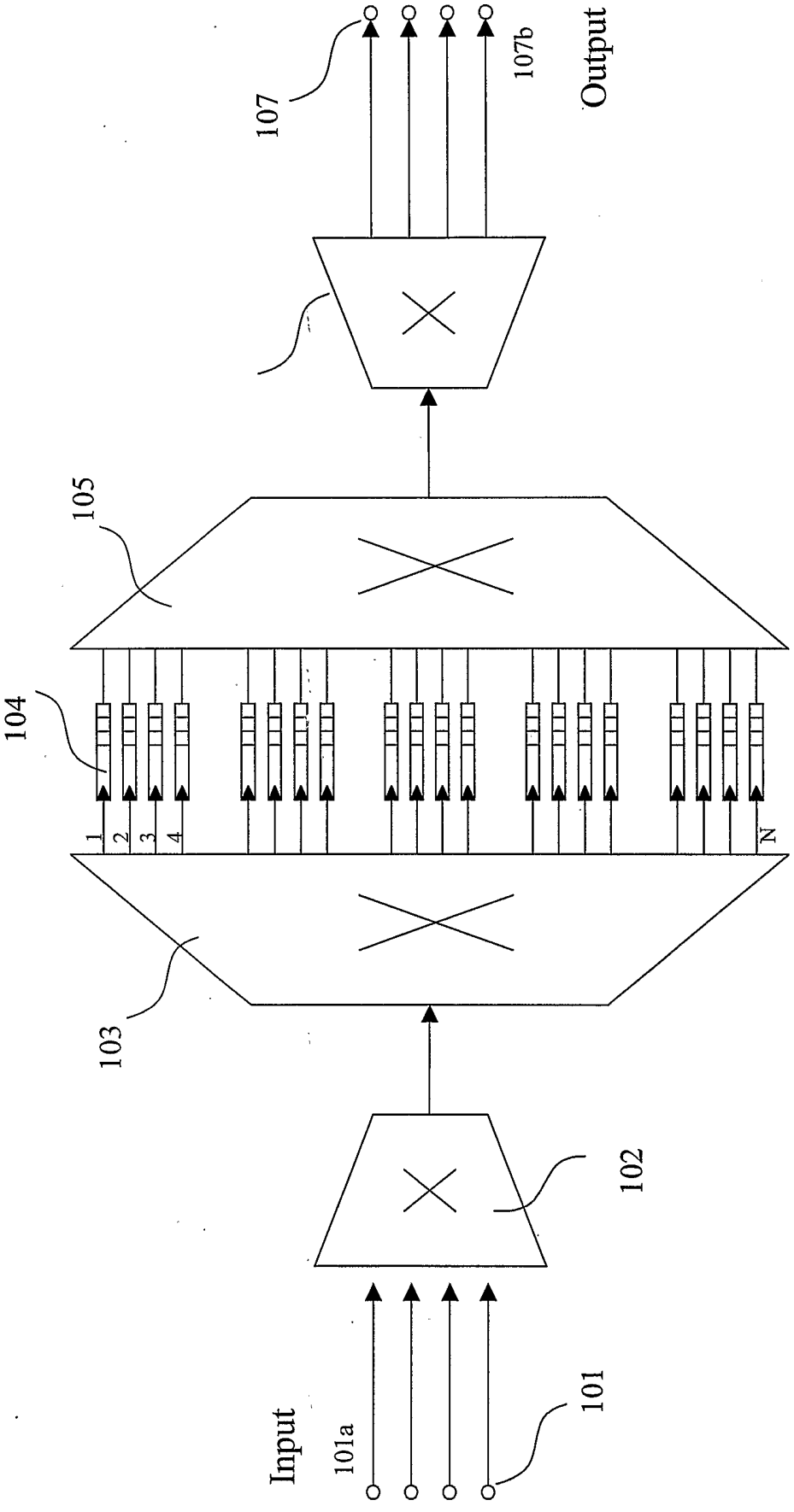


Fig 1

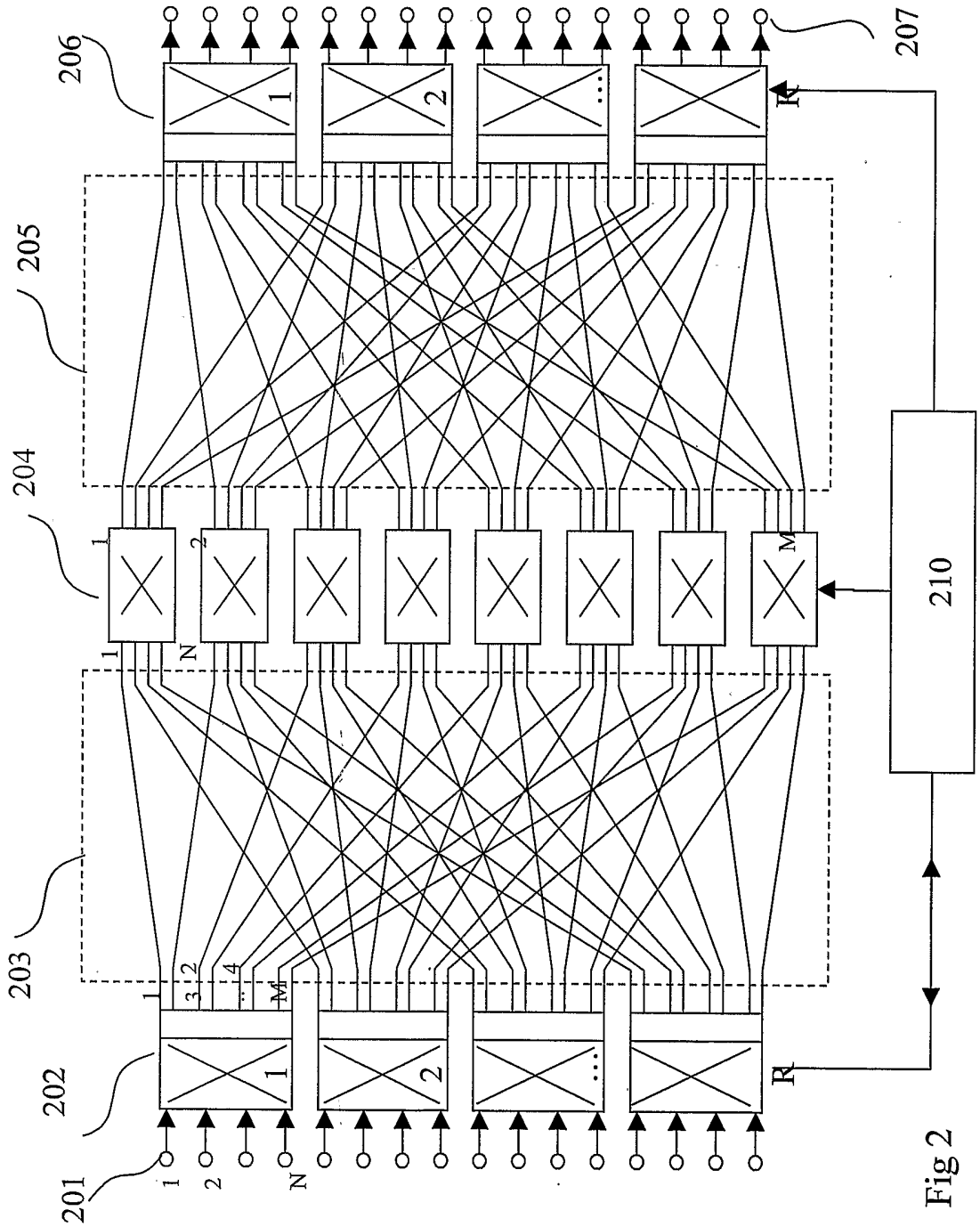


Fig 2

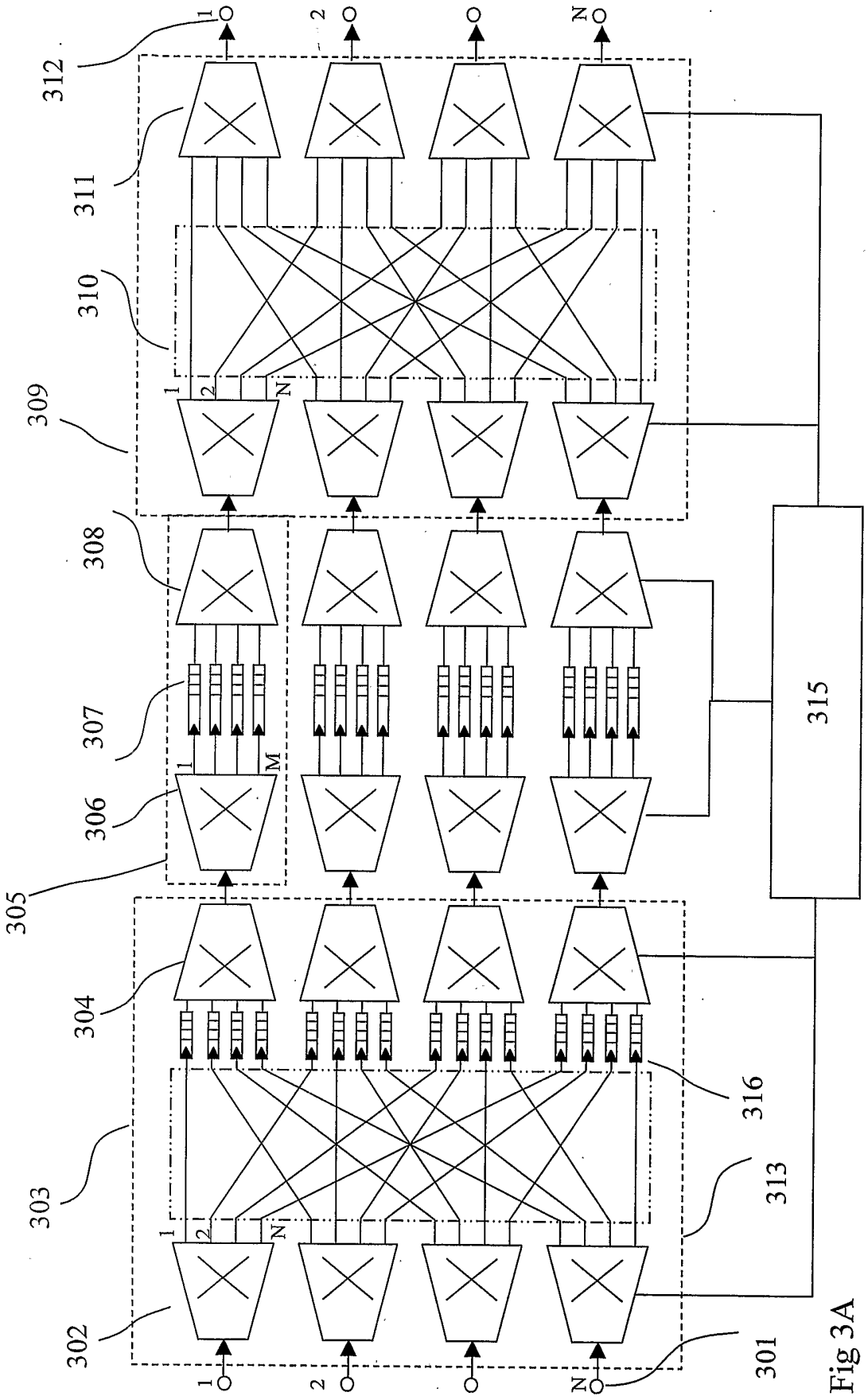


Fig 3A

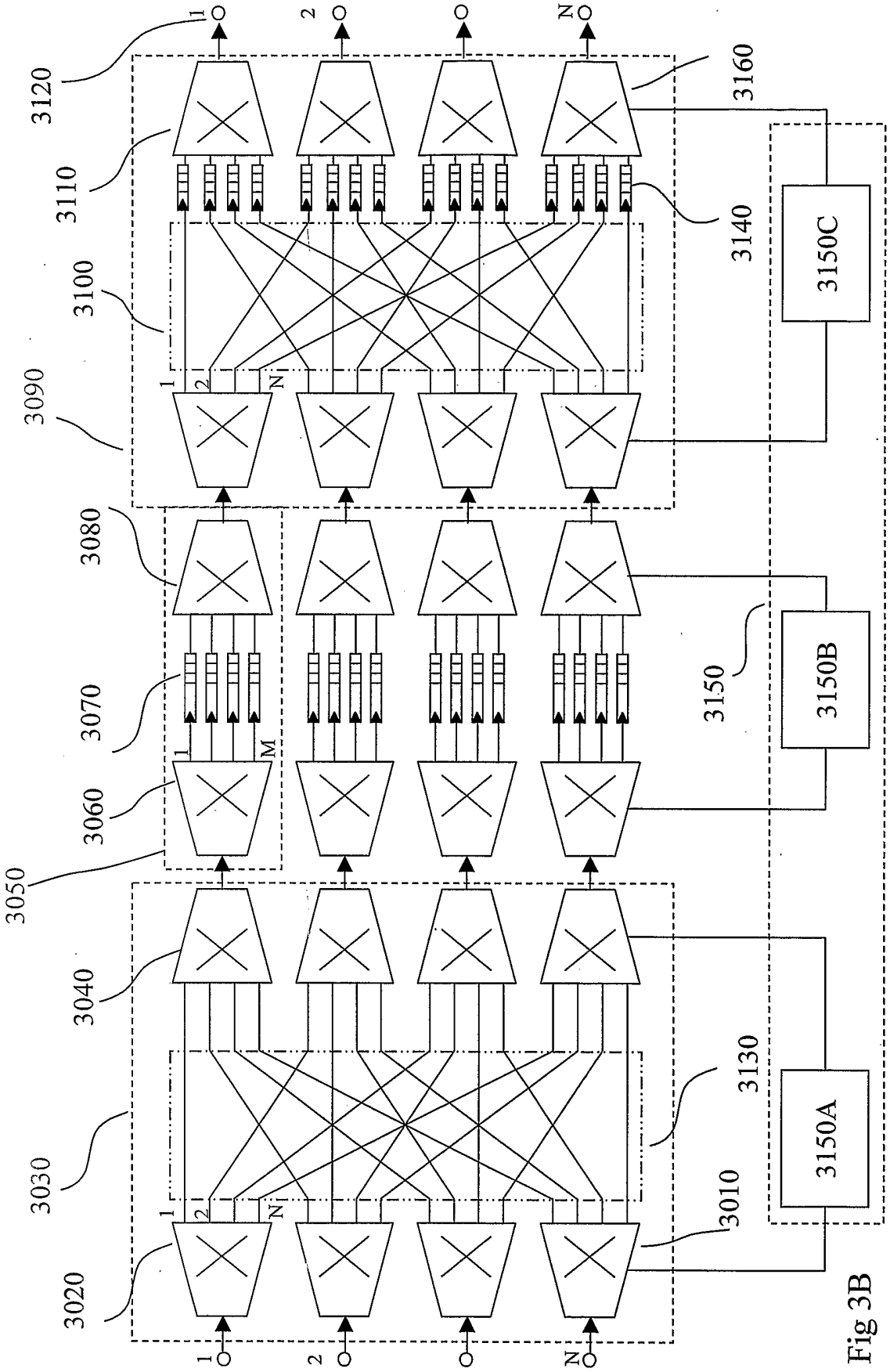

3120
3110
3100
3090
3080
3070
3060
3050
3040
3030
3020
3160
3140
3150C
3150
3150B
3130
3150A
3010
1
2
N
M

Fig 3B

