



(86) Date de dépôt PCT/PCT Filing Date: 2000/08/14
(87) Date publication PCT/PCT Publication Date: 2001/03/08
(45) Date de délivrance/Issue Date: 2007/07/03
(85) Entrée phase nationale/National Entry: 2002/02/27
(86) N° demande PCT/PCT Application No.: US 2000/022368
(87) N° publication PCT/PCT Publication No.: 2001/017180
(30) Priorité/Priority: 1999/08/31 (US09/387,047)

(51) Cl.Int./Int.Cl. *H04L 12/56* (2006.01)

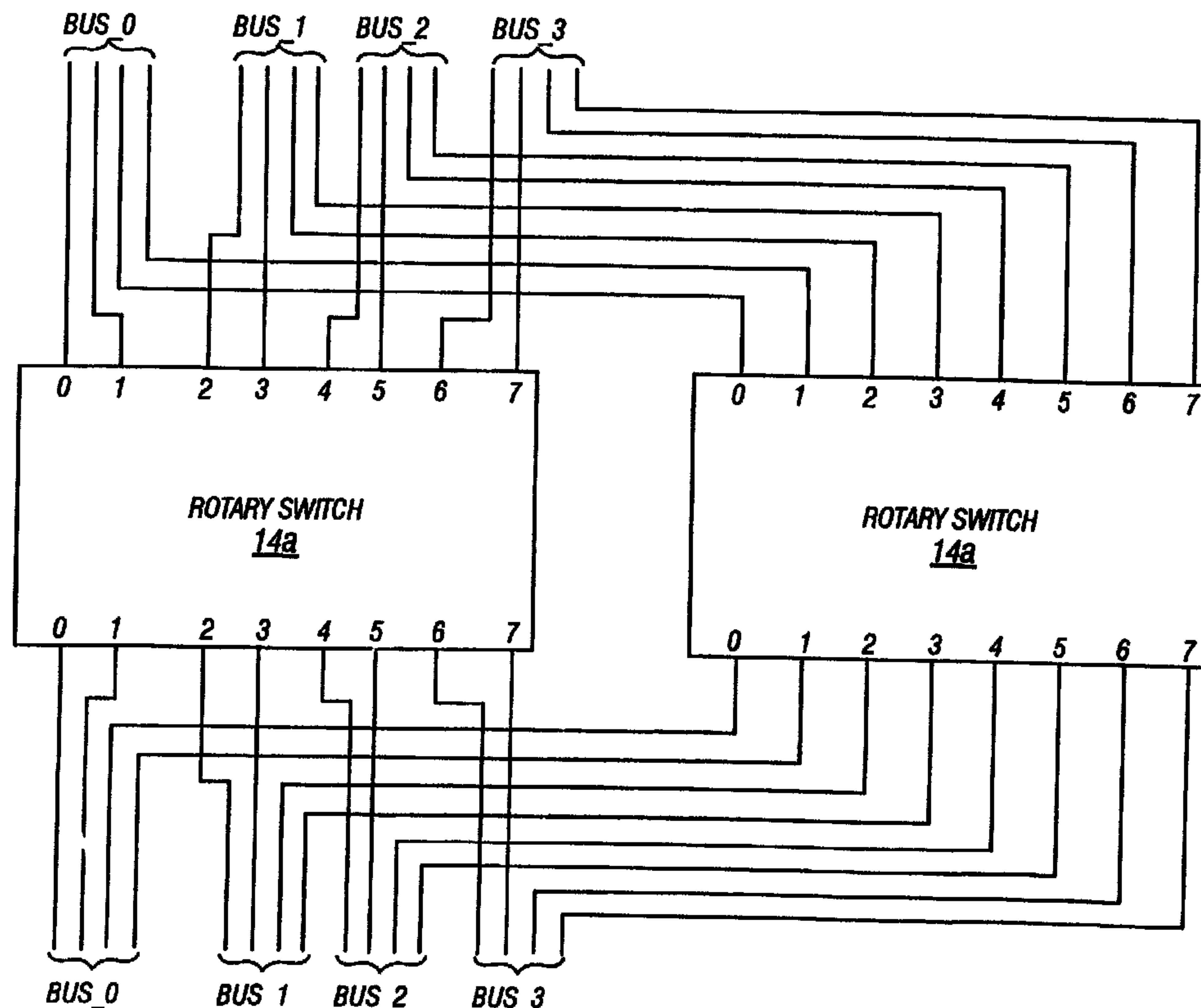
(72) Inventeurs/Inventors:
WOLRICH, GILBERT, US;
ADILETTA, MATTHEW J., US;
CYR, JOHN, US

(73) Propriétaire/Owner:
INTEL CORPORATION, US

(74) Agent: SMART & BIGGAR

(54) Titre : MATRICE DE COMMUTATION EVOLUTIVE

(54) Title: SCALABLE SWITCHING FABRIC



(57) Abrégé/Abstract:

A switch fabric includes a first plurality of data switches each having a plurality of input ports and a plurality of output ports the plurality of switches capable of switching any of its input ports to any of its output ports with the plurality of data switches having inputs coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to a first one of the plurality of switches, and a succeeding byte of the first input bus is coupled to a succeeding one of the plurality of switches.

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

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
8 March 2001 (08.03.2001)

PCT

(10) International Publication Number
WO 01/17180 A1(51) International Patent Classification⁷: **H04L 12/56**

(21) International Application Number: PCT/US00/22368

(22) International Filing Date: 14 August 2000 (14.08.2000)

(25) Filing Language: English

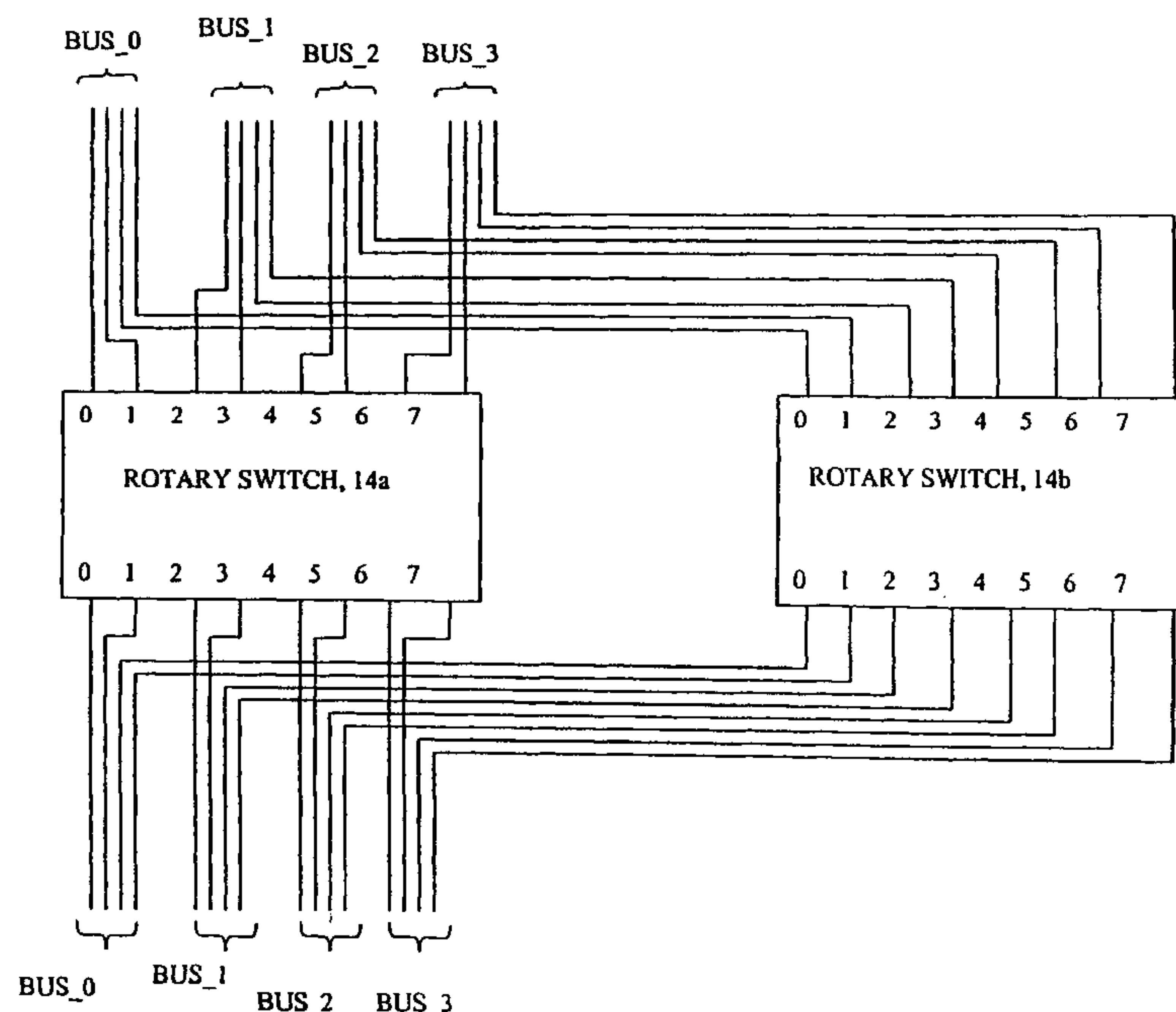
(26) Publication Language: English

(30) Priority Data:
09/387,047 31 August 1999 (31.08.1999) US(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 09/387,047 (CON)
Filed on 31 August 1999 (31.08.1999)(71) Applicant (for all designated States except US): **INTEL CORPORATION** [US/US]; 2150 Mission College Boulevard, Santa Clara, CA 95052 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **WOLRICH,**Gilbert [US/US]; 4 Cider Mill Road, Framingham, MA 01701 (US). **ADILETTA, Matthew, J.** [US/US]; 20 Monticello Drive, Worcester, MA 01603 (US). **CYR, John** [US/US]; 36 Leonard Road, Boxborough, MA 01719 (US).(74) Agent: **HARRIS, Scott, C.**; Fish & Richardson P.C., Suite 500, 4350 La Jolla Village Drive, San Diego, CA 92122 (US).(81) Designated States (*national*): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.(84) Designated States (*regional*): ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: **SCALABLE SWITCHING FABRIC**

(57) Abstract: A switch fabric includes a first plurality of data switches each having a plurality of input ports and a plurality of output ports the plurality of switches capable of switching any of its input ports to any of its output ports with the plurality of data switches having inputs coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to a first one of the plurality of switches, and a succeeding byte of the first input bus is coupled to a succeeding one of the plurality of switches.

