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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
14 November 2013 (14.11.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/169948 A1

(51) International Patent Classification:
G06F 9/44 (2006.01) *H04L 29/12* (2006.01)

(21) International Application Number:
PCT/US2013/040210

(22) International Filing Date:
8 May 2013 (08.05.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

61/646,107	11 May 2012 (11.05.2012)	US
13/889,088	7 May 2013 (07.05.2013)	US
13/889,123	7 May 2013 (07.05.2013)	US

(71) Applicant: **ORACLE INTERNATIONAL CORPORATION** [US/US]; 500 Oracle Parkway, M/S 5op7, Redwood Shores, CA 94065 (US).

(72) Inventors: **BOGDANSKI, Bartosz**; Hoff Terrasse 15, H0203, N-0275 Oslo (NO). **JOHNSEN, Bjorn, Dag**; Vilberggrenda 9, N-0687 Oslo (NO).

(74) Agents: **MEYER, Sheldon, R.** et al.; Fliesler Meyer LLP, 650 California Street, Fourteenth Floor, San Francisco, CA 94108 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

(54) **Title**: SYSTEM AND METHOD FOR ROUTING TRAFFIC BETWEEN DISTINCT INFINIBAND SUBNETS BASED ON SOURCE ROUTING

400

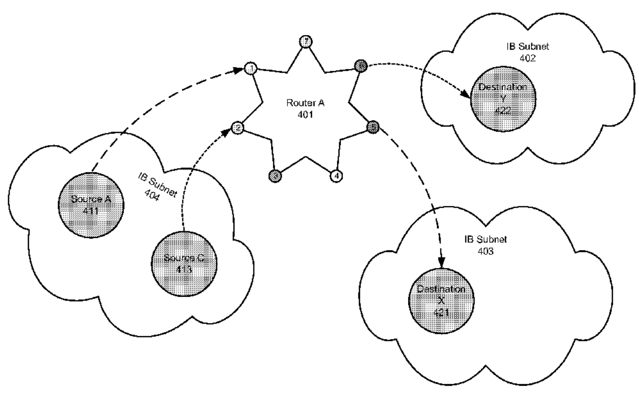


FIGURE 4

(57) **Abstract**: A system and method can rout traffic between distinct subnets in a network environment. A router that connects the distinct subnets, such as InfiniBand (IB) subnets, can receive a list of destinations that the router is responsible for routing one or more packets to. Then, the router can generate a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets, and use a modulo based hash to select one router port from a plurality of output router ports of the router.

WO 2013/169948 A1

SYSTEM AND METHOD FOR ROUTING TRAFFIC BETWEEN DISTINCT INFINIBAND SUBNETS BASED ON SOURCE ROUTING

Copyright Notice:

Field of Invention:

[0002] The present invention is generally related to computer systems, and is particularly related to a middleware machine environment.

Background:

[0003] As larger cloud computing architectures are introduced, the performance and administrative bottlenecks associated with the traditional network and storage have become a significant problem. The InfiniBand (IB) technology has seen increased deployment as the foundation for a cloud computing fabric. This is the general area that embodiments of the invention are intended to address.

Summary:

[0004] Described herein is a system and method that can rout traffic between distinct InfiniBand (IB) subnets. A router that connects the distinct subnets, such as InfiniBand (IB) subnets, can receive a list of destinations that the router is responsible for routing one or more packets to. Then, the router can generate a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets, and use a modulo based hash to select one router port from a plurality of output router ports of the router.

[0005] Also described herein is a system for routing traffic between distinct subnets in a network environment. The system comprises means for receiving, at a router, a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment. The system further comprises means for generating a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets. The system further comprises means for using a modulo based hash to select one router port from a plurality of output router ports of the router.

[0006] Further, also described herein is a system for routing traffic between distinct subnets in a network environment. The system comprises a router configured to receive a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at

least one subnet in the network environment. The system is also configured to generate a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets and to use a modulo based hash to select one port from the output router ports of the router.

Brief Description of the Figures:

[0007] **Figure 1** shows an illustration of routing traffic between distinct InfiniBand (IB) subnets in a network environment, in accordance with an embodiment of the invention.

[0008] **Figure 2** shows an illustration of supporting packet forwarding on a router in a network environment, in accordance with an embodiment of the invention.

[0009] **Figure 3** shows an illustration of choosing ingress router ports for different destinations in a network environment, in accordance with an embodiment of the invention.

[0010] **Figure 4** shows an illustration of selecting an egress router port for a received packet at a router in a network environment, in accordance with an embodiment of the invention.

[0011] **Figure 5** illustrates an exemplary flow chart for supporting the inter-subnet source routing (ISSR) algorithm at a router in a network environment, in accordance with an embodiment of the invention.

[0012] **Figure 6** shows an illustration of supporting the inter-subnet fat-tree routing (ISFR) algorithm at a router in a network environment, in accordance with an embodiment of the invention.

[0013] **Figures 7-9** illustrates routing in a network environment with different topologies, in accordance with an embodiment of the invention.

[0014] **Figure 10** illustrates an exemplary flow chart for supporting the inter-subnet fat-tree routing (ISFR) algorithm at a router in a network environment, in accordance with an embodiment of the invention.

[0015] **Figure 11** illustrates a functional block diagram to show features in accordance with an embodiment of the invention.

Detailed Description:

[0016] The invention is illustrated, by way of example and not by way of limitation, in the figures of the accompanying drawings in which like references indicate similar elements. It should be noted that references to “an” or “one” or “some” embodiment(s) in this disclosure are not necessarily to the same embodiment, and such references mean at least one.

[0017] The description of the invention as following uses the InfiniBand (IB) network as an example for a high performance network. It will be apparent to those skilled in the art that other types of high performance networks can be used without limitation. Also, the description of the invention as following uses the fat-tree as an example for a network topology model. It will be apparent to those skilled in the art that other types of network topology models can be used without limitation.

[0018] Described herein are systems and methods that can support routing traffic in between distinct networks.

InfiniBand (IB) Architecture

[0019] The IB Architecture is a serial point-to-point full-duplex technology. As IB clusters grow in size and complexity, the network can be segmented into manageable sections, which can be referred to as subnets. An IB subnet can include a set of hosts interconnected using switches and point to point links. An IB subnet can also include at least one subnet manager (SM), which is responsible for initializing and bringing up the network, including the configuration of all the switches, routers and host channel adapters (HCAs) in the subnet.

[0020] IB supports a rich set of transport services in order to provide both remote direct memory access (RDMA) and traditional send/receive semantics. Independent of the transport service used, the IB HCAs communicate using queue pairs (QPs). A QP is created during the communication setup, and can have a set of initial attributes such as QP number, HCA port, destination LID, queue sizes, and transport service that are supplied. An HCA can handle many QPs, each QP consists of a pair of queues, such as a send queue (SQ) and a receive queue (RQ), and there is one such pair present at each end-node participating in the communication. The send queue holds work requests to be transferred to the remote node, while the receive queue holds information on what to do with the data received from the remote node. In addition to the QPs, each HCA has one or more completion queues (CQs) that are associated with a set of send and receive queues. The CQ holds completion notifications for the work requests posted to the send and receive queue. Even though the complexities of the communication are hidden from the user, the QP state information is kept in the HCA.

Routing Traffic between Distinct InfiniBand (IB) Subnets

[0021] **Figure 1** shows an illustration of routing traffic between distinct InfiniBand (IB) subnets in a network environment, in accordance with an embodiment of the invention. As shown in Figure 1, a network environment 100 can be based on the InfiniBand Architecture (IBA), and supports a two-layer topological division.

[0022] The lower layer of the IB network, or IB fabric 100, can be referred to as subnets, e.g. subnets 101-104, each of which includes a set of hosts interconnected using switches and point-to-point links. At the higher level, the one or more subnets 101-104 in an IB fabric 100 can be interconnected using routers, e.g. routers 105-106. Furthermore, each subnet 101-104 can run its own subnet manager (SM) that configures only the ports on the local subnet and the routers 105-106 are non-transparent to the subnet managers (SM).

[0023] The hosts and switches within each of the subnets101-104 can be addressed using the designated local identifiers (LIDs). The size of the large installations may be limited by the number of available local identifiers (LIDs). For example, a single subnet may be limited to 49151 unicast LIDs. One approach to expand the IB address space is expanding the LID addressing space to 32 bits, the usability of which can be limited since this approach may not be backward compatible with older hardware.

[0024] In accordance with an embodiment of the invention, the IB routers 105-106 can provide address space scalability in the IB fabric 100. As shown in Figure 1, when more end-ports are desired, multiple subnets 101-104 can be combined together using one or more IB routers 105-106. Each of the subnets 101-104 can use a local identifier (LID) address space 111-114. Since LID addresses have local visibility within each LID address space 111-114, the LID addresses can be reused in the different subnets101-104 connected by routers 105-106. Thus, address space scalability can be provided for the IB fabric 100, and this approach can theoretically yield an unlimited addressing space.

[0025] Furthermore, by segmenting a large and complex network 100 into multiple subnets 101-104, the system can provide fabric management containment. The fabric management containment can be beneficial in providing: 1) fault isolation, 2) increased security, and 3) intra-subnet routing flexibility.

[0026] First, by dividing a large network 100 into several smaller subnets 101-104, faults or topology changes can be contained to a single subnet, and the subnet reconfiguration may not pass through the routers 105-106 to other subnets. This shortens the reconfiguration time and limits the impact of a fault.

[0027] Second, from a security point of view, segmenting a large fabric100 into subnets 101-104 using routers 105-106 limits the scope of most attacks to the attacked subnet.

[0028] Third, from a routing point of view, fabric management containment can be beneficial in supporting more flexible routing schemes, which can be particularly advantageous in case of a hybrid fabric that comprises two or more regular topologies.