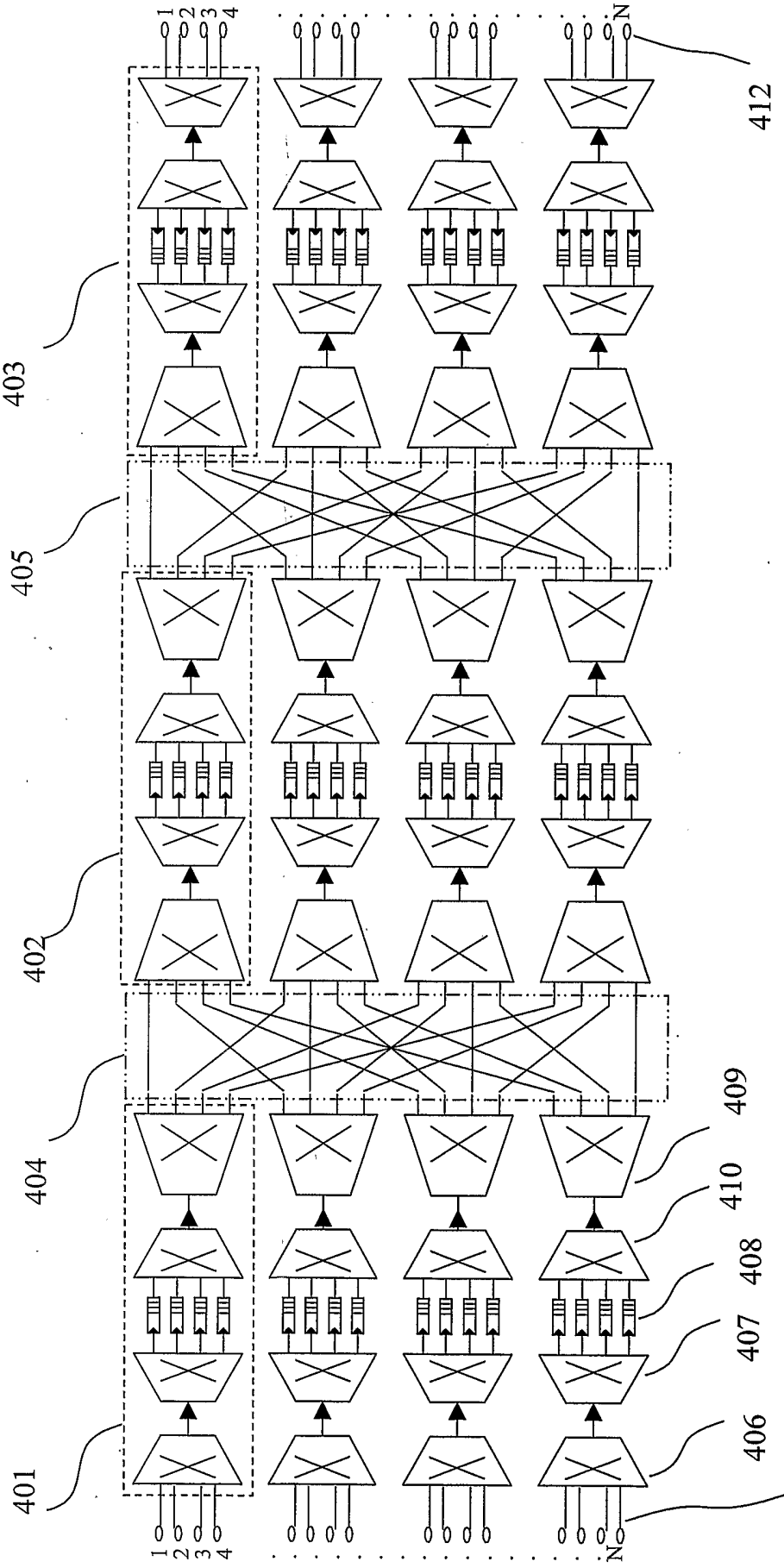


Fig 4A

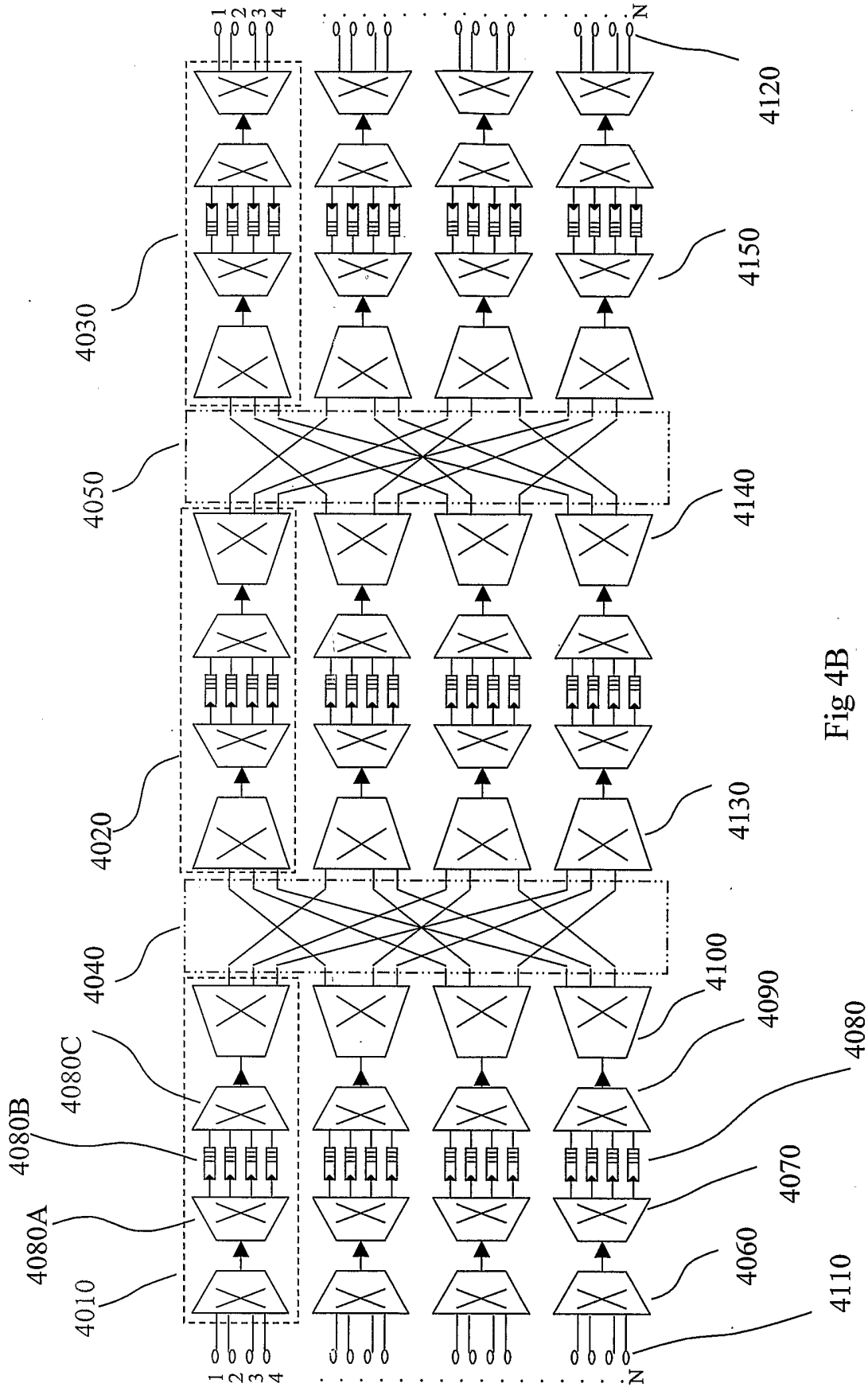

1
2
3
4
N
4120
4150
4030
4050
4140
4020
4130
4040
4100
4090
4080C
4080
4080B
4070
4080A
4060
4010
4110

Fig 4B

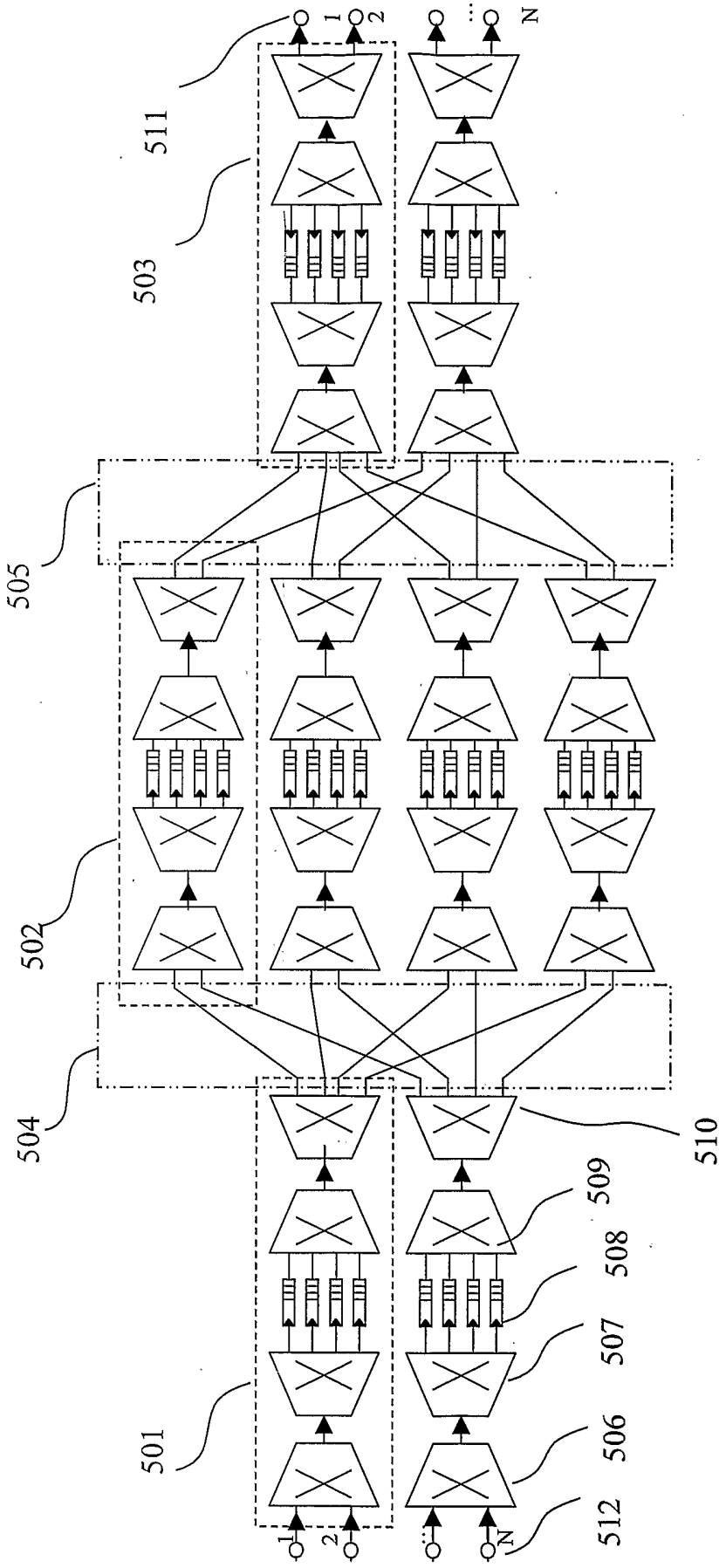


Fig 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CA2005/001913

A. CLASSIFICATION OF SUBJECT MATTER