WO 01/17180 A1

WO 01/17180 A1



Published:

- *With international search report.*
- *Before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments.*

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.

76307-71

SCALABLE SWITCHING FABRIC

BACKGROUND

This invention relates to switching fabrics used to switch data in computer networks and other data moving applications.

Crossbars are one type of switching fabric used to switch data between pluralities of devices. They can be thought of as a switch that has a plurality of vertical paths interconnected by switching elements to a plurality of horizontal paths in a manner that the switch elements can interconnect any one of the vertical paths to any one of the horizontal paths. Generally such crossbars are implemented with custom application specific integrated circuits (ASIC's).

SUMMARY

According to an aspect of the present invention, there is provided a switch fabric, comprising: first and second data switches each having a plurality of input ports and a plurality of output ports, the first and second data switches capable of switching any of its input ports to any of its output ports, each of the first and second data switches comprising: a plurality of virtual queues; input segment logic coupled to at least one bus, said input segment logic to determine to which virtual queue incoming data should be sent; output segment logic to select which virtual queue should be connected to an output port; and a distributed switching arrangement to provide a non-blocking cross-bar switching fabric arrangement capability over a series of byte sliced buses, the distributed switching arrangement coupling the virtual queues to the output

76307-71

segment logic through the non-blocking cross-bar switching arrangement, the non-blocking cross-bar switching fabric arrangement capability comprising a plurality of multiplexers coupled to outputs of the plurality of the
5 virtual queues to select ones of the virtual queues to feed to a second plurality of multiplexers that produce inputs to the output segment logic; said first and second data switches being coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to
10 one of the input ports of the first switch, and a last byte of the first input bus is coupled to one of the input ports of the second switch.

According to another aspect the invention provides a switch fabric, comprising: a plurality of data switches
15 each having a plurality of input ports and a plurality of output ports, the plurality of switches capable of switching any of its input ports to any of its output ports, the plurality of data switches, comprising: a plurality of virtual queues; input segment logic coupled to at least one
20 bus, said input segment logic to select to which virtual queue incoming data should be sent; output segment logic to select which virtual queue should be connected to an output port; and a distributed switching arrangement to provide a non-blocking cross-bar switching fabric arrangement
25 capability over a series of byte sliced buses, the distributed switching arrangement coupling the virtual queues to the output segment logic through the non-blocking cross-bar switching arrangement, the non-blocking cross-bar switching fabric arrangement capability comprising a
30 plurality of multiplexers coupled to outputs of the plurality of the virtual queues to select ones of the virtual queues to feed to a second plurality of multiplexers that produce inputs to the output segment logic; said input

76307-71

ports of plurality of data switches being coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to one of the input ports of a first one of the plurality of switches, and a succeeding
5 byte of the first input bus is coupled to one of the input ports of a succeeding one of the plurality of switches.

According to another aspect the invention provides a switch fabric comprising: a first data switch having a plurality of inputs and outputs; a second data switch having
10 a plurality of inputs and outputs, each data switch, comprises: a plurality of virtual queues; input segment logic coupled to at least one bus, said input segment logic to select to which virtual queue incoming data should be sent; output segment logic to select which virtual queue
15 should be connected to an output port; and a distributed switching arrangement to provide a non-blocking cross-bar switching fabric arrangement capability over a series of byte sliced buses, the distributed switching arrangement coupling the virtual queues to the output segment logic
20 through the non-blocking cross-bar switching arrangement, the non-blocking cross-bar switching fabric arrangement capability comprising a plurality of multiplexers coupled to outputs of the plurality of the virtual queues to select ones of the virtual queues to feed to a second plurality of
25 multiplexers that produce inputs to the output segment logic, wherein the distributed switching arrangement has the inputs of the first and second data switches coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to one of the inputs of the
30 first switch, and a last byte of the first input bus is coupled to one of the outputs of the second switch.

76307-71

One or more of the following advantages may be provided by one or more aspects of the invention.

A high performance, scalable switching fabric is provided for scaling a rotary switch for a multitude of
5 ports. The rotary switch uses virtual queuing providing the rotary switch controller (RSC) full crossbar capability, such that any of its input queues can couple to any of its output queues without blocking. The RSC permits dynamic configuration of additional ports. The RSC is a modular
10 concept allowing a switch to grow from e.g., 32 ports to 64 ports to 128 ports using a passive backplane.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a network system including a rotary switch.

15 FIG. 2 is a block diagram showing an implementation of the rotary switch.

FIG. 3 is a block diagram of the rotary switch.

FIG. 4A is a block diagram of a rotary switch coupled in a byte sliced configuration.

FIG. 4B is a block diagram of two rotary switches coupled in a byte sliced configuration.

5 FIG. 5A is a chart diagram showing byte mapping for a single rotary switch of FIG. 4A.

FIG. 5B is a chart diagram showing byte mapping in the device of FIG. 4B.

10 FIG. 6 is a block diagram showing structures used in arbitration in the rotary switch.

DESCRIPTION

Referring to FIG. 1, a networked system 10 includes a rotary switch 12 transferring data from input ports to
15 output ports in a non-blocking manner. For instance, the switch can be used for sending packet data at plurality of network processors 14 to network devices 16 coupled to separate 32 bit FIFO busses. The rotary switch 12 includes a data path that is sliced or partitioned on an 8 bit (i.e.
20 byte) basis to allow system 10 to be expanded from e.g., a 2 x 2 FIFO bus system to an 8 x 8 FIFO bus system and so forth as described below. The rotary switch 12 includes a plurality of 8-bit wide Virtual Input Queues ("VIQ") 18 that are distributed into input segments with each of 8 byte
25 wide input segments coupled to 16 of the VIQ's 18. The rotary switch device 12 also includes a plurality of Output Segments 20. The rotary switch 12 also includes a switching fabric network 24 which in combination with the Virtual Input Queues and output segment logic 20 can move
30 byte wide data from any of the plurality of Input Virtual

76307-71

Queues 18 to an output FBUS byte without restricting the access of any other input segment to any of Input Virtual Queues 18. Data switching is controlled by an arbiter 22.

The internal fabric 24 of the RSC 12 provides full input to output connectivity that is, a full of any and all of the inputs to any and all of the outputs. In an exemplary rotary switch 12, there are 128 8-bit wide Virtual Input Queues ("VIQ"), distributed over 8 segments e.g., 16 VIQ's per segment and 8 output segments, thus providing a device 12 having 128 input ports to any of the 128 output ports. The fabric is an independently queued structure that does not require symmetric switching. The switching is a distributed function between a loose arrangement among processors 14 and the RSC fabric 24. The RSC arbiter 22 provides a fair round robin service for received packets. The processors 14 can provide packets either through a simple round robin or weighted fair queuing to the RSC 12. The output port switching is based on a PULL arbitration scheme.

Referring to FIG. 2, an implementation of the system 10 of FIG. 1, is shown. A rotary switch device 12 is shown coupled to a pair of network processors 14. The network processors are preferably parallel-based multithreaded processors. One example of such a processor is described in US Patent No. 6,606,704. Each of the processors 14 communicate with data supplying devices 13 e.g., here Media Access Controllers (MAC'S) that are coupled to the physical layer of a network.

The system 10 also includes a passive backplane 30. The passive backplane 30 employs tri-state

76307-71

steering logic to enable dynamic reconfiguration of the system 10 based on the number of ports supported. This system 10 is a byte slice arrangement. As a byte sliced arrangement, when new ports are added, all ports stop
5 transmitting to RSC devices 12. Depending on buffering and steering initialization time, input ports may or may not have to be paused. The passive backplane includes nine (9) main busses 30a-30b. The first bus 30a is a computer bus e.g., a Personal Computer Interconnect (PCI) bus. While
10 this is bridged bus, and therefore strictly speaking the backplane is not passive, the bridge and microprocessor unit that is commonly associated with busses such as the PCI bus can be provided as a daughter card to maintain a passive backplane 30. The other 8 busses on the backplane are used
15 to interconnect the RSC blades Blade_0-Blade_3. A RSC blade is an arrangement of RSC devices 12, network processors 14 and network devices 16.

Since the FBUS data (32 bit Unidirectional busses) is sectioned into up to four 8 bit segments. The segments
20 can be sized based on the number of equivalent ports supported by the system 10. If a single blade is used i.e., a 32 port system, the backplane steering logic connects Bus 30b to Bus 30e and Bus 30i to Bus 30f to provide input data, i.e., two 24 bit FBUS data busses to the RSC 12.
25 These two busses along with immediate feedback of 8 bits each from Bus 30b and Bus 30i form two 32 bit input busses to the RSC 12. Bus 30d is steered to Bus 30c for transmit to the MAC

device 16. The 24 data bits of FBUS buses along with the 8 bits from the RSC 12 form 32 bits to the MAC. Similarly, Bus 30g is steered to Bus 30h and merged with 8 bits of immediate feedback data from the RSC 12 to form 32 bits to the MACs 16.

Each blade e.g., Blade_0 to Blade_3 is similar in construction and are scaled together via the passive backplane 32. Thus, if system 10 has 64 ports supported, the steering logic (not shown) selects the appropriate bytes from each of the busses and connects them to their respective destination busses in 16 bit sections. If 128 ports are instantiated, then the section size is 8 bits. It should be noted that a 64 port system can be configured into a 96 port system where the section size is 8 bits, and the three RSC blades are used, with only 6 input segments (instead of the possible 8). In this type of configuration, a fourth RSC is required to complete the 32 bit byte-slice. Therefore, 96 port systems require the use of a simple add-in board which is populated with only an RSC using bits (31:8) on both input sections.

Referring to FIG. 3, the RSC 12 includes input segment logic devices (ISL) 40a-40h that handles incoming FBUS data and distributes the incoming data to an appropriate virtual input queue in virtual queue logic (VQL) devices 42a-42p. The RSC 12 also includes output segment logic devices (OSL) 44a-44h. The OSL devices 44 pull data from the VQL logic devices 42 and deliver the data to the output side of the FBUS for distribution to appropriate MAC devices 14. The input virtual queues 42₀ to 42₁₂₇ are coupled to the output segment logic 44 via a series of multiplexers 47₀ to 47₁₂₇.