[0029] For example, a hybrid fabric 100 may include a fat-tree part interconnected with a mesh or a torus part (or any other regular topology). There may be no straightforward way to route the different parts of the hybrid fabric 100 separately, because the intra-subnet routing algorithms can only have a subnet scope. Moreover, there are no general purpose agnostic routing algorithms for IB that can provide optimal performance for a hybrid topology.

[0030] In accordance with an embodiment of the invention, the hybrid topology can be divided into smaller regular subnets101-104, each of which can be routed using a different routing algorithm that is optimized for a particular subnet. For example, a fat-tree routing algorithm can route the fat-tree part and the dimension-order routing can route the mesh part of the topology.

[0031] Furthermore, a super subnet manager can be used to coordinate between the local subnet managers and can establish the path through the transit subnet, in the case of more irregular networks where the final destination is located behind another subnet (e.g. at least two router hops required).

[0032] Thus, by using routing between IB subnets, the system can provide address space scalability and fabric management containment in an IB network.

Native InfiniBand(IB) Routers

[0033] **Figure 2** shows an illustration of supporting packet forwarding on a router in a network environment, in accordance with an embodiment of the invention. As shown in Figure 2, the network environment 200 can include multiple subnets, e.g. IB subnets A-B 201-202, that are interconnected using a router 210. Furthermore, the IB subnet A 201 connects to an end node X 203, which is associated with host channel adapter (HCA) 213, and the IB subnet B 202 connects to an end node Y 204, which is associated with host channel adapter (HCA) 214.

[0034] In accordance with an embodiment of the invention, the IB router 210 can operate at layer-3 of the IB addressing hierarchy. Furthermore, in addition to LIDs, each IB device may also have a 64-bit global unique identifier (GUID), which can be used to form a GID, an IB layer-3 address. For example, a GID can be created by concatenating a 64-bit subnet ID with the 64-bit GUID to form an IPv6-like 128-bit address. Additionally, the term GUID may be also used to refer to a port GUIDs, i.e. the GUIDs assigned to every port in the IB fabric, and the GUID can be burned into the non-volatile memory.

[0035] As shown in Figure 2, an end-node, e.g. end node X 203, can send a packet to another end-node, e.g. end node Y 204, via the router 210. The address resolution mechanism can make the local router 210 visible to the end nodes X-Y 203-204.

[0036] For example, the packet received at the router 210 can include an incoming routing header 221, which includes a local routing header (LRH) 223 and a global routing header (GRH) 225. Before forwarding the packet, the router 210 can modify the incoming routing header 221 into an outgoing routing header 222, which includes a LRH 224 and a GRH 226.

[0037] As shown in Figure 2, the end-node X 203 can put the local HCA's LID address, e.g. X, as the source LID, srcLID, and put the local router's LID address, e.g. A, as the destination LID, dstLID, in the LRH 223. Furthermore, the end-node X 203 can put its layer-3 address (GID), e.g. 1234, as the source GID, srcGID, and can put the final destination address (GID), e.g. 5678, as the destination GID, dstGID, in the GRH 225.

[0038] When the packet reaches at the router 210, a packet forwarding mechanism 220 can update and replace the packet fields. For example, the system can replace the source LID, srcLID, in the local routing header (LRH) 224 with the LID of the router's egress port, e.g. B. Furthermore, the system can replace the destination LID, dstLID, with the LID address of the destination, e.g. Y.

[0039] Alternatively, the system can use the LID of the egress port of the previous-hop router as the source LID, srcLID, if the packet is forwarded in from another router. Also, the system can replace the destination LID, dstLID, with the LID of the next-hop port, if further packet forwarding is necessary. In each case, the system can recompute the cyclic redundancy checks (CRCs) before forwarding the packet to the next hop.

[0040] In accordance with an embodiment of the invention, different methods can be used for routing traffic between distinct InfiniBand (IB) subnets. These methods can be used to answer various problems with inter-subnet routing, such as which router should be chosen for a particular destination (first routing phase) and which path should be chosen by the router to reach the destination (second routing phase).

[0041] For example, these methods can include a simple classic routing method, a source routing method that provides good performance for various topologies, and an optimized routing method that is specialized and provides optimal performance for fat-tree based topologies. Both the source routing method and the optimized routing method can potentially deliver better performance than the classic routing method. Furthermore, the optimized routing method allows for obtaining the optimal performance for an underlying fat-tree topology. Additionally, both methods support the use of arbitrary multi-port routers, whereas the classic routing may not utilize all the ports effectively, even when it does not restrict the number of available ports.

[0042] Thus, using these methods, the native IB routers allow building more complex IB fabrics by connecting subnets together without a significant decrease in performance.

Inter-Subnet Source Routing (ISSR)

[0043] In accordance with an embodiment of the invention, a general purpose routing algorithm, such as the inter-subnet source routing (ISSR) routing algorithm, can be used for routing complex network with hybrid subnets. The ISSR routing algorithm can include two phases, a first phase for choosing an ingress router port for a particular destination, and a second phase for selecting an egress router port.

[0044] **Figure 3** shows an illustration of choosing ingress router ports for different destinations in a network environment, in accordance with an embodiment of the invention. As shown in Figure 3, a network environment 300 can include a subnet 310 that is connected with one or more routers, e.g. router A 301 and router B 302.

[0045] A subnet manager (SM), e.g. SM 303 in the subnet 310, can include a mapping file 304, which can be used to choose ingress router ports for different destinations. Furthermore, the choosing of an ingress router port can be based on round-robin path distribution among all available routers that are connected to the local subnet. For example, the find_router() function, which chooses the local router port for the ISSR algorithm, can be implemented in a way similar to the OpenSM routing algorithm.

[0046] The following is a high-level example of a mapping file 304.

Table 1: A high-level example of a mapping file for ISSR and ISFR algorithms

1: dst_gid1	router A port 1 guid
2: dst_gid2	router B port 1 guid
3: dst_gid3	router A port 2 guid
4: dst_gid4	router B port 2 guid
5: #default route	
6: *	router A port 1 guid
7: *	router B port 1 guid

[0047] Unlike OpenSM routing, which can only match a whole subnet to a single router port, the system can map the destination end-ports to different router ports. Furthermore, the setup of the mapping file can be different from the mapping file provided in the OpenSM. As shown in the above Table 1, the mapping file 304 contains a fully qualified port GID, instead of only a subnet prefix as for the OpenSM inter-subnet routing, which allows the system to provide full granularity in choosing ingress router ports.

[0048] Furthermore, as shown in Figure 3, an equal (or similar) number of destinations can be mapped to a number of ports, e.g. in a round robin manner. For example, destination A 311 (dst_gid1) and destination C 313 (dst_gid3) can be routed through port 1 and port 2 on router A 301, and destination B 312 (dst_gid2) and destination D 314 (dst_gid4) can be routed through port 1 and port 2 on router B 302. Additionally, the mapping file 304 can specify backup and default routes.

[0049] In accordance with an embodiment of the invention, the selecting an egress router port can be implemented using a modulo based hash in the router firmware. The hash can take a random number generated using the source and destination LIDs (or any other useful parameter such as the ingress port number on the router). Using the hash, one of the output router ports can be selected. Furthermore, the ISSR routing algorithm is a deterministic oblivious routing algorithm that always uses the same path for the same pair of nodes.

[0050] Furthermore, a two-step port verification method can be used to select the egress router port, since a router may be attached to more than two subnets. First, the system can choose a set of possible ports that are attached to the subnet (or in the direction of the subnet) in which the destination is located. Then, the system can use a simple hash based on a modulo function to choose the egress port.

[0051] **Figure 4** shows an illustration of selecting an egress router port for a received packet at a router in a network environment, in accordance with an embodiment of the invention. As shown in Figure 4, a network environment 400 can include a router, e.g. router A 401, which can route one or more packets to various destinations in different subnets.

[0052] For example, router A 401 can receive, at port 1, a packet from source A 411 which is destined to destination X 421 in the subnet 403. Additionally, router A 401 can receive, at port 2, another packet from source C 413, which is destined to destination Y 422 in the subnet 402.

[0053] The following Algorithm 1 can be implemented both in the SM and the router firmware for selecting an egress router port.

Algorithm 1: choose_egress_port() function in ISSR

```
1: if received_intersubnet_packet() then
2:        dstLID = get_next_LID(dGID)
3:        srand(srcLID + dstLID)
4:        port_set = choose_possible_out_ports()
5:        e_port = port_set[(rand()%por_set:size)]
6: end if
```

[0054] The routing decision can be based both on the source LID and the destination LID. The source LID can be either of the original source or the egress port of the previous-hop router in a transit subnet scenario, while the destination LID can be either the final destination LID or the LID of the next-hop ingress router port.

[0055] The router A 401 knows the values for both the source LID and the destination LID because it can see the subnets attached. To obtain the destination LID, a mapping function, such as the get_next_LID() in Algorithm 1 (line 2), can be used to map the destination GID to a destination LID or returning the next-hop LID based on the subnet prefix located in the GID is required. The mapping function can perform one or more of the following: a content-addressable memory lookup, decoding the final destination local identifier (LID) from the global routing header (GRH) field, decoding the next-hop local identifier from the global routing header (GRH) field.

[0056] In accordance with an embodiment of the invention, the algorithm selecting an egress router port can calculate a random number based on the source and destination LIDs. This random number can be used to select a single egress port from a set of possible ports. The generation of the random number can be done in a deterministic manner so that a given source-destination pair can always generate the same number. Thus, the system can prevent out of order delivery when routing between subnets and, unlike round-robin method for selecting the egress port, makes sure that each source-destination pair always uses the same path through the network.

[0057] **Figure 5** illustrates an exemplary flow chart for supporting the inter-subnet source routing (ISSR) algorithm at a router in a network environment, in accordance with an embodiment of the invention. As shown in Figure 5, at step 501, a router can receive a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment. Then, at step 502, the router can generate a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets. Furthermore, at step 503, the router can use a modulo based hash to select one port from the output router ports of the router.

