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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 April 2019 (18.04.2019)

WIPO | PCT

(10) International Publication Number
WO 2019/073452 A1

(51) International Patent Classification:
H04L 12/66 (2006.01)

(21) International Application Number:
PCT/IB2018/057945

(22) International Filing Date:
12 October 2018 (12.10.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
2982147 12 October 2017 (12.10.2017) CA

(71) Applicant: **ROCKPORT NETWORKS INC.** [CA/CA]; Ste. 600, 515 Legget Drive, Ottawa, Ontario K2K 3G4 (CA).

(72) Inventor: **WILLIAMS, Matthew Robert**; 7 Insmill Crescent, Kanata, Ontario K2T 1G5 (CA).

(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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, 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, SA, 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, ST, 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, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *as to the identity of the inventor (Rule 4.17(i))*
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

Published:
— *with international search report (Art. 21(3))*
— *in black and white; the international application as filed contained color or greyscale and is available for download from PATENTSCOPE*

(54) **Title:** DIRECT INTERCONNECT GATEWAY

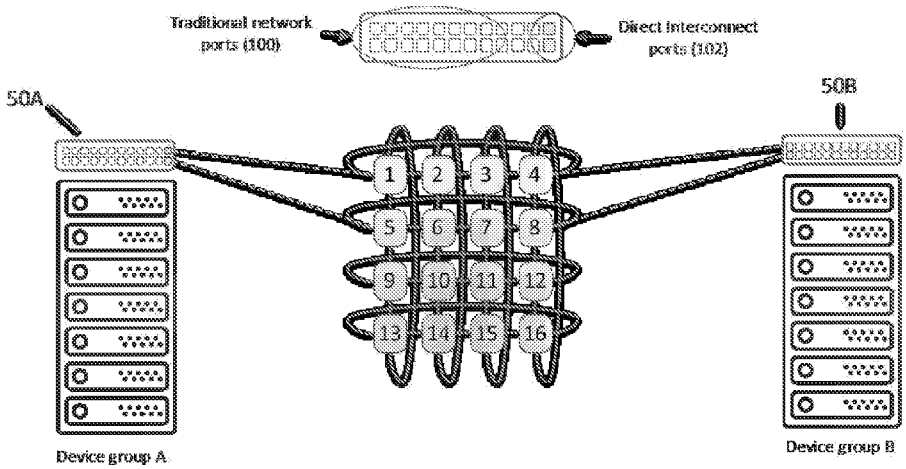

FIGURE 6

(57) **Abstract:** A dedicated network gateway device that is capable of bridging, switching or routing network traffic between traditional network and direct interconnect networks, comprising: a first set of one or more traditional network ports with a single link per port generally comprising one or more of SFP+, QSFP, and QSFP+ connectors, such ports being connected to switches or devices that form a traditional network; and a second set of one or more direct interconnect ports with a high number of links per port (two or more) generally comprising one or more of MXC, MTP, and MTO connectors, such ports being connected to a direct interconnect network.

WO 2019/073452 A1

DIRECT INTERCONNECT GATEWAY

FIELD OF THE INVENTION

[0001] The present invention relates to I/O (Input/Output) traffic in a computer network. More specifically, the present invention relates to a dedicated device that bridges, switches or routes data between traditional networks and direct interconnect networks.

BACKGROUND OF THE INVENTION

[0002] Computer networks allow a multitude of nodes to route or otherwise exchange data with each other. As a result, computer networks are able to support an immense number of applications and services such as the shared use of storage servers, access to the World Wide Web, use of email, etc.

[0003] Nodes themselves can often be characterized into three types based on the specialized tasks that they perform: computation nodes, such as servers having CPUs that perform calculations (but that generally have little to no local disk space); I/O nodes that contain the system's secondary storage and provide parallel file-system services; and gateway nodes that provide connectivity to external data servers and mass storage systems. Some nodes can even serve more than one function, such as, for instance, handling both I/O and gateway functions.

[0004] I/O for parallel and distributed systems, however, has become a huge concern for both users and designers of computer systems. In this respect, while the speeds of CPUs have been increasing at an exponential rate virtually every year, the speed of I/O devices has unfortunately increased at a slower pace, often due to the fact that they can be more limited by the speed of mechanical components. I/O performance, a measure of I/O data traffic between nodes, is therefore often a limiting factor in network performance. Indeed, the mismatch in speed between CPUs and I/O is accentuated in parallel and distributed computer systems, leaving I/O as a bottleneck that can severely limit scalability. This is especially the case when the network is involved with commercial applications involving multimedia and scientific modelling, for instance, each of which has huge I/O requirements.

[0005] Direct interconnect networks, such as those disclosed in PCT Patent Application Publication No. WO 2015/027320 A1 (which describes a novel torus or

higher radix interconnect topology for connecting network nodes in a mesh-like manner in parallel computer systems), generally restrict traffic to nodes that are part of the direct interconnect. While the novel system and architecture disclosed in PCT Patent Application Publication No. WO 2015/027320 A1 is particularly beneficial and practical for commercial deployment in data centers and cloud data centers, most data centers in operation today are still based, unfortunately, on a traditional legacy three-tier architecture, a fat tree architecture, or a DCell server-centric architecture, among others. With data centers based on these architectures, it is unfortunately either undesirable or impossible for them to join a direct interconnect, and they are therefore unable to exploit the benefits of such a network topology. Some prior art direct interconnect architectures have provided a system wherein each node, or a subset of nodes (i.e. gateway nodes), have dual connectivity, both to the direct interconnect and to the traditional network, but such nodes are difficult to manage and load the resources of the device as they bridge or route between the two networks.