Each of the multiplexers 42_0 to 42_{127} is coupled to each of the virtual input queues 42 in its row e.g., VIQ 42_0 - 42_{112} for multiplexer 42_0). There are sixteen of said multiplexers in each column. The output of the
5 multiplexers 47_0 to 47_{127} for each of the columns (e.g., multiplexer 47_0 - 47_{15} for column 0) feed corresponding output multiplexers 49_0 to 49_7 which are coupled to the output segment logic 44a -44h.

The RSC 12 also includes Input Ready logic 46 that
10 samples the virtual input queues in the virtual queue logic (VQL) devices 42a-42p to report back the status of the virtual input queues to mapped input segments, so that the devices 14 (FIG. 1) that supply data can track buffer fullness. The RSC 12 also includes output ready logic 48.
15 The Output Ready Logic (ORL) 48 is analogous to the Input ready logic. However, the ORL 48 samples the network devices 16 e.g., MAC transmit ready bits to determine if the network devices 16 are ready to accept more transmit data. The RSC 12 also includes Output Segment Arbitration
20 Logic 50. The Output Segment Arbitration Logic 50 as will be described in FIG. 6, is used to determine which virtual input queue should provide data to its output segment in an appropriate timeslot. One preferred scheme has the Output Segment Arbitration Logic 50 using a round robin time
25 multiplexed arbitration algorithm.

The Input Segment Logic (ISL) 40 interfaces with the network processor 14 and determines which virtual queue an incoming mpkt (64 byte payload) should be sent to. The RSC 12 has a plurality of virtual queues. In one example,
30 there are 16 virtual queues to which each input segment can

direct incoming mpkts. The RSC 12 is arranged into segments e.g., 8 input segments. If fewer input segments are required (i.e., the RSC 12 is configured for fewer ports), then logically contiguous input segments are joined
5 to form either a 32 bit datapath (32 ports), or a 16 bit datapath (64 ports).

The ISL 40 uses in-band information to control virtual queue loading and port arbitration. In-band information is used to minimize pin cost as would be associated with
10 explicit out-of-band control. The in-band information includes a destination output port (8 bits), an SOP bit, a "Transmit As Is" control bit, a byte enable control bit, a CSR enable, and a Virtual Queue Identifier (4 bits). Since, there is only 8 bits of in-band data available per
15 cycle, and there are 16 bits required, 2 in-band cycles are required. There is an optimization, in 64 or 32 port modes, 16 bits/32 bits of in-band information are possible, therefore only one in-band information cycle is required.

Input data to the crossbar includes two 32 bit header
20 words containing control information followed by up to eight (8) 32 bit words of packet data. The 32 bit input word is connected to 4 input segments. The header is segmented into 2 bytes of control for each input segment, and specifies the VIQ to load, the output destination, byte
25 count, end of packet, and byte enable for the last 32 bit word being transmitted. The 4 output segments specified receive the VIQ addresses of new packets being loaded in a "pending" output FIFO. All packets being sent to the same output port are loaded into a similar "pending" FIFO so that
30 all four output byte segments begin to send data to the

output FIFO bus on the same cycle. Byte data from the 4
output segments is combined to form the 32 bit output FIFO
bus.

Two output segments per RSC are enabled to drive
5 control signals (start of packet, end of packet, transmit
as is, transmit error), while all segments drive byte
enable signals. The output control logic samples the ready
signals for all output destinations. All output segments
update their ready bit status in lock step so that the 4
10 sliced bytes of the input FIFO bus can be switch at the
same time.

The inputs to the input segment logic 42 include the
FBUS data bits (7:0), and control signals, TxSel, EOP, and
NewQHdr. The FBUS data bits are as above, the TxSel bits
15 are used to frame the FBUS data bits as valid, whereas, EOP
is used to explicitly identify the end of a packet. The
NewQHdr bit indicates to the ISL 420 that a new set of
Virtual Queue information is coming. An optimization may
be that if EOP AND NewQhdr are asserted this would require
20 only a single prepend cycle to indicate a target Virtual
Queue. The implication in that case is that the transfer
is not a new packet but rather continuation data from a
currently transmitting packet. In this case an in-band EOP
is required.

25 The Input Ready Logic (IRL) 46 samples the status of
the 16 virtual queues in virtual queue logic 42a-42p
associated with each input segment. If a VIQ has available
space then the IRL will report that to a requesting network
processor 14 (FIG. 1), via VIQ transmit ready bits.

30

The network processor 14 can use this information to schedule transfers to the virtual queues.

The virtual queues VIQ's are associated with a particular output while there is valid data maintained in the virtual queue. The virtual queues can have a suitable
5 storage depth e.g., 4 mpkts for 14 of the queues and to 8 mpkts for two, where each mpkt is 64 bytes. There are 16 Virtual queues associated with each input/output segment.

Other arrangements are possible. Each VIQ has an
10 input pointer and an output pointer. The input pointer is used by the Input Segment Logic 40 (ISL) to push data into the VIQ, while the output pointer is used by the output Segment Logic to "pull" data from the VIQ for distribution out the transmit FBUS. In one implementation, the VIQ's
15 are single ported random access memory devices. Since a read and write may be concurrently required for the full crossbar operation, the VIQ's are cycled twice as fast as the input fill rate. For example, if the input fill rate is 66-80Mhz from the input segment FBUSES, the output drain
20 rate would be a decoupled 66-80Mhz FBUS drain rate then the VIQ's would operate at 133 to 166 MHZ that is twice as fast as the faster of the Input or Output FBUS rates. Alternatively, the queues can be orgainized as 2 bytes in width and accessed on alternate cycles.

25 In order to maximize the efficiency of the Rotary switch 12, the switch fabric operates at twice the output FBUS drain rate. One way to accomplish this would be to cycle the VIQ's twice again as fast. Another way would be to have the VIQ's be twice as wide. Thus, if the VIQ's
30 input segment is 8 bits wide the VA's are buffered to form

16 bits of write data. Read operations will fetch 16 bits of read data which will be supplied to the switch fabric at 133-160Mhz 8 bit chunks.

The Output Segment Logic (OSL) 44 is a timeslot filler. The Output Segment Logic 44 uses Output Segment Arbitration 50 results to select which new Virtual Queue should be "connected" to an output port. The OSL examines Transmit ready bits which are collected by the Output Ready bit Logic (ORL) 48 to determine if the output port is ready to accept a new mpkt (64 bytes). The Output Segment Logic decouples the VIQ crossbar logic from the output drain rate, by employing a 16 mpkt queue at each output segment ($16 * 64B * 8 = 8KB$). This decoupling allows the crossbar to operate at a higher frequency. The OSL 44 includes a 16 entry timeslot queue. Each VIQ to Output Port has an explicit timeslot entry. If a VIQ is not available, its slot may be compressed. Up to n slots may be compressed (most likely $n = 2$) before filling is "wait stated" until skipped VIQ's are available.

The Output Ready Logic (ORL) 48 interrogates the destination network devices 16 (FIG. 1) for transmit ready bits. The transmit ready bits are used by the RSC 12 in the Output Segment Logic 44 (OSL) to promote data from the RSC 12 to the appropriate output segment FBUS 31. The ORL 48 is a ready bus Master. It cycles through all attached MAC's fetching the transmit ready bits. The ORL 48 assembles all the transmit ready bits and provides them to their respective output segment. The OSL 44 uses these bits to determine if the tail of the queue should be filled

30

with that output port's mpkt. This is done to avoid head of queue blocking.

The Output Segment Arbitration (OSA) 50 is used to link a virtual queue 42 (VIQ) to an output port. The RSC 12 employs a distributed crossbar selection scheme. The network processor 14 performs weighted fair queuing and provides the top elements for transmission to the RSC 12. The RSC 12 in turn uses a fair service algorithm and a non-blocking scheme so that efficiency is maintained.

10

Switch Arrangements

Referring to FIG. 4A, a rotary switch 14a, is coupled to provide a 2 x 2 FIFO BUS switching fabric. The rotary switch 14a is fed by buses B0-B1 which are each 32 bit byte sliced busses. The output of the rotary switch 14a is coupled to output busses e.g., FBUS_0, FBUS_1. On the input side, the four bytes of each bus B0-B1 are coupled in sequence to the rotary switch 14a and on the output side the first four output segments of each rotary switch provide the bytes of FBUS_0, the next four output segments provide the bytes of FBUS_1. The mapping for this arrangement is shown in FIG. 5A. In this manner a byte sliced architecture is provide. This byte sliced architecture, is non-blocking. That is, any input port can be connected to any output port without blocking any other input port from connecting to any other output port. In any one cycle, all input ports can couple data to different ones of all of the output ports.

Referring to FIG. 4B, a pair of rotary switches 14a, 14b are coupled to provide a 4 x 4 FIFO BUS switching

30

fabric. The rotary switches 14a, 14b have input segments coupled by buses B0-B3 which are each 32 bit byte sliced buses. The output segments of the rotary switches 14a, 14b are coupled to output busses e.g., FBUS_0 to FBUS_3. On
5 the input side, the first two bytes of each bus B0-B3 are coupled to the input segments of the first rotary switch 14a, whereas, the last two bytes of each bus are coupled to the input segments of the second rotary switch 14b. On the output side the first two output segments of each rotary
10 switch provide the bytes of FBUS_0, the next two output segments the bytes of FBUS_1 and so forth. The mapping for this arrangement is shown in FIG. 5B. In this manner a byte sliced architecture is provide. This byte sliced architecture, is a 4 x 4 architecture and is non-blocking.
15 That is, any input port can be connected to any output port and not block any other input port from connecting to any other output port. In any one cycle, all input ports can couple data to different ones of all of the output ports.

Thus, rotary switches can be coupled to provide larger
20 switching fabrics. Four switches (a mapping of which is set out below) can be coupled such that eight, 4 byte buses could be coupled to the four switches with first bytes of a each bus coupled to the first switch, second bytes of each bus coupled to the second switch, third bytes of each bus
25 coupled to the third switch and fourth bytes of each bus coupled to the fourth switch. Moreover, with larger rotary switches i.e., that can interface to larger buses, e.g., 8 byte buses, even larger configurations could be provided in a similar manner. On the output side a similar connection
30 arrangement is provided.