Inter-Subnet Fat-Tree Routing (ISFR)

[0058] In accordance with an embodiment of the invention, an optimized routing method, such as the inter-subnet fat-tree routing (ISFR) routing algorithm, can be used for routing between subnets with fat-tree topologies.

[0059] The optimized routing method can include three phases, such as a first phase of round-robin path distribution among all available router that are connected to the local subnet, a second phase that queries downward switches connected to each router to learn which switch act as the primary path towards a particular destination, and a third phase that builds a routing table using the information obtained from the first phase and the second phase.

[0060] **Figure 6** shows an illustration of supporting the inter-subnet fat-tree routing (ISFR) algorithm at a router in a network environment 600, in accordance with an embodiment of the invention. As shown in Figure 6, a router 601 is responsible for routing a packet 650 to a subnet 610 that connects to one or more end nodes 621-624. The subnet 610 can include one or more switches, such as SWs 611-616 in a fat-tree topology.

[0061] The router 601 is able to learn which of its ports can be used for routing packets to a particular destination accordingly to the associated GID. For example, the router 601 can receive port mappings 630 from a subnet manager (SM) 620 in a subnet 610. The port mappings 630 can include a list of destination that the router 601 is responsible for routing packets originating from the subnet 610.

[0062] Furthermore, the router 601 can learn from the switches 611-612 which downward output ports to use for achieving maximum performance when routing the packet 650. For example, the router 601 can query switches 611-612 in the subnet 610 for a primary path to a destination node in the subnet 610. The query allows the router to choose the primary path (either switch 611 or switch 612) for the Packet 650, because only one of those switches may have a dedicated primary path towards the desired destination.

[0063] Then, the router 601 can build a routing table 640 using the information obtained above. For example, the local OpenSM routing can mark the primary path to a destination node in the routing table of either switch 611 or switch 612.

[0064] The following Algorithm 2 can be implemented in the router firmware for selecting an egress port on the router.

Algorithm 2: query_down_for_egree_port() function in ISSR

```
1: if received_mapping_files then
2:        for all_port_in_down ports do
3:                down_switch = get_node(port)
4:                lid = get_LID_by_GID(GID)
```

```
            if down_switch:routing_table[lid] == primary_path then
                    e_port = port
            end if
    end for
end
```

[0065] The above ISFR algorithm can use the previously defined file format containing the GID-to-router port mappings in Table 1. Like the ISSR algorithm, the ISFR algorithm can be implemented in the router device. Furthermore, the ISFR algorithm may work only on fat-trees and with fat-tree routing running locally in every subnet. Also, the ISFR algorithm can fall back to ISSR if necessary.

[0066] As shown in Figure 6, the router 601 can be represented as locating on the top of a proper fat-tree topology. Thus, after the query is performed, the router 601 can have one path per port in the downward direction for each destination located in a particular subnet. In the cases of oversubscribed fat-trees, the number of paths can be equal to the oversubscription ratio.

[0067] Additionally, there can be more than one routers connected to the subnet 610. The property of the ISFR routing is such that if all spine (top) routers were replaced with switches, the routing tables for ISFR routing (with routers) and for local routing (with switches) would be the same.

[0068] In accordance with an embodiment of the invention, the ISFR routing algorithm can be implemented based on the communication established between subnet managers (SMs) in different subnets that are connected through the non-transparent routers. For example, an interface can be provided in the routers through which the SMs can communicate, and handshaking can be implemented between two SMs located in neighboring subnets.

[0069] **Figures 7-9** illustrates routing in a network environment with different topologies, in accordance with an embodiment of the invention. Each of the topologies can be represented as a multi-stage fat-tree, e.g. a three-stage fat-tree with routers placed on the top. For example, the three-stage fat-tree can have three routing/switching stages and one node stage, with each subnet (a two-stage fat-tree) appearing as a branch. Furthermore, larger topologies can also be supported based on the ISFR routing algorithm.

[0070] In accordance with an embodiment of the invention, the network environment can be configured such that each subnet is a fat-tree topology that can be directly attached to the other subnets without any transit subnets in between.

[0071] For example, in Figure 7, the system 700 can use six routers 710 to connect two fat-tree subnets, e.g. subnets A-B 701-702. In Figure 8, the system 800 can use six routers 810 to connect three fat-tree subnets, e.g. subnets A-C 801-803. In Figure 9, the system 900 can create a 648-port three-stage fat-tree using eighteen routers 910 to connect six fat-tree subnets 901-906.

[0072] **Figure 10** illustrates an exemplary flow chart for supporting the inter-subnet fat-tree routing (ISFR) algorithm at a router in a network environment, in accordance with an embodiment of the invention. As shown in Figure 10, at step 1001, a router can receive a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment. Then, at step 1002, the router can obtain information, from one or more switches in the at least one subnet, on which downward output ports of the router can be used for routing the one or more packets. Furthermore, at step 1003, the router can build a routing table based on the obtained information.

[0073] **Figure 11** illustrates a functional block diagram to show the present feature. The present feature may be implemented as a system 1100 for routing traffic between distinct subnets in a network environment. The system 1100 includes one or more microprocessors and a router 1110 running on the one or more microprocessors. The one or more microprocessors includes: a receiving unit 1120 to receive a list of destinations that the router 1110 is responsible for routing one or more packets to, wherein the router 1110 connects to at least one subnet in the network environment; a generating unit 1130 to generate a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets; and a using unit 1140 to use a modulo based hash to select one port from the output router ports of the router.

[0074] According to an embodiment, there is disclosed a system for routing traffic between distinct subnets in a network environment operating on one or more microprocessors. The system comprises means for receiving, at a router, a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment. The system comprises means for generating a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets. The system comprises means for using a modulo based hash to select one router port from a plurality of output router ports of the router.

[0075] Preferably the system comprises means for performing round-robin path distribution among all available router that are connected to the at least one subnet.

[0076] Preferably the system comprises means for implementing the modulo based hash in a router firmware, wherein the modulo based hash can take the random number generated using the source LID and destination LID.

[0077] Preferably the system comprises means for receiving, from a subnet manager, a port mapping between one or more destination global identifiers (GIDs) and one or more router ports.

[0078] Preferably the system comprises means for allowing the subnets connected by the router to be configured in different network topologies.

[0079] Preferably the system comprises means for allowing the source LID to be associated with an egress port of an previous-hop router.

[0080] Preferably the system comprises means for generating a same random number for a given source and destination pair.

[0081] Preferably the system comprises means for obtaining a destination local identifier (LID) based on a destination global identifier (GID) associated with the one or more packets.

[0082] Preferably the system comprises means for returning, via a mapping function, a next-hop LID based on a subnet prefix located in a global identifier (GID).

[0083] Preferably the system comprises means for allowing the at least one subnet to be configured in a fat-tree topology.

[0084] According to one embodiment, there is disclosed a system for routing traffic between distinct subnets in a network environment. The system comprises a router, said router is configured to receive a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment. The router is configured to generate a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets and is also configured to use a modulo based hash to select one port from the output router ports of the router.

[0085] Preferably the system is capable of having the round-robin path distribution performed among all available router that are connected to the at least one subnet.

[0086] Preferably the system is capable of having the modulo based hash implemented in a router firmware, wherein the modulo based hash can take the random number generated using the source LID and destination LID.

[0087] Preferably the system is capable of having said router operate to receive, from a subnet manager, a port mapping between one or more destination global identifiers (GIDs) and one or more router ports.

[0088] Preferably the system is capable of having said router operate to allow the subnets connected by the router to be configured in different network topologies.

[0089] Preferably the system is capable of having the source LID associated with an egress port of an previous-hop router.

[0090] Preferably the system is capable of having said router operate to generate a same random number for a given source and destination pair.

[0091] Preferably the system is capable of having said router operate to obtain a destination local identifier (LID) based on a destination global identifier (GID) associated with the one or more packets.

[0092] Preferably the system is capable of having said router operate to obtain a next-hop LID based on a subnet prefix located in a global identifier (GID).

[0093] The present invention may be conveniently implemented using one or more conventional general purpose or specialized digital computer, computing device, machine, or microprocessor, including one or more processors, memory and/or computer readable storage media programmed according to the teachings of the present disclosure. Appropriate software coding can readily be prepared by skilled programmers based on the teachings of the present disclosure, as will be apparent to those skilled in the software art.

[0094] In some embodiments, the present invention includes a computer program product which is a storage medium or computer readable medium (media) having instructions stored thereon/in which can be used to program a computer to perform any of the processes of the present invention. The storage medium can include, but is not limited to, any type of disk including floppy disks, optical discs, DVD, CD-ROMs, microdrive, and magneto-optical disks, ROMs, RAMs, EPROMs, EEPROMs, DRAMs, VRAMs, flash memory devices, magnetic or optical cards, nanosystems (including molecular memory ICs), or any type of media or device suitable for storing instructions and/or data.

[0095] The foregoing description of the present invention has been provided for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise forms disclosed. Many modifications and variations will be apparent to the practitioner skilled in the art. The embodiments were chosen and described in order to best explain the principles of the invention and its practical application, thereby enabling others skilled in the art to understand the invention for various embodiments and with various modifications that are suited to the particular use contemplated. It is intended that the scope of the invention be defined by the following claims and their equivalence.

Claims:

What is claimed is:

1. A method for routing traffic between distinct subnets in a network environment, comprising:

receiving, at a router, a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment;

generating a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets; and

using a modulo based hash to select one router port from a plurality of output router ports of the router.

2. The method according to Claim 1, further comprising:

performing round-robin path distribution among all available router that are connected to the at least one subnet.

3. The method according to Claims 1 or 2, further comprising:

implementing the modulo based hash in a router firmware, wherein the modulo based hash can take the random number generated using the source LID and destination LID.