IPC: ***H04L 12/56*** (2006.01)
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC: H04L 12/56 (2006.01)

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic database(s) consulted during the international search (name of database(s) and, where practicable, search terms used)
Delphion, EPAT, Canadian Patent Database, US Patent Database.
Keywords such as: switching fabric, switch matrix, memory switch, multi stage switch

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US2003/0133465 (ALFANO, Vic) 17 July 2003 (17-07-2003) -abstract -disclosure page 3, paragraph 34 -figure 4	26-35
X	US2004/0131068 (KORBER et al.) 08 July 2004 (08-07-2004) -abstract -disclosure page 3, paragraph 0053 -figures 2 and 3	26-35
A	WO2004/006517 (DUBOIS, Michel) 15 January 2004 (15-01-2004) -see whole document	1-53

[ ] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
20 March 2006 (20-03-2006)	29 March 2006 (29-03-2006)

Name and mailing address of the ISA/CA	Authorized officer
Canadian Intellectual Property Office Place du Portage I, C114 - 1st Floor, Box PCT 50 Victoria Street Gatineau, Quebec K1A 0C9 Facsimile No.: 001(819)953-2476	Hassan Bayaa (819) 997-7810

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/CA2005/001913

Patent Document Cited in Search Report	Publication Date	Patent Family Member(s)	Publication Date
US2003133465	17-07-2003	NONE	
US2004131068	08-07-2004	EP1432179 A1	23-06-2004
WO2004006517	15-01-2004	AU2003247125 A1	23-01-2004
		CA2491035 A1	15-01-2004
		CN1682497 A	12-10-2005
		EP1535429 A1	01-06-2005
		JP2005532729T T	27-10-2005
		US2004008674 A1	15-01-2004