This switching fabric is scalable, i.e., easily expanded from a 2 x 2 FIFO bus configuration (32 ports to 32 ports) up to an 8 x 8 FIFO bus configuration (128 ports to 128 ports) without adding additional hierarchical levels of switches. That is, expansion occurs on a single level of switches which reduces latency and complexity.

Referring to FIG. 5A, the mapping as a 2 x 2 FIFO bus switching fabric requires one RSC 12, with bytes output mapped as follows:

10

1. output segment 0 - byte 0 (bits(31:24)) output FIFO bus 0
2. output segment 1 - byte 1 (bits(23:16)) output FIFO bus 0
3. output segment 2 - byte 2 (bits(15:08)) output FIFO bus 0
4. output segment 3 - byte 3 (bits(07:00)) output FIFO bus 0

15

5. output segment 4 - byte 0 (bits(31:24)) output FIFO bus 1
6. output segment 5 - byte 1 (bits(23:16)) output FIFO bus 1
7. output segment 6 - byte 2 (bits(15:08)) output FIFO bus 1
8. output segment 7 - byte 3 (bits(07:00)) output FIFO bus 1

20

and the input segments mapped as;

25

1. input segment 0 - byte 0 (bits(31:24)) input FIFO bus 0
2. input segment 1 - byte 1 (bits(23:16)) input FIFO bus 0
3. input segment 2 - byte 2 (bits(15:08)) input FIFO bus 0
4. input segment 3 - byte 3 (bits(07:00)) input FIFO bus 0

30

5. input segment 4 - byte 0 (bits(31:24)) input FIFO bus 1
6. input segment 5 - byte 1 (bits(23:16)) input FIFO bus 1
7. input segment 6 - byte 2 (bits(15:08)) input FIFO bus 1
8. input segment 7 - byte 3 (bits(07:00)) input FIFO bus 1

where FBUS_xY corresponds to byte "x" of FBUS "Y." Thus, the output segments 0, 1, 2, 3 connect concurrently to corresponding Virtual Queues either in input segments 0, 1, 2, 3 or 4, 5, 6, 7, respectively.

Referring to FIG. 5B, to expand to a 4 x 4 FIFO bus switching fabric requires two RSC devices 14. FIG. 5B shows the output mapping with the output mapped as,

5	1.	RSC_0	output	segment 0	- byte 0 (bits(31:24))	output FIFO bus 0
	2.	RSC_0	output	segment 1	- byte 1 (bits(23:16))	output FIFO bus 0
	3.	RSC_0	output	segment 2	- byte 0 (bits(31:24))	output FIFO bus 1
	4.	RSC_0	output	segment 3	- byte 1 (bits(23:16))	output FIFO bus 1
10	5.	RSC_0	output	segment 4	- byte 0 (bits(31:24))	output FIFO bus 2
	6.	RSC_0	output	segment 5	- byte 1 (bits(23:16))	output FIFO bus 2
	7.	RSC_0	output	segment 6	- byte 0 (bits(31:24))	output FIFO bus 3
	8.	RSC_0	output	segment 7	- byte 1 (bits(23:16))	output FIFO bus 3
15	9.	RSC_1	output	segment 0	- byte 2 (bits(15:08))	output FIFO bus 0
	10.	RSC_1	output	segment 1	- byte 3 (bits(07:00))	output FIFO bus 0
	11.	RSC_1	output	segment 2	- byte 2 (bits(15:08))	output FIFO bus 1
	12.	RSC_1	output	segment 3	- byte 3 (bits(07:00))	output FIFO bus 1
20	13.	RSC_1	output	segment 4	- byte 2 (bits(15:08))	output FIFO bus 2
	14.	RSC_1	output	segment 5	- byte 3 (bits(07:00))	output FIFO bus 2
	15.	RSC_1	output	segment 6	- byte 2 (bits(15:08))	output FIFO bus 3
	16.	RSC_1	output	segment 7	- byte 3 (bits(07:00))	output FIFO bus 3

The input would be mapped in a similar manner (not shown in FIG. 5B). The input segments mapped as;

25	1.	RSC_0	input	segment 0	- byte 0 (bits(31:24))	input FIFO bus 0
	2.	RSC_0	input	segment 1	- byte 1 (bits(23:16))	input FIFO bus 0
	3.	RSC_0	input	segment 2	- byte 0 (bits(31:24))	input FIFO bus 1
	4.	RSC_0	input	segment 3	- byte 1 (bits(23:16))	input FIFO bus 1
30	5.	RSC_0	input	segment 4	- byte 0 (bits(31:24))	input FIFO bus 2
	6.	RSC_0	input	segment 5	- byte 1 (bits(23:16))	input FIFO bus 2
	7.	RSC_0	input	segment 6	- byte 0 (bits(31:24))	input FIFO bus 3
	8.	RSC_0	input	segment 7	- byte 1 (bits(23:16))	input FIFO bus 3
35	9.	RSC_1	input	segment 0	- byte 2 (bits(15:08))	input FIFO bus 0
	10.	RSC_1	input	segment 1	- byte 3 (bits(07:00))	input FIFO bus 0
	11.	RSC_1	input	segment 2	- byte 2 (bits(15:08))	input FIFO bus 1
	12.	RSC_1	input	segment 3	- byte 3 (bits(07:00))	input FIFO bus 1
40	13.	RSC_1	input	segment 4	- byte 2 (bits(15:08))	input FIFO bus 2
	14.	RSC_1	input	segment 5	- byte 3 (bits(07:00))	input FIFO bus 2
	15.	RSC_1	input	segment 6	- byte 2 (bits(15:08))	input FIFO bus 3
	16.	RSC_1	input	segment 7	- byte 3 (bits(07:00))	input FIFO bus 3

where in FIG. 5B, FBUS_xY corresponds to byte "x" of FBUS "Y." Thus output segments (RSC_0 0,1/ RSC_1 0,1) , (RSC_0 2,3/ RSC_1 2,3) , (RSC_0 4,5/ RSC_1 4,5) , (RSC_0 6,7/ RSC_1 6,7) representing output FIFO busses 0, 1, 2, and 3 respectively, connect concurrently to corresponding Virtual

Input Queues in VIL 42 for the input segments of input FIFO busses 0, 1, 2, and 3.

Thus, an 8 x 8 FIFO bus crossbar requires 4 RSC chips, with the output mapped as:

5

1. RSC_0 output segment 0 - byte 0 (bits(31:24)) output FIFO bus 0
2. RSC_0 output segment 1 - byte 0 (bits(31:24)) output FIFO bus 1
3. RSC_0 output segment 2 - byte 0 (bits(31:24)) output FIFO bus 2
4. RSC_0 output segment 3 - byte 0 (bits(31:24)) output FIFO bus 3

10

5. RSC_0 output segment 4 - byte 0 (bits(31:24)) output FIFO bus 4
6. RSC_0 output segment 5 - byte 0 (bits(31:24)) output FIFO bus 5
7. RSC_0 output segment 6 - byte 0 (bits(31:24)) output FIFO bus 6
8. RSC_0 output segment 7 - byte 0 (bits(31:24)) output FIFO bus 7

15

9. RSC_1 output segment 0 - byte 1 (bits(23:16)) output FIFO bus 0
10. RSC_1 output segment 1 - byte 1 (bits(23:16)) output FIFO bus 1
11. RSC_1 output segment 2 - byte 1 (bits(32:16)) output FIFO bus 2
12. RSC_1 output segment 3 - byte 1 (bits(23:16)) output FIFO bus 3

20

13. RSC_1 output segment 4 - byte 1 (bits(23:16)) output FIFO bus 4
14. RSC_1 output segment 5 - byte 1 (bits(23:16)) output FIFO bus 5
15. RSC_1 output segment 6 - byte 1 (bits(23:16)) output FIFO bus 6
16. RSC_1 output segment 7 - byte 1 (bits(23:16)) output FIFO bus 7

25

17. RSC_2 output segment 0 - byte 2 (bits(15:08)) output FIFO bus 0
18. RSC_2 output segment 1 - byte 2 (bits(15:08)) output FIFO bus 1
18. RSC_2 output segment 2 - byte 2 (bits(15:08)) output FIFO bus 2
19. RSC_2 output segment 3 - byte 2 (bits(15:08)) output FIFO bus 3

30

20. RSC_2 output segment 4 - byte 2 (bits(15:08)) output FIFO bus 4
21. RSC_2 output segment 5 - byte 2 (bits(15:08)) output FIFO bus 5
22. RSC_2 output segment 6 - byte 2 (bits(15:08)) output FIFO bus 6
23. RSC_2 output segment 7 - byte 2 (bits(15:08)) output FIFO bus 7

35

24. RSC_3 output segment 0 - byte 3 (bits(07:00)) output FIFO bus 0
25. RSC_3 output segment 1 - byte 3 (bits(07:00)) output FIFO bus 1
26. RSC_3 output segment 2 - byte 3 (bits(07:00)) output FIFO bus 2
27. RSC_3 output segment 3 - byte 3 (bits(07:00)) output FIFO bus 3

40

28. RSC_3 output segment 4 - byte 3 (bits(07:00)) output FIFO bus 4
29. RSC_3 output segment 5 - byte 3 (bits(07:00)) output FIFO bus 5
30. RSC_3 output segment 6 - byte 3 (bits(07:00)) output FIFO bus 6
31. RSC_3 output segment 7 - byte 3 (bits(07:00)) output FIFO bus 7

45

The input segments would be mapped as follows:

5	1.	RSC_0	input	segment 0	- byte 0 (bits(31:24))	input FIFO bus 0
	2.	RSC_0	input	segment 1	- byte 0 (bits(31:24))	input FIFO bus 1
	3.	RSC_0	input	segment 2	- byte 0 (bits(31:24))	input FIFO bus 2
	4.	RSC_0	input	segment 3	- byte 0 (bits(31:24))	input FIFO bus 3
10	5.	RSC_0	input	segment 4	- byte 0 (bits(31:24))	input FIFO bus 4
	6.	RSC_0	input	segment 5	- byte 0 (bits(31:24))	input FIFO bus 5
	7.	RSC_0	input	segment 6	- byte 0 (bits(31:24))	input FIFO bus 6
	8.	RSC_0	input	segment 7	- byte 0 (bits(31:24))	input FIFO bus 7
15	9.	RSC_1	input	segment 0	- byte 1 (bits(23:16))	input FIFO bus 0
	10.	RSC_1	input	segment 1	- byte 1 (bits(23:16))	input FIFO bus 1
	11.	RSC_1	input	segment 2	- byte 1 (bits(32:16))	input FIFO bus 2
	12.	RSC_1	input	segment 3	- byte 1 (bits(23:16))	input FIFO bus 3
20	13.	RSC_1	input	segment 4	- byte 1 (bits(23:16))	input FIFO bus 4
	14.	RSC_1	input	segment 5	- byte 1 (bits(23:16))	input FIFO bus 5
	15.	RSC_1	input	segment 6	- byte 1 (bits(23:16))	input FIFO bus 6
	16.	RSC_1	input	segment 7	- byte 1 (bits(23:16))	input FIFO bus 7
25	17.	RSC_2	input	segment 0	- byte 2 (bits(15:08))	input FIFO bus 0
	18.	RSC_2	input	segment 1	- byte 2 (bits(15:08))	input FIFO bus 1
	18.	RSC_2	input	segment 2	- byte 2 (bits(15:08))	input FIFO bus 2
	19.	RSC_2	input	segment 3	- byte 2 (bits(15:08))	input FIFO bus 3
30	20.	RSC_2	input	segment 4	- byte 2 (bits(15:08))	input FIFO bus 4
	21.	RSC_2	input	segment 5	- byte 2 (bits(15:08))	input FIFO bus 5
	22.	RSC_2	input	segment 6	- byte 2 (bits(15:08))	input FIFO bus 6
	23.	RSC_2	input	segment 7	- byte 2 (bits(15:08))	input FIFO bus 7
35	24.	RSC_3	input	segment 0	- byte 3 (bits(07:00))	input FIFO bus 0
	25.	RSC_3	input	segment 1	- byte 3 (bits(07:00))	input FIFO bus 1
	26.	RSC_3	input	segment 2	- byte 3 (bits(07:00))	input FIFO bus 2
	27.	RSC_3	input	segment 3	- byte 3 (bits(07:00))	input FIFO bus 3
40	28.	RSC_3	input	segment 4	- byte 3 (bits(07:00))	input FIFO bus 4
	29.	RSC_3	input	segment 5	- byte 3 (bits(07:00))	input FIFO bus 5
	30.	RSC_3	input	segment 6	- byte 3 (bits(07:00))	input FIFO bus 6
	31.	RSC_3	input	segment 7	- byte 3 (bits(07:00))	input FIFO bus 7

For the 8 x 8 FIFO crossbar configuration each input/output segment of the RSC 12 switches 1 byte of the 32 bit FIFO bus concurrent with the other RSC 12 slices.

Referring to FIG. 6, distribution of the Output Segment Arbitration Logic (OSA) 50a-50h is shown. Each of

the OSA logic elements 50 determine which virtual queue 42 to link to which of the output segment logic 44a-44h. The input FBUS segments at a start of a new packet provides a start of packet "SOP" flag, the destination port of this new packet and a virtual queue number. At this starting point the destination port is known, so a physical mapping to the output segment logic is performed. This mapping is stored in an output port Map Queue 60. Each physical port has an output port Map Queue 60₀-60₁₂₇. These queues maintain pointers to the next virtual queue which has a packet for the port. Each Map Queue 60 maintains up to 8 entries (one for each input segment). The entry has the VIQ# of the next packet to be transmitted.

When the Output Segment Logic 44 completes the transmission of an packet to a particular port, the Output Segment Logic requests the Output Arbitration Logic 50 to supply the Output Segment Logic 44 with the NEXT_VQ number that is stored in that output ports' map queue 60 NEXT_VQ. This NEXT_VQ number is an address that the Output Segment Logic 44 uses to control 8:1 multiplexers 62a-62h that feed the Output Segment Logic's 44 16 mpkt entry output queues. The VIQ number is a 7 bit value. The three most significant bits indicate which Input Segment and the least significant 4 bits indicate which VIQ within that input segment. Using these 7 bits the OSL can completely specify the next packet to be transmitted.

Disposed between the Input Segment Logic (ISL) 40 and the output port Map Queue 60 is a time division multiplex bus 70 used by the eight input segments to transmit destination information to the selected output segments.

76307-71

This multiplexing is straightforward since there are at a maximum 8 new packets to be sorted and 16 cycles to promote their VIQ# to their respective Map Queue. Therefore, by employing a very simple round robin technique the input
5 segment Destination ports are sorted over the next 8 cycles. The Map Queues drain using a first in first out algorithm.

Other Embodiments

It is to be understood that while the invention has been described in conjunction with the detailed
10 description thereof, the foregoing description is intended to illustrate and not limit the scope of the invention, which is defined by the scope of the appended claims.

76307-71

CLAIMS:

1. A switch fabric, comprising:

first and second data switches each having a plurality of input ports and a plurality of output ports,
5 the first and second data switches capable of switching any of its input ports to any of its output ports, each of the first and second data switches comprising:

a plurality of virtual queues;

input segment logic coupled to at least one bus,
10 said input segment logic to determine to which virtual queue incoming data should be sent;

output segment logic to select which virtual queue should be connected to an output port; and

a distributed switching arrangement to provide a
15 non-blocking cross-bar switching fabric arrangement capability over a series of byte sliced buses, the distributed switching arrangement coupling the virtual queues to the output segment logic through the non-blocking cross-bar switching arrangement, the non-blocking cross-bar
20 switching fabric arrangement comprising a plurality of multiplexers coupled to outputs of the plurality of the virtual queues to select ones of the virtual queues to feed to a second plurality of multiplexers that produce inputs to the output segment logic;

25 said first and second data switches being coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to one of the input ports of the first switch, and a last byte of the first

76307-71

input bus is coupled to one of the input ports of the second switch.

2. The switch fabric of claim 1 wherein said output ports of the first and second data switches are coupled to a plurality of output buses so that a first byte of a first one of the output buses is coupled to one of the output ports of the first switch and a last byte of the first output bus is coupled to one of the output ports of the second switch.

10 3. The switch fabric of claim 1 wherein the first and second data switches comprise:

logic to control mapping of bytes of the input buses to the input segment logic and mapping of bytes of the output segment logic to the output buses.

15 4. A switch fabric, comprising:

a plurality of data switches each having a plurality of input ports and a plurality of output ports, the plurality of switches capable of switching any of its input ports to any of its output ports, the plurality of data switches, comprising:

a plurality of virtual queues;

input segment logic coupled to at least one bus, said input segment logic to select to which virtual queue incoming data should be sent;

25 output segment logic to select which virtual queue should be connected to an output port; and

a distributed switching arrangement to provide a non-blocking cross-bar switching fabric arrangement

·76307-71

capability over a series of byte sliced buses, the distributed switching arrangement coupling the virtual queues to the output segment logic through the non-blocking cross-bar switching arrangement, the non-blocking cross-bar
5 switching fabric arrangement comprising a plurality of multiplexers coupled to outputs of the plurality of the virtual queues to select ones of the virtual queues to feed to a second plurality of multiplexers that produce inputs to the output segment logic;

10 said input ports of plurality of data switches being coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to one of the input ports of a first one of the plurality of switches, and a succeeding byte of the first input bus is coupled to
15 one of the input ports of a succeeding one of the plurality of switches.

5. The switch fabric of claim 4 wherein the plurality of switches are two and the succeeding byte is a third byte of a four byte bus.

20 6. The switch fabric of claim 5 wherein the switches couple four byte sliced buses, and with the first byte and the second byte of each of the four buses being coupled to respective input ports of the first switch and the third
byte and fourth byte of each of the four buses being coupled
25 to respective input ports of the second switch.

7. The switch fabric of claim 4 wherein the plurality of switches are four and the succeeding byte is a second byte of a four byte bus.

8. The switch fabric of claim 7 wherein the switches
30 couple to eight byte sliced buses, and with the first byte

76307-71

of each of the eight buses being coupled to respective input ports of a first one of the switches, a second byte of each of the eight buses being coupled to respective input ports of a second one of the switches, a third byte of each of the eight buses being coupled to respective input ports of a third one of the switches, and a fourth byte of each of the eight buses being coupled to respective input ports of a fourth one of the switches.

9. A switch fabric comprising:

10 a first data switch having a plurality of inputs and outputs;

a second data switch having a plurality of inputs and outputs, each data switch, comprises:

a plurality of virtual queues;

15 input segment logic coupled to at least one bus, said input segment logic to select to which virtual queue incoming data should be sent;

output segment logic to select which virtual queue should be connected to an output port; and

20 a distributed switching arrangement to provide a non-blocking cross-bar switching fabric arrangement capability over a series of byte sliced buses, the distributed switching arrangement coupling the virtual queues to the output segment logic through the non-blocking cross-bar switching arrangement, the non-blocking cross-bar switching fabric arrangement comprising a plurality of multiplexers coupled to outputs of the plurality of the virtual queues to select ones of the

25

· 76307-71

virtual queues to feed to a second plurality of multiplexers that produce inputs to the output segment logic,

wherein the distributed switching arrangement has the inputs of the first and second data switches coupled to a plurality of input buses so that a first byte of a first one of the input buses is coupled to one of the inputs of the first switch, and a last byte of the first input bus is coupled to one of the outputs of the second switch.

10. The switching fabric of claim 9 wherein the distributed switching arrangement has the outputs of the first and second data switches coupled to a plurality of output buses so that a first byte of a first one of the output buses is coupled to one of the outputs of the first switch, and a last byte of the first output bus is coupled to one of the outputs of the second switch.