4. The method according to any preceding Claim, further comprising:

receiving, from a subnet manager, a port mapping between one or more destination global identifiers (GIDs) and one or more router ports.

5. The method according to any preceding Claim, further comprising:

allowing the subnets connected by the router to be configured in different network topologies.

6. The method according to any preceding Claim, further comprising:

allowing the source LID to be associated with an egress port of an previous-hop router.

7. The method according to any preceding Claim, wherein:

generating a same random number for a given source and destination pair.

8. The method according to any preceding Claim, further comprising:

obtaining a destination local identifier (LID) based on a destination global identifier (GID) associated with the one or more packets.

9. The method according to any preceding Claim, further comprising:
returning, via a mapping function, a next-hop LID based on a subnet prefix located in a global identifier (GID).

10. The method according to any preceding Claim, further comprising:
allowing the at least one subnet to be configured in a fat-tree topology.

11. A computer program comprising program instructions for running on one or more microprocessors to perform all the steps of the method of any preceding claim.

12. A computer program product comprising the computer program of claim 11 provided on a machine-readable medium.

13. A system for routing traffic between distinct subnets in a network environment, comprising:
one or more microprocessors,
a router running on the one or more microprocessors, wherein said router operates to
receive a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment;
generate a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets; and
use a modulo based hash to select one port from the output router ports of the router.

14. The system according to Claim 13, wherein:
round-robin path distribution is performed among all available router that are connected to the at least one subnet.

15. The system according to any of Claims 13 and 14, wherein:
the modulo based hash is implemented in a router firmware, wherein the modulo based hash can take the random number generated using the source LID and destination LID.

16. The system according to any of Claims 13-15, wherein:
said router operates to receive, from a subnet manager, a port mapping between one or more destination global identifiers (GIDs) and one or more router ports.

17. The system according to any of Claims 13-16, wherein:

said router operates to allow the subnets connected by the router to be configured in different network topologies.

18. The system according to any of Claims 13-17, wherein:

the source LID is associated with an egress port of an previous-hop router.

19. The system according to any of Claims 13-18, wherein:

said router operates to generate a same random number for a given source and destination pair.

20. The system according to any of Claims 13-19, wherein:

said router operates to obtain a destination local identifier (LID) based on a destination global identifier (GID) associated with the one or more packets.

21. The system according to any of Claims 13-20, wherein:

said router operates to obtain a next-hop LID based on a subnet prefix located in a global identifier (GID).

22. A non-transitory machine readable storage medium having instructions stored thereon that when executed cause a system to perform the steps comprising:

receiving, at a router, a list of destinations that the router is responsible for routing one or more packets to, wherein the router connects to at least one subnet in the network environment;

generating a random number based on a source local identifier (LID) and a destination LID associated with the one or more packets; and

using a modulo based hash to select one router port from a plurality of output router ports of the router.

23. A computer program for causing a computer to implement the method recited in any one of Claims 1 to 10.

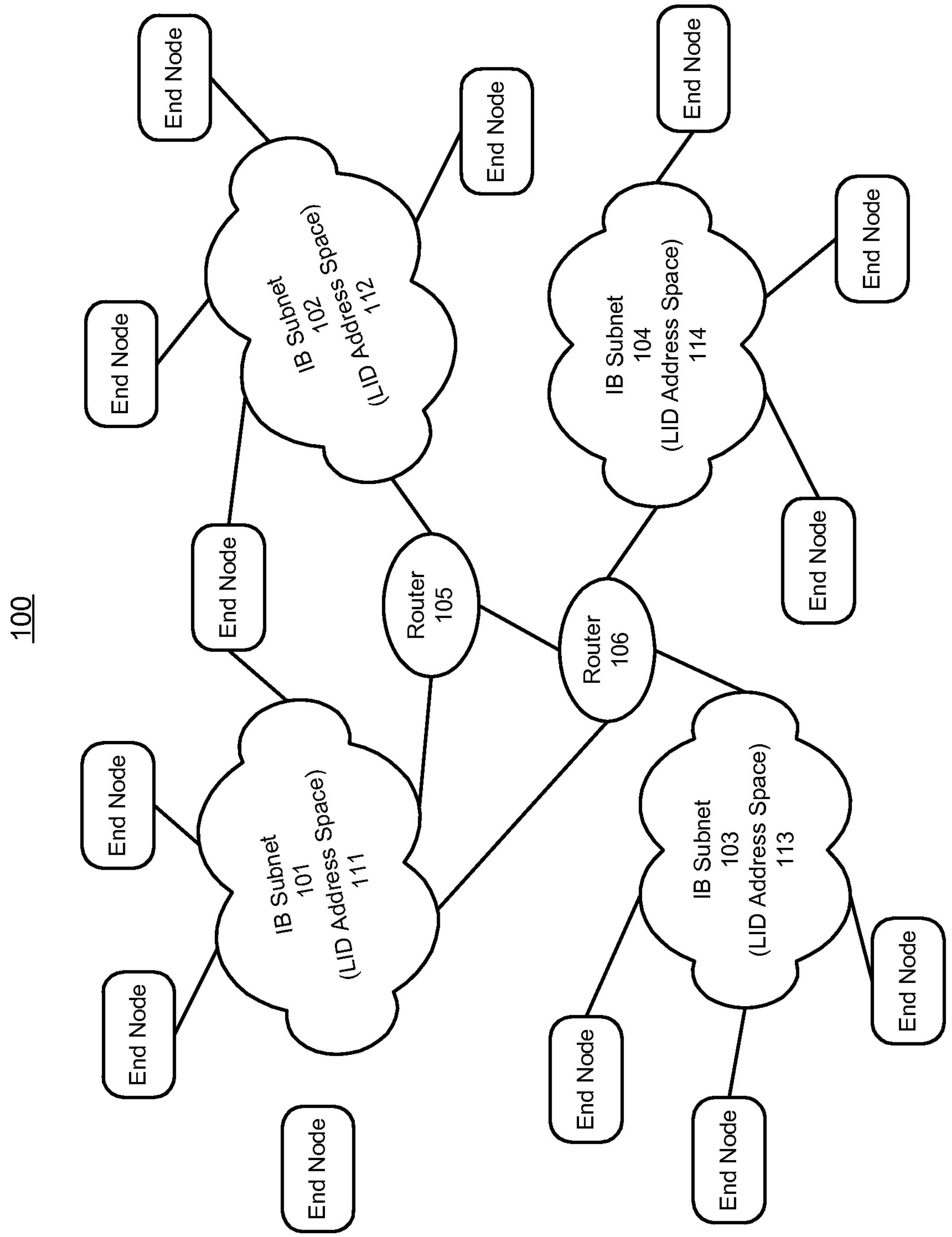

100
End Node
End Node
End Node
End Node
IB Subnet
102
(LID Address Space)
112
End Node
Router
105
Router
106
IB Subnet
101
(LID Address Space)
111
End Node
End Node
End Node
IB Subnet
104
(LID Address Space)
114
End Node
End Node
End Node
IB Subnet
103
(LID Address Space)
113
End Node
End Node
End Node