SMART & BIGGAR

OTTAWA, CANADA

PATENT AGENTS

PCT/US 00/22368
NOV 13 NOV 2000

CA 02383473 2002-02-27

1/8

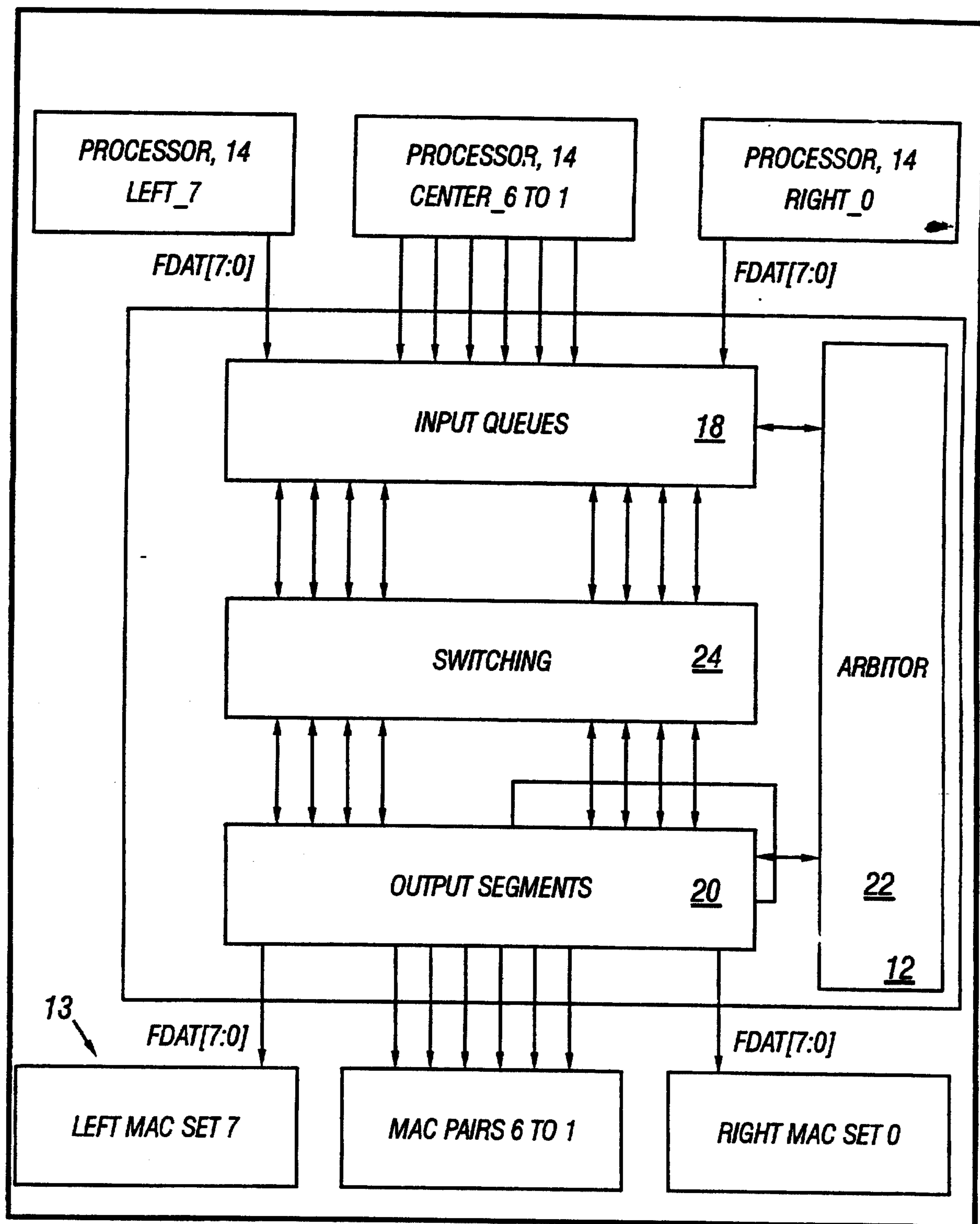


FIG. 1

PCT/US 00/22368
 RO/US 13 NOV 2000

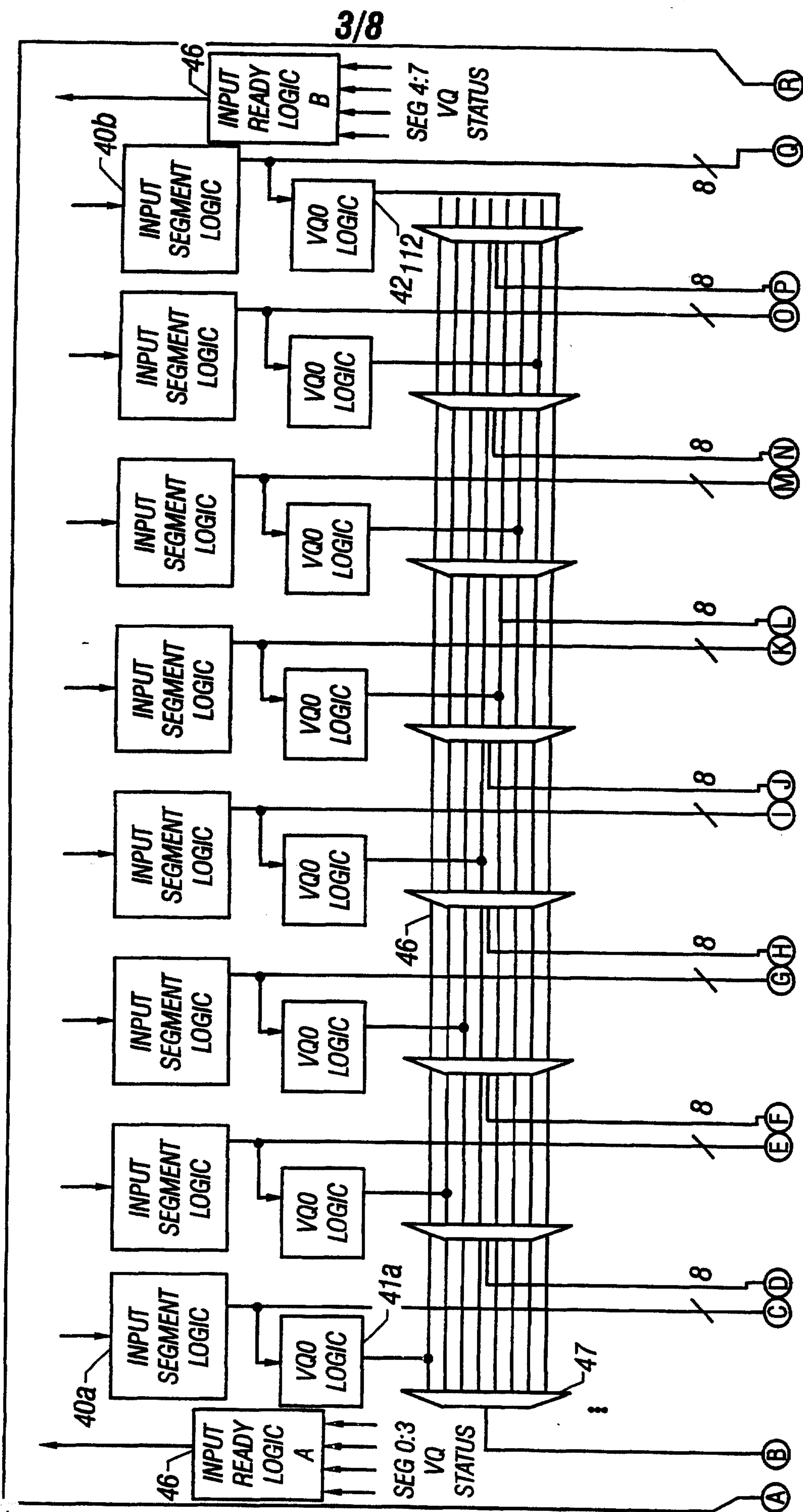