FIGURE 1

200

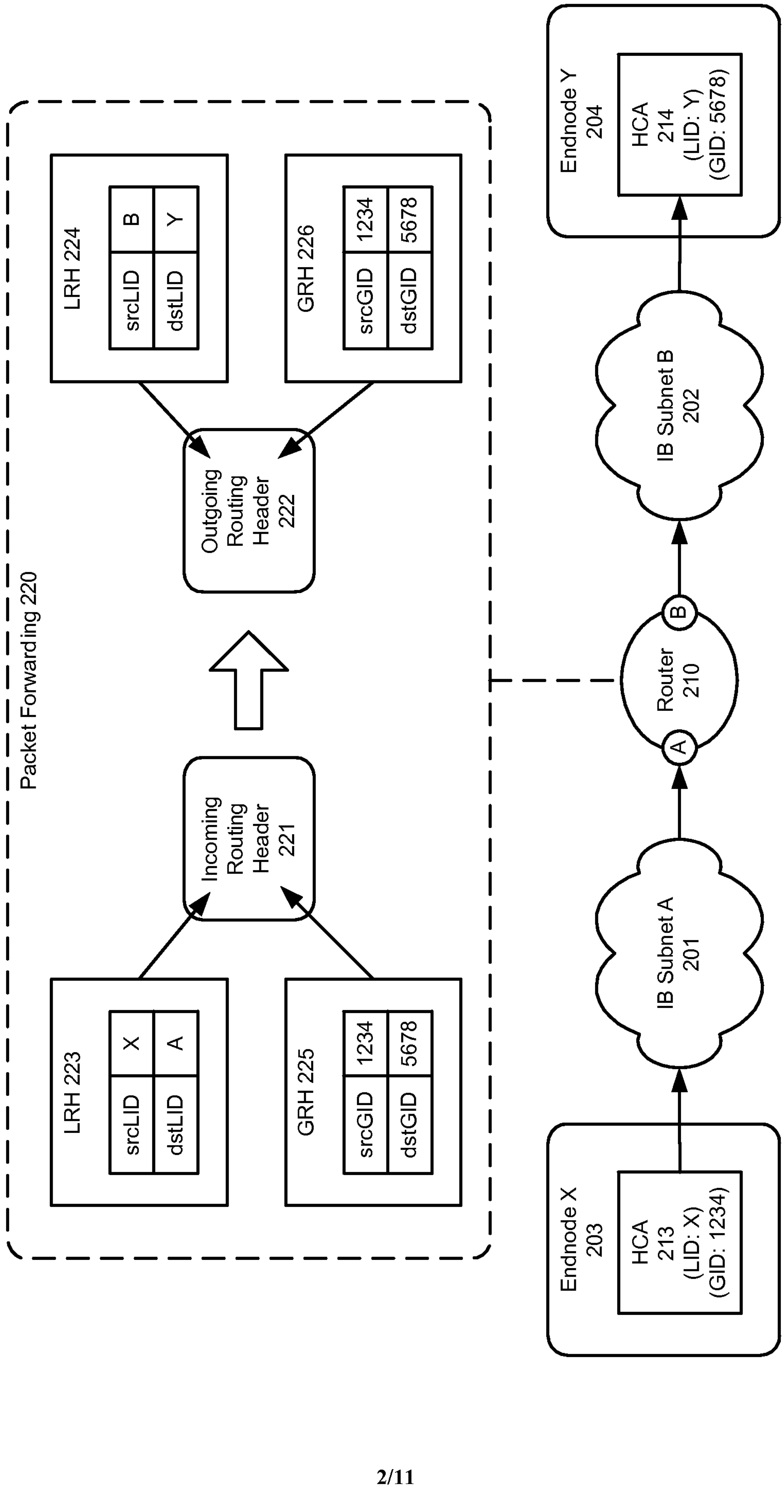

Packet Forwarding 220
LRH 223
srcLID
X
dstLID
A
GRH 225
srcGID
1234
dstGID
5678
Incoming Routing Header 221
Outgoing Routing Header 222
LRH 224
srcLID
B
dstLID
Y
GRH 226
srcGID
1234
dstGID
5678
Endnode X 203
HCA 213 (LID: X) (GID: 1234)
IB Subnet A 201
A
Router 210
B
IB Subnet B 202
Endnode Y 204
HCA 214 (LID: Y) (GID: 5678)

FIGURE 2

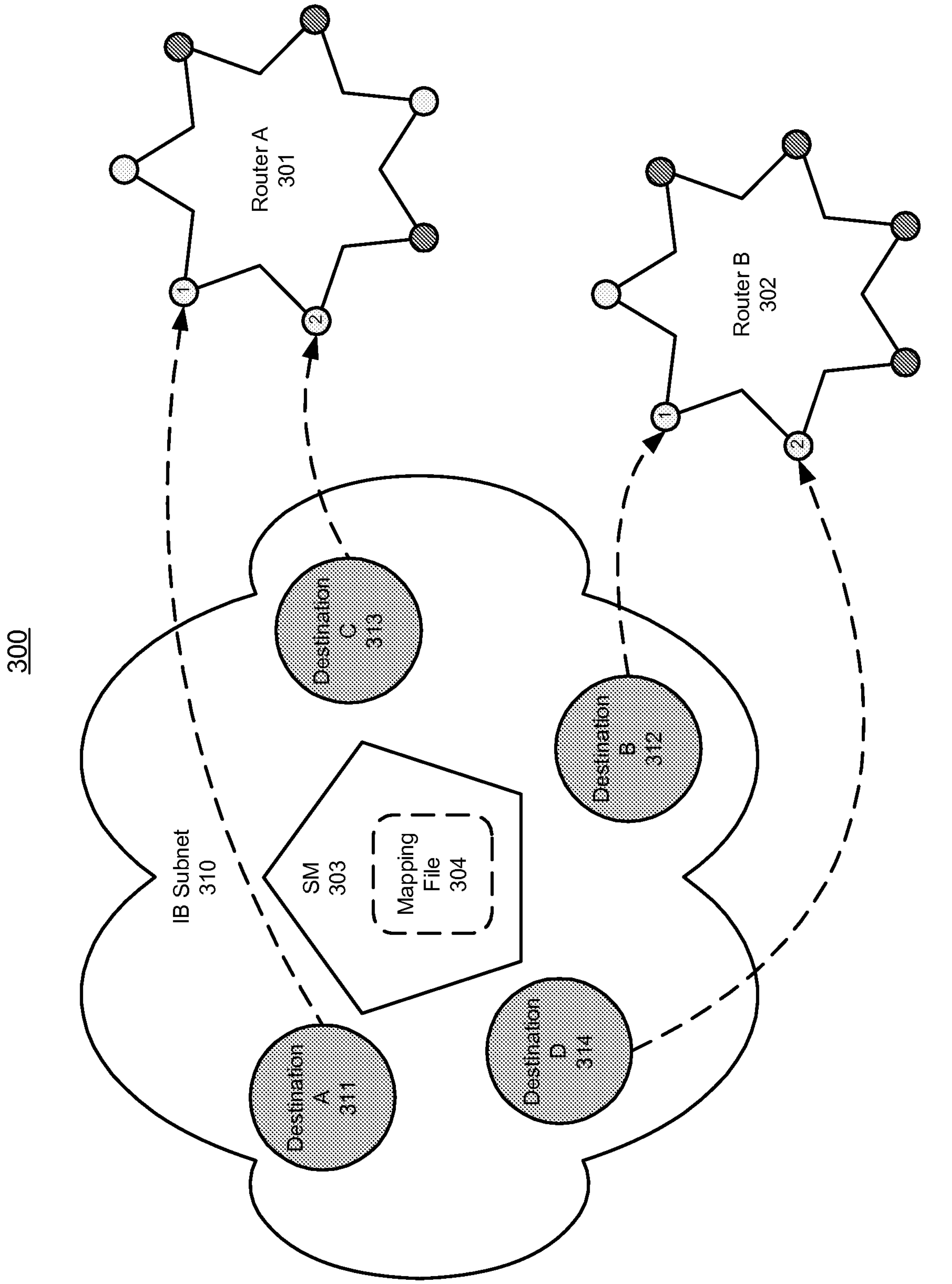

300
Router A
301
Router B
302
IB Subnet
310
SM
303
Mapping
File
304
Destination
A
311
Destination
B
312
Destination
C
313
Destination
D
314
1
2
1
2

FIGURE 3

400

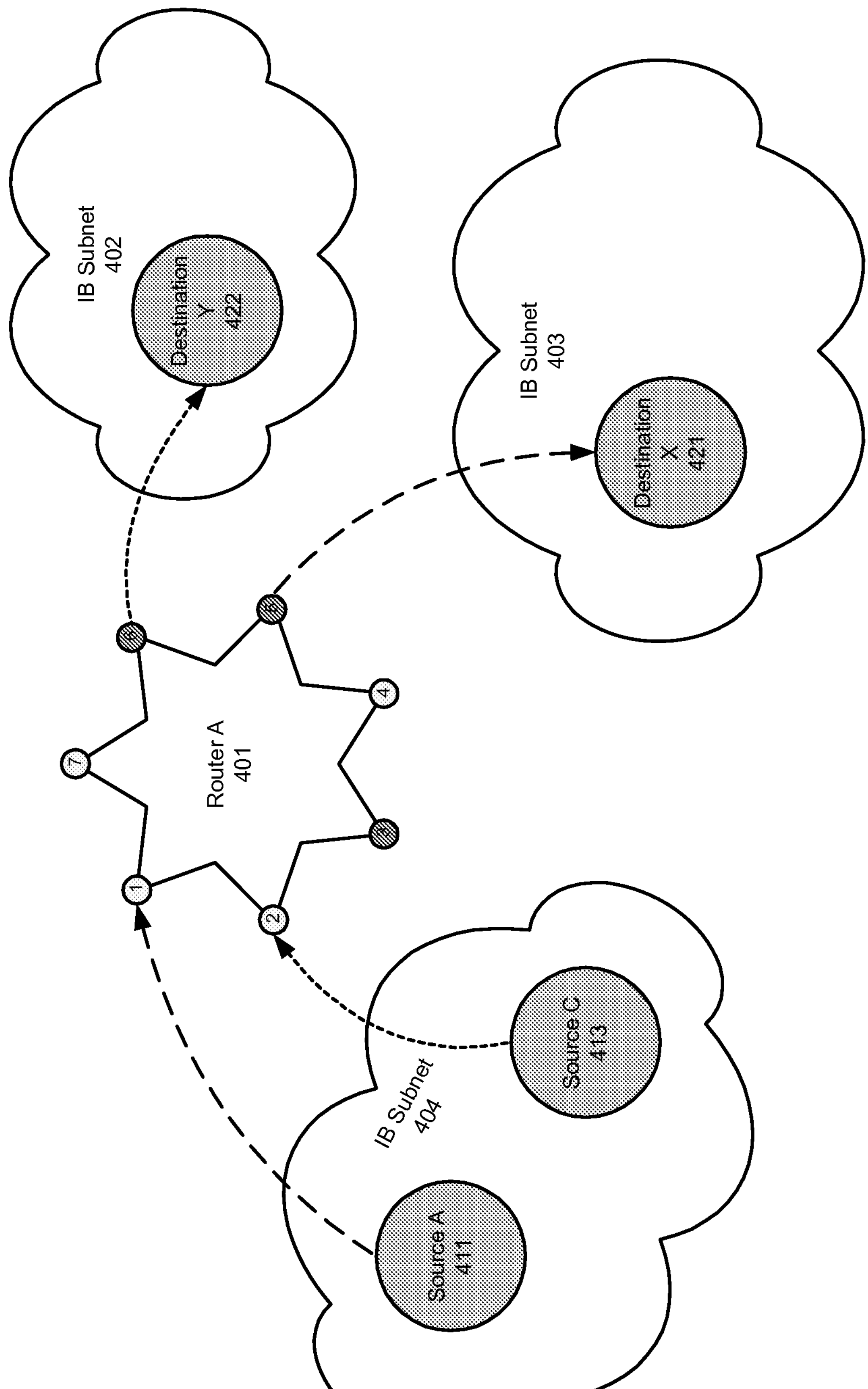

IB Subnet 402
Destination Y 422
IB Subnet 403
Destination X 421
Router A 401
7
1
2
4
IB Subnet 404
Source A 411
Source C 413

FIGURE 4

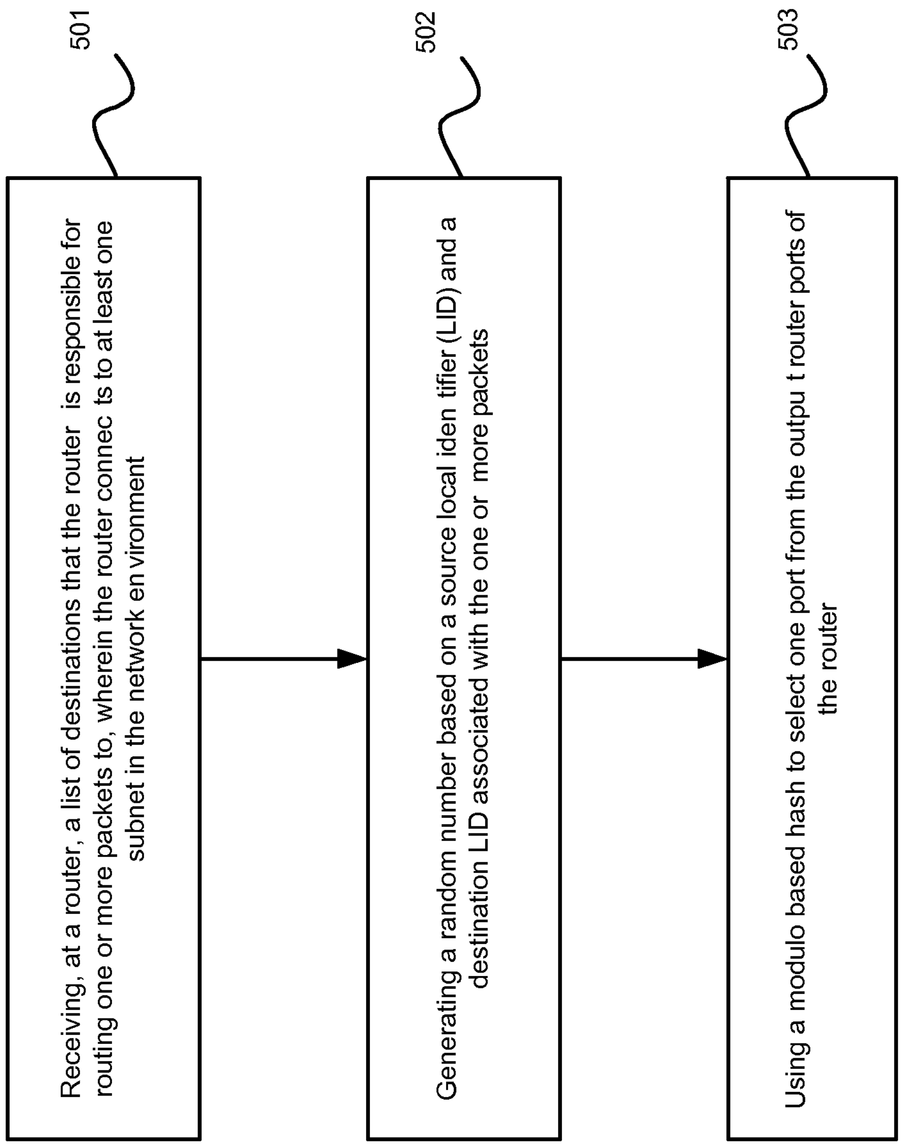

501
Receiving, at a router, a list of destinations that the router is responsible for routing one or more packets to, wherein the router connec ts to at least one subnet in the network en vironment
502
Generating a random number based on a source local iden tifier (LID) and a destination LID associated with the one or more packets
503
Using a modulo based hash to select one port from the outpu t router ports of the router

FIGURE 5

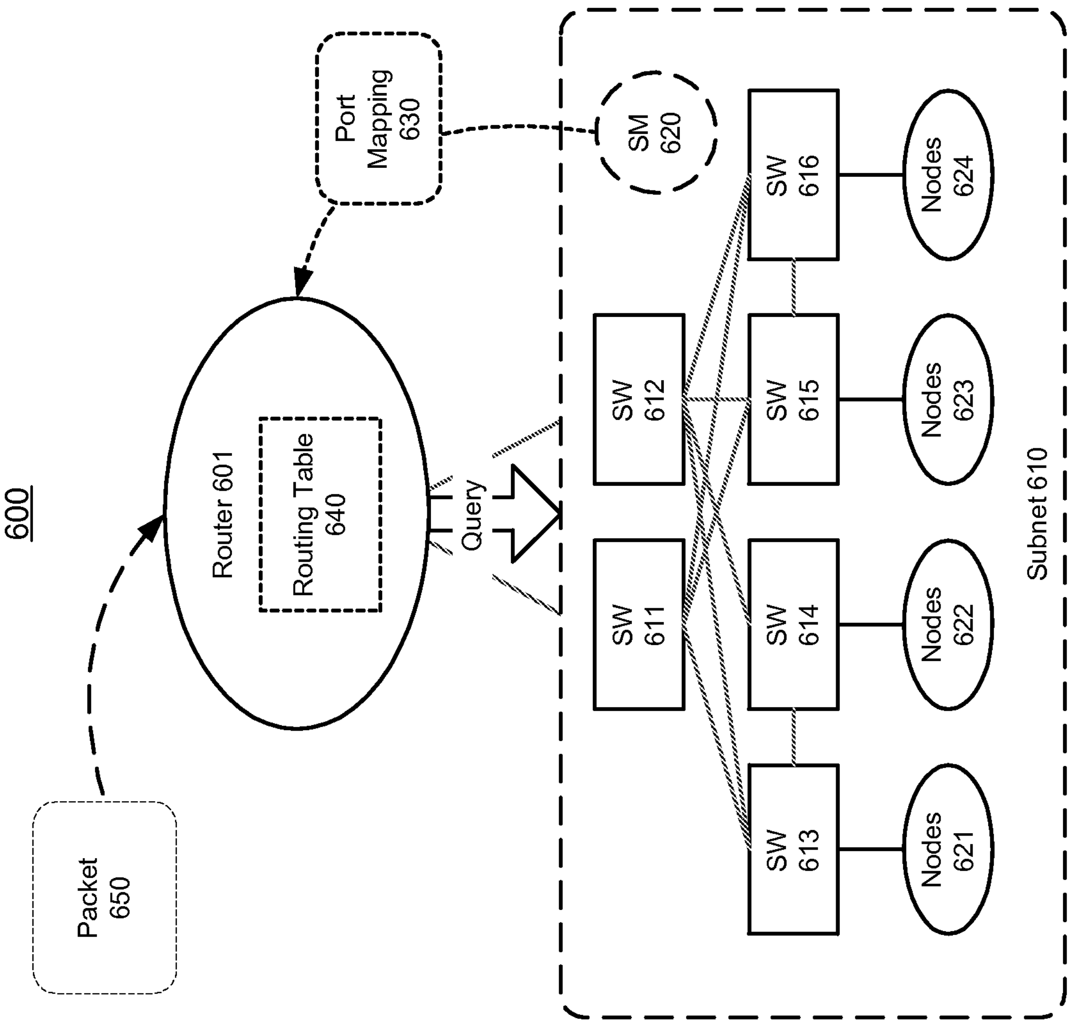

600
Packet 650
Router 601
Routing Table 640
Port Mapping 630
Query
SM 620
SW 611
SW 612
SW 613
SW 614
SW 615
SW 616
Nodes 621
Nodes 622
Nodes 623
Nodes 624
Subnet 610