FIG. 3A

4/8

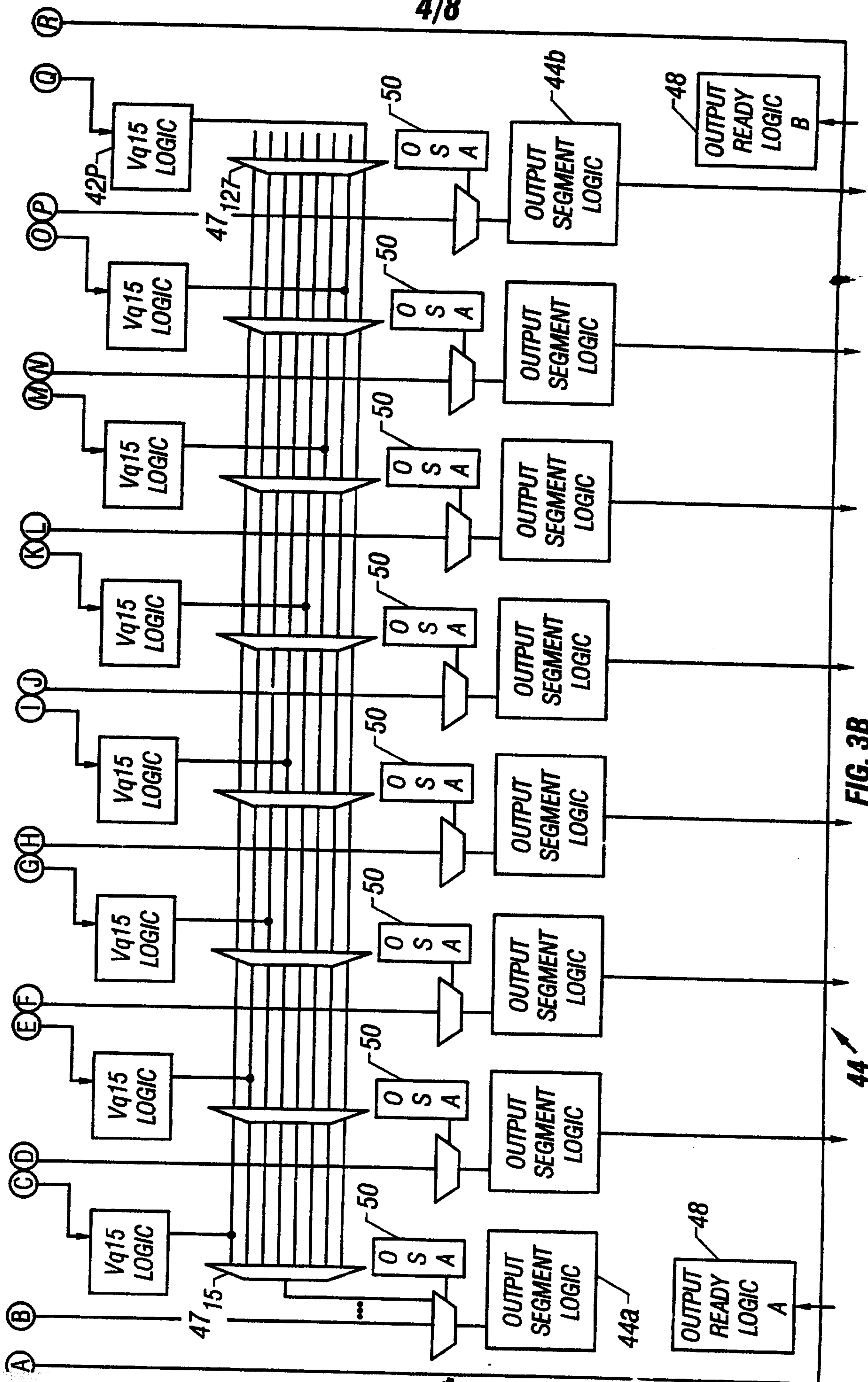


FIG. 3B

CA 02383473 2002-02-27

PCT/US 00/22368
RO/US 13 NOV 2000

5/8

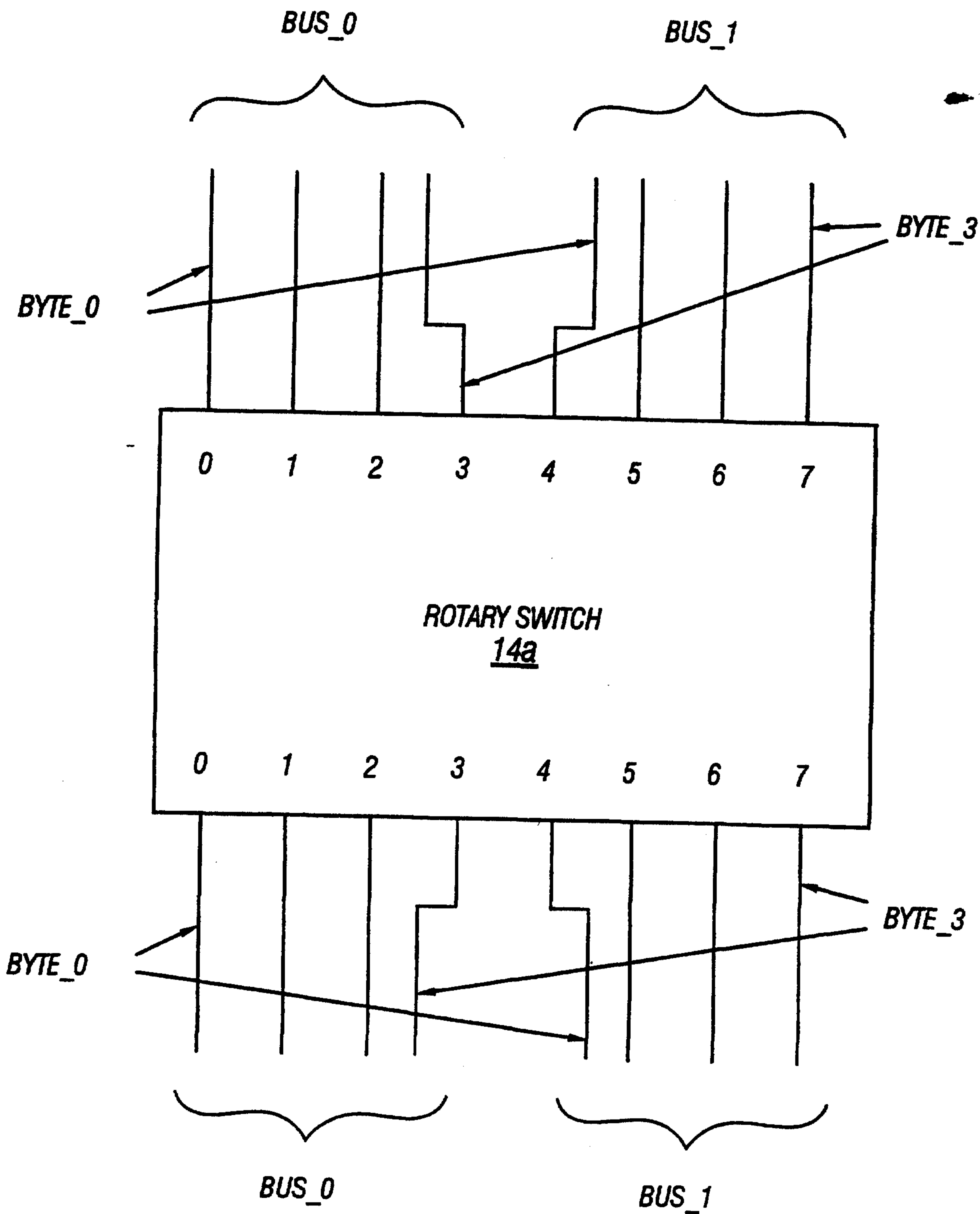


FIG. 4A

PCT/US 00/22368
RO/US 13 NOV 2000

6/8

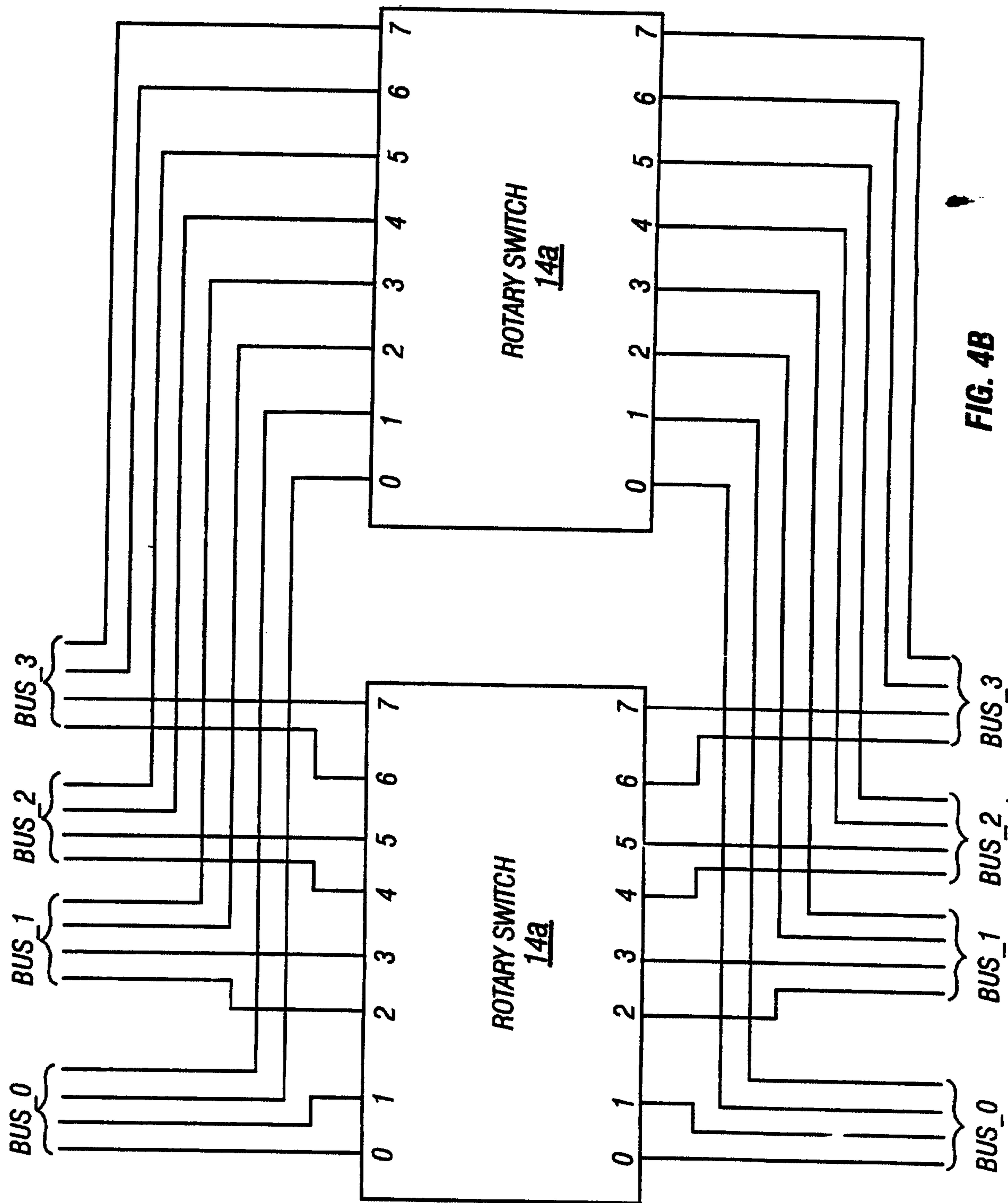


FIG. 4B

7/8

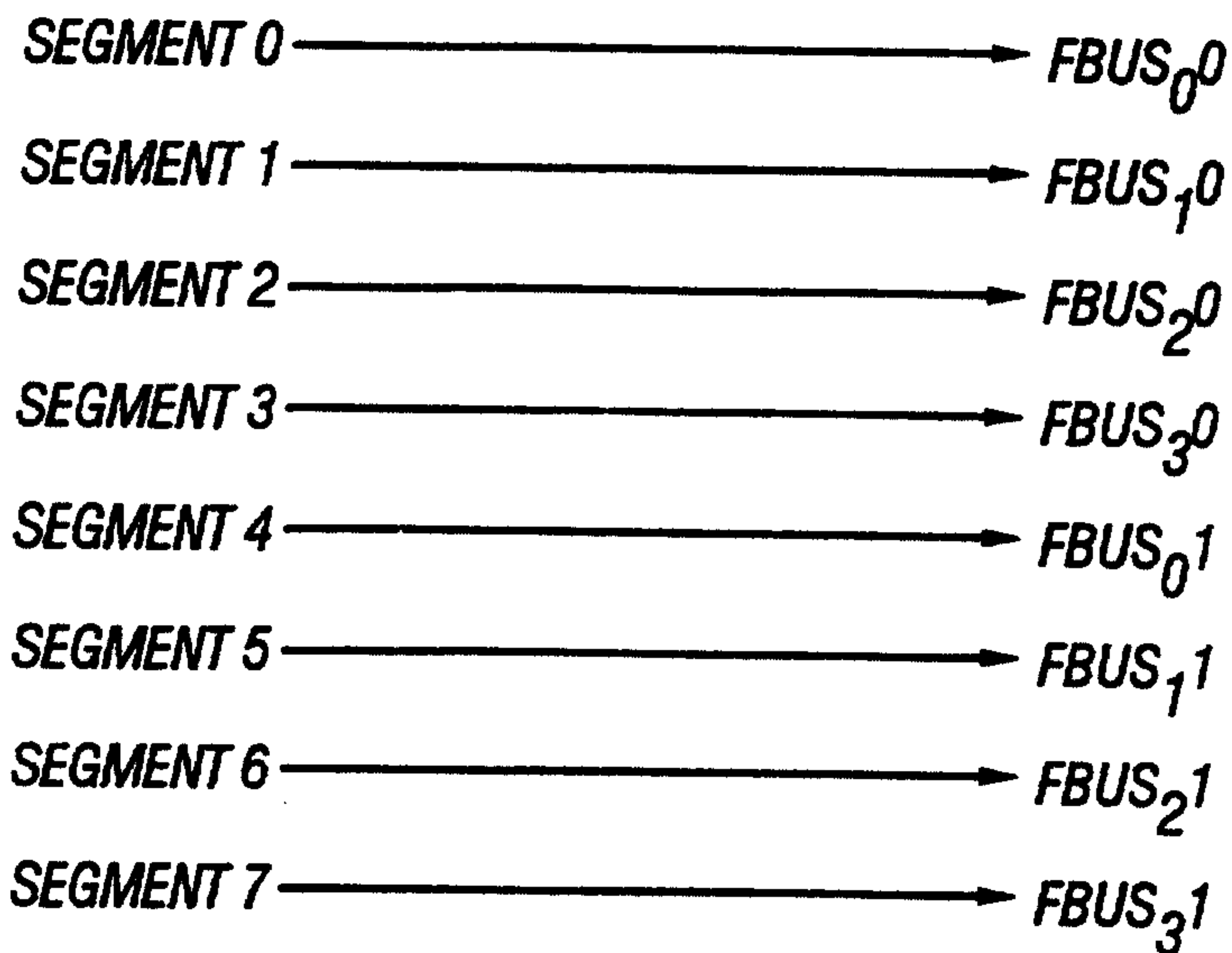


FIG. 5A

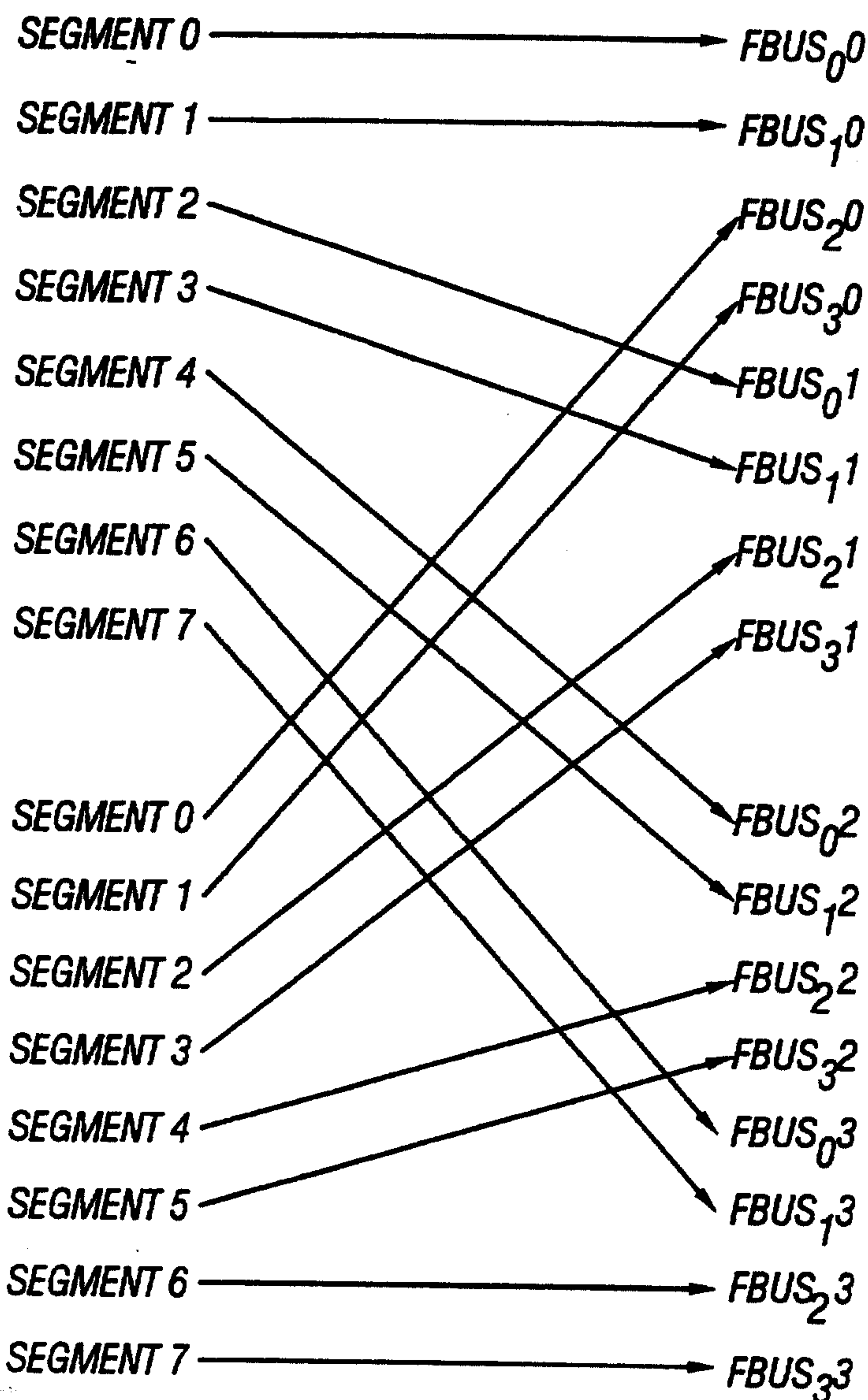


FIG. 5B

ISA-DRAW
PCT/US 00/22368
R0/US 13 NOV 2000

8/8

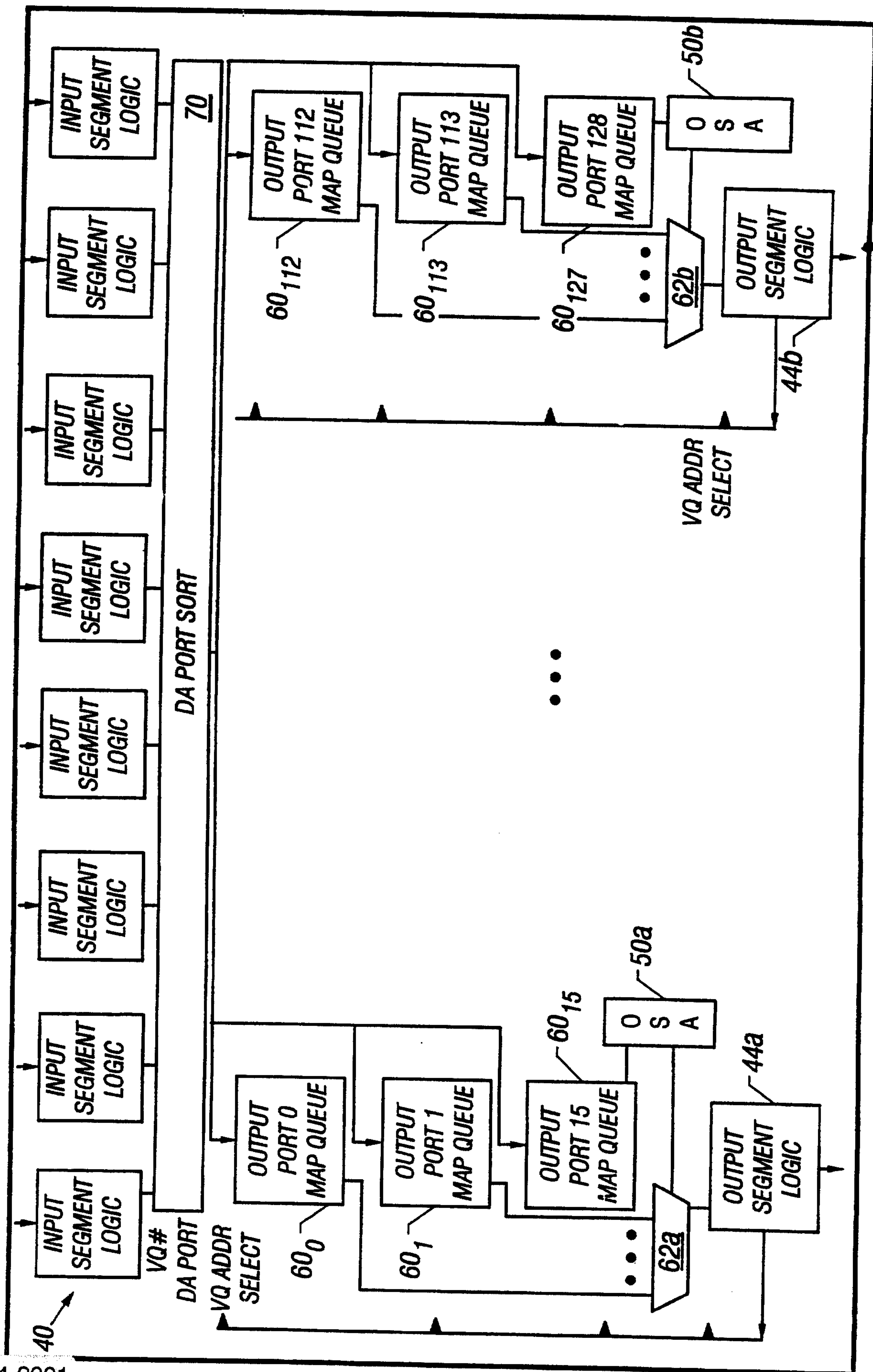


FIG. 6