FIGURE 6

700

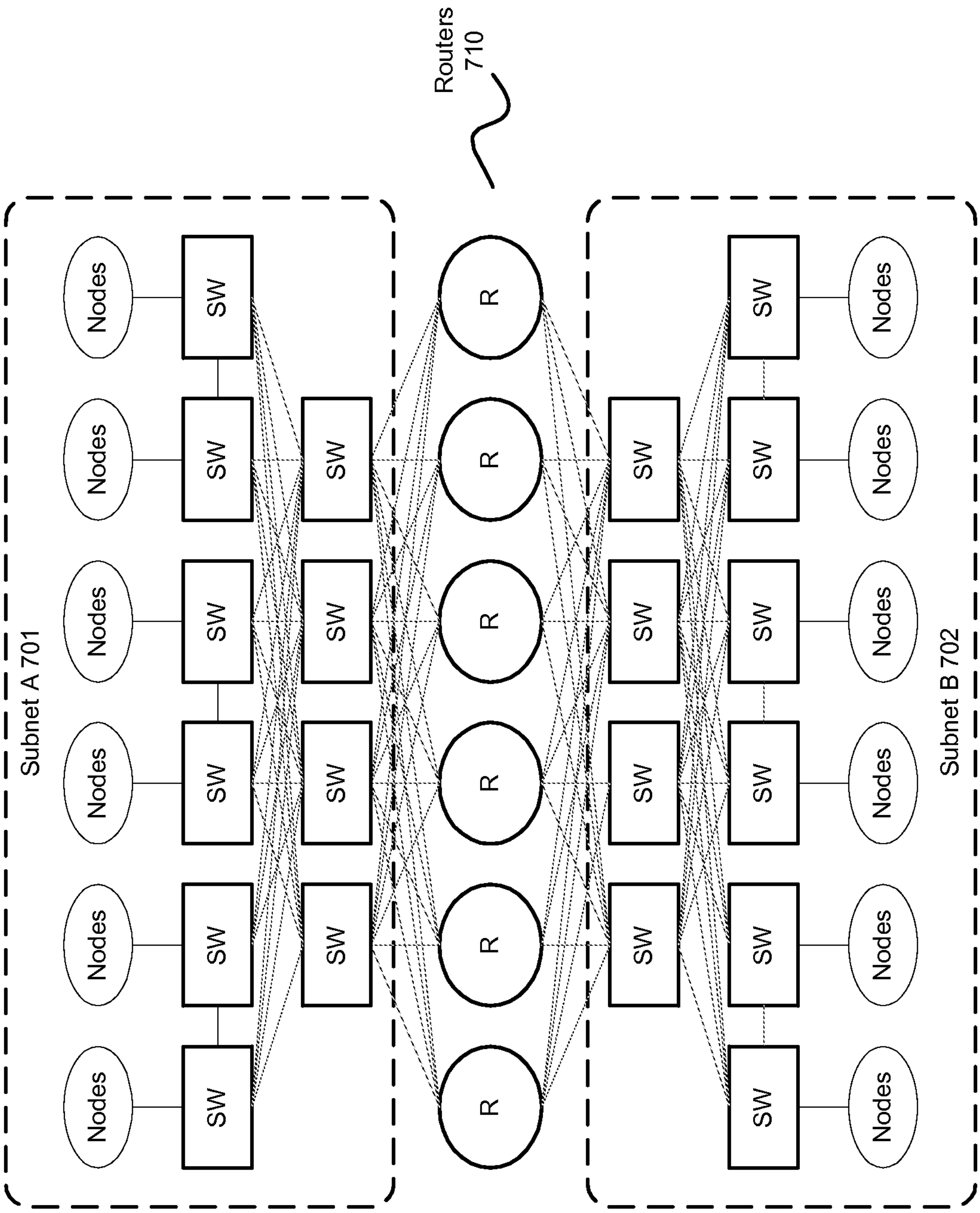
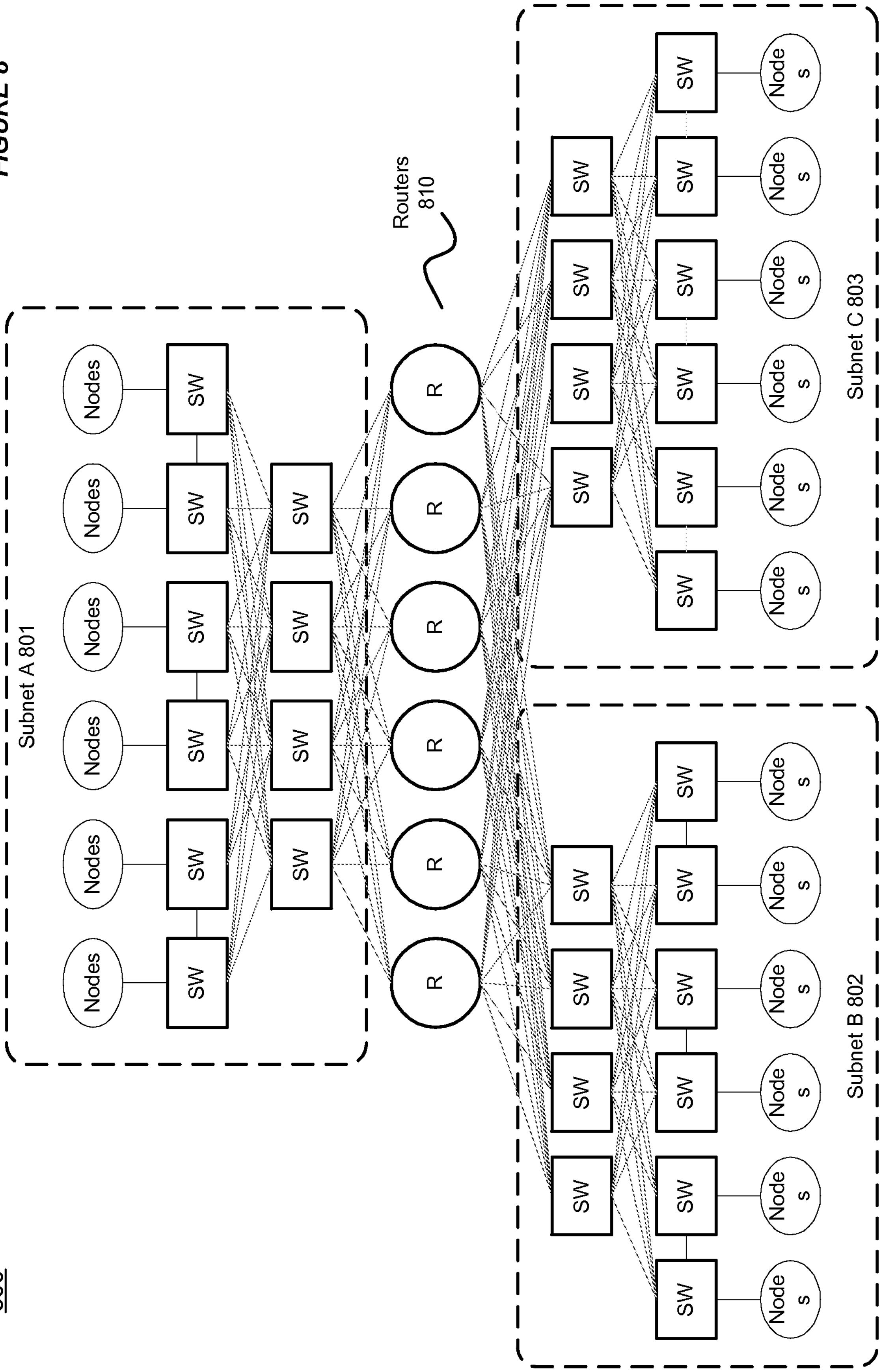


FIGURE 7

FIGURE 8

800


Subnet A 801
Nodes
SW
R
Routers
810
Subnet B 802
Subnet C 803
Node
s

900

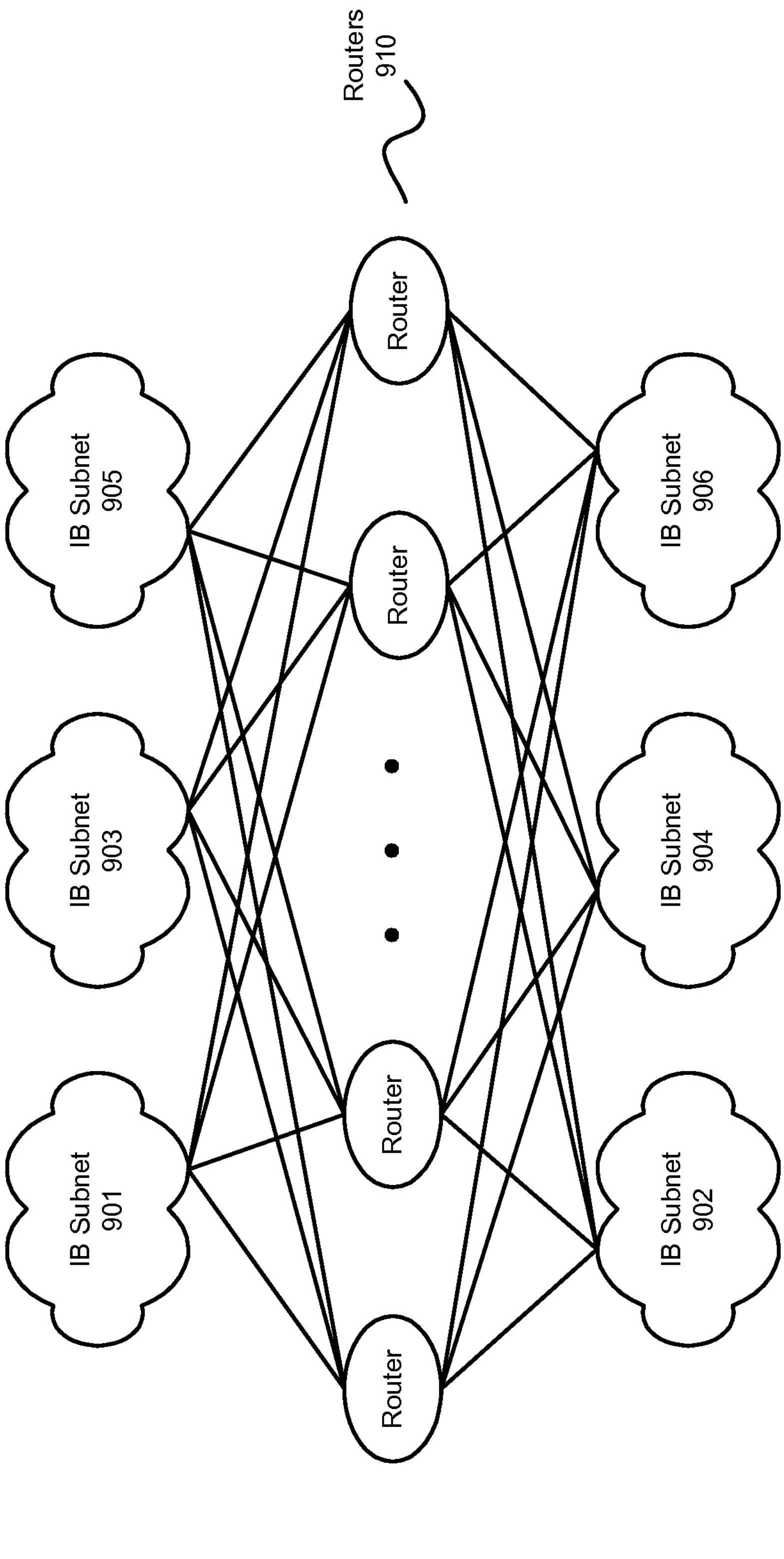

Routers
910
IB Subnet
905
IB Subnet
903
IB Subnet
901
Router
Router
Router
Router
IB Subnet
906
IB Subnet
904
IB Subnet
902

FIGURE 9

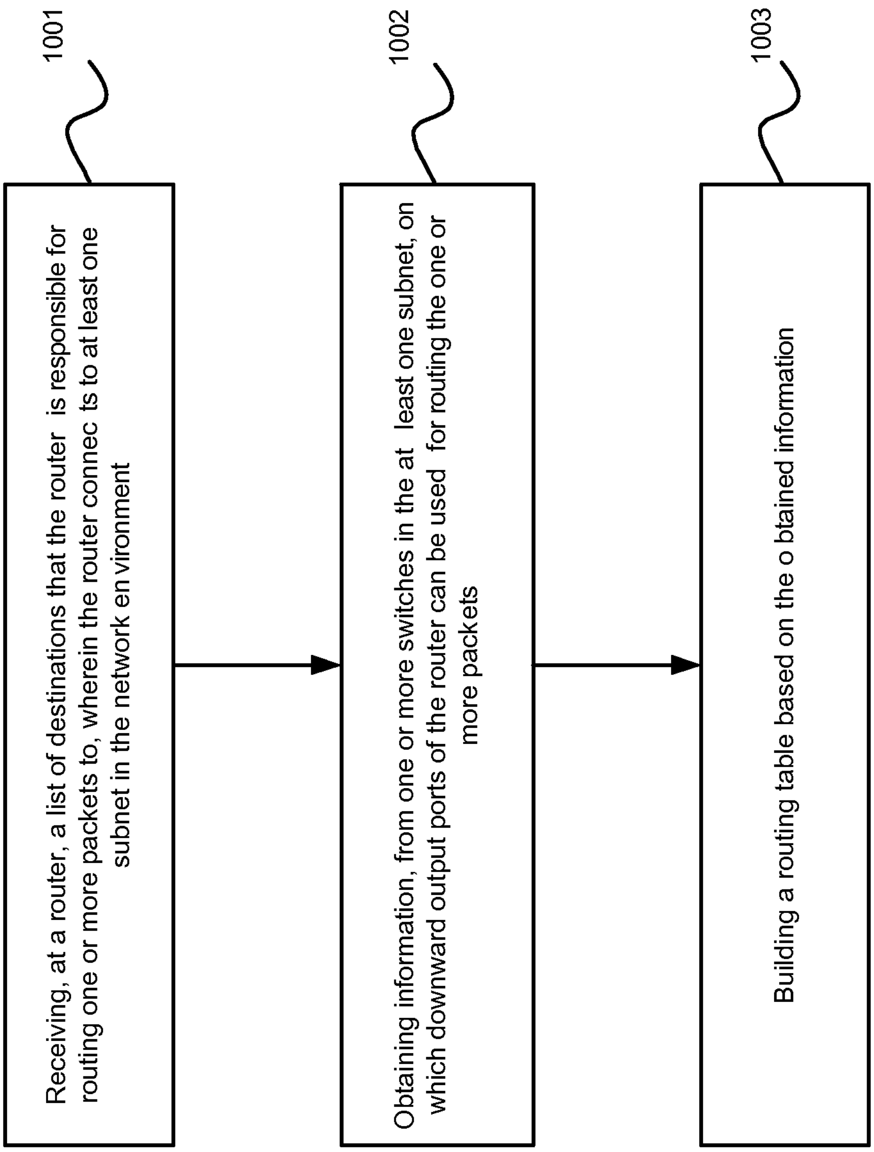

1001
Receiving, at a router, a list of destinations that the router is responsible for routing one or more packets to, wherein the router connec ts to at least one subnet in the network en vironment
1002
Obtaining information, from one or more switches in the at least one subnet, on which downward output ports of the router can be used for routing the one or more packets
1003
Building a routing table based on the o btained information

FIGURE 10

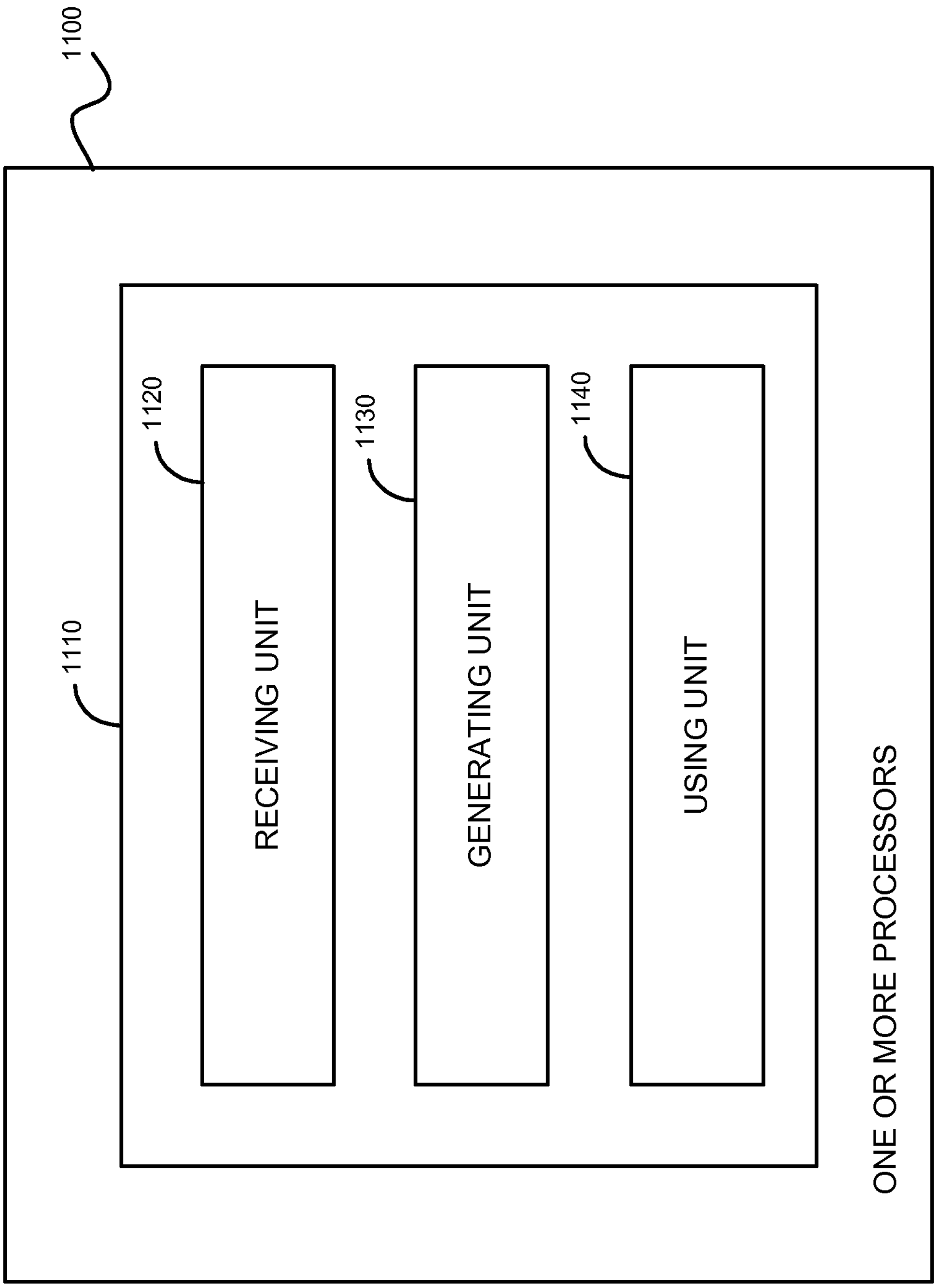

1100
1110
1120
1130
1140
RECEIVING UNIT
GENERATING UNIT
USING UNIT
ONE OR MORE PROCESSORS

FIGURE 11

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/040210

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F9/44 H04L29/12
ADD.

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)
H04L G06F

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2009/216853 A1 (BURROW STEPHEN R [US] ET AL) 27 August 2009 (2009-08-27) paragraph [0030] - paragraph [0032] paragraph [0048] - paragraph [0053] -----	1-23
A	VISHNU A ET AL: "Performance Modeling of Subnet Management on Fat Tree InfiniBand Networks using OpenSM", PARALLEL AND DISTRIBUTED PROCESSING SYMPOSIUM, 2005. PROCEEDINGS. 19TH IEEE INTERNATIONAL DENVER, CO, USA 04-08 APRIL 2005, PISCATAWAY, NJ, USA,IEEE, 4 April 2005 (2005-04-04), pages 296b-296b, XP010785940, DOI: 10.1109/IPDPS.2005.339 ISBN: 978-0-7695-2312-5 paragraph [2.background] paragraph [3.subnet.management.mechanism] -----	1-23

-/--

[X] 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
5 August 2013	12/08/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Lefebvre, Laurent

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2013/040210

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 7 401 157 B2 (COSTATINO LINO [US] ET AL COSTANTINO LINO [US] ET AL) 15 July 2008 (2008-07-15) column 3, line 30 - column 5, line 18 -----	1-23
A	BOGDANSKI B ET AL: "sFtree: A fully connected and deadlock free switch-to-switch routing algorithm for fat-trees", ACM TRANSACTIONS ON ARCHITECTURE AND CODE OPTIMIZATION, ASSOCIATION FOR COMPUTING MACHINERY, NEW YORK, NY, US , vol. 8, no. 4 14 January 2012 (2012-01-14), pages 1-20, XP002692976, ISSN: 1544-3566, DOI: 10.1145/2086696.2086734 Retrieved from the Internet: URL:http://simula.no/publications/Simula.simula.864 [retrieved on 2013-02-27] paragraph [2.the.infiniband.architecture] paragraph [3.fat.tree.routing] -----	1-23
A	NO AUTHOR NAME SUPPLIED IN SOURCE DATA: "Local Area Network (LAN) Emulation over InfiniBand", IP.COM JOURNAL, IP.COM INC., WEST HENRIETTA, NY, US, 18 September 2003 (2003-09-18), XP013012867, ISSN: 1533-0001 the whole document -----	1-23

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2013/040210

Patent document cited in search report		Publication date	Patent family member(s)	Publication date
US 2009216853	A1	27-08-2009	NONE	
US 7401157	B2	15-07-2008	NONE	