[0006] It would therefore be desirable to have a direct interconnect gateway that is designed and capable of allowing direct interconnect devices and non-direct interconnect devices to communicate. Moreover, it would be beneficial to have a gateway that could assist in overcoming some of the shortcomings described above for I/O traffic.

SUMMARY OF THE INVENTION

[0007] In one aspect, the present invention provides for a dedicated device, namely a gateway device, that is capable of bridging, switching or routing between traditional and direct interconnect networks.

[0008] In another aspect, the present invention provides a highly manageable gateway device that can be managed by network management systems.

[0009] In yet another aspect, the present invention provides a gateway device that allows for the coordination of MAC tables and ARP, broadcast, multicast and anycast responses between multiple direct interconnect ports.

[00010] In one embodiment, the present invention provides a dedicated network gateway device that is capable of bridging, switching or routing network traffic between traditional and direct interconnect networks, comprising: a first set of one or more traditional network ports with a single link per port, such ports being connected to

switches or devices that form a traditional network; and a second set of one or more direct interconnect ports with two or more links per port, such ports being connected to a direct interconnect network. The traditional network ports may comprise one or more of SFP+, QSFP, and QSFP+ connectors, and may be connected to switch or router ports in the traditional network, while the direct interconnect ports may comprise one or more of MXC, MTP, and MTO connectors, and may be connected to a passive patch panel/hub used in the implementation of the direct interconnect network. Alternatively, the direct interconnect ports may be each connected to their own dedicated direct interconnect application-specific integrated circuit (ASIC), or they may each be connected to one or more shared application-specific integrated circuits (ASICs). The bridging, switching or routing function may be performed by a network switch ASIC or by a network controller ASIC. The ASICs may be capable of acting as a direct interconnect node with locally destined/sourced traffic sent over a traditional network interface line. In addition, in other embodiments, the direct interconnect ports may take the place of a device within the direct interconnect network, and they may even be connected to multiple direct interconnect networks.

[00011] In another embodiment, the present invention provides a dedicated network gateway device that is capable of bridging, switching or routing network traffic between traditional and direct interconnect networks, wherein said device comprises two ports, namely a first port that is a direct interconnect port that is capable of being connected to a direct interconnect network, and a second port that is a standard network port that is capable of being connected to switches or devices that form a traditional network. The first port may comprise one of a MXC, MTP, or MTO connector, and may be connected to a passive patch panel/hub used in the implementation of the direct interconnect network. The second port may comprise one of a SFP+, QSFP, or QSFP+ connector, and may be connected to switch or router ports in the traditional network.

[00012] In yet another embodiment, the present invention provides a dedicated network gateway device that is capable of bridging or routing network traffic between a traditional network and a direct interconnect network, comprising: a first set of traditional network ports with a single link per port, such ports being connected to end devices that form a first traditional network; and a second set of direct interconnect ports with two or more links per port, such ports being connected to the direct interconnect network, wherein said direct interconnect network acts as a backbone

that allows network traffic to route from the dedicated network gateway device to another dedicated network gateway device, said another dedicated network gateway device comprising: a first set of traditional network ports with a single link per port, such ports being connected to end devices that form a second traditional network; and a second set of direct interconnect ports with two or more links per port, such ports being connected to the direct interconnect network.

[00013] In yet a further embodiment, the present invention provides a dedicated network gateway device that is capable of bridging, switching or routing network traffic between traditional network and direct interconnect networks, comprising: a first set of one or more traditional network ports with a single link per port, such ports being connected to switches or devices that form a traditional network; a second set of one or more direct interconnect ports with one or more links per port, such ports being connected to a direct interconnect network; and a plurality of direct interconnect ports that are logically associated to act as a single direct interconnect node.

[00014] In another embodiment, the present invention provides a computer-implemented method of bridging, switching or routing network traffic between a traditional network and a direct interconnect network, comprising the steps of: connecting the dedicated gateway device to the traditional network and the direct interconnect network, said dedicated gateway device acting as one or more nodes within the direct interconnect network; and forwarding the network traffic by means of the gateway device between the traditional network and the direct interconnect network based on headers or content of the network traffic.

[00015] In a further embodiment, the present invention provides a computer-implemented method of coordinating which gateway device should provide access to a resource located in a traditional network when said resource is accessible by more than one gateway device, comprising the steps of: (i) receiving an ARP, broadcast, multicast, or anycast traffic at a direct interconnect port, wherein said traffic is requesting access to the resource located in the traditional network, and wherein said direct interconnect port is linked via one or more hops to the more than one gateway devices, each of which is capable of providing access to the resource; (ii) calculating an optimum gateway device port out of the more than one gateway devices that should provide access to the resource; (iii) creating an association between the traffic, the direct interconnect node, and the calculated optimum gateway device port; and (iv) communicating the association with each of the more than one gateway devices to

ensure that the calculated optimum gateway device port provides access to the resource. The step of calculating the optimum gateway device port that should provide access to the resource may comprise determining which of the more than one gateway device ports is closest to the direct interconnect port or may comprise employing a consensus algorithm to ensure consistency of the traffic. The step of communicating the association may be handled by a dedicated or shared coordination bus.

BRIEF DESCRIPTION OF THE DRAWINGS

[00016] The embodiment of the invention will now be described, by way of example, with reference to the accompanying drawings in which:

[00017] **FIGURE 1** is a high-level overview of a gateway device in accordance with one embodiment of the present invention.

[00018] **FIGURE 2** is an overview of a gateway device in accordance with one embodiment of the present invention, comprising a traditional network switch ASIC with certain ports connected to traditional network connectors and other ports connected to their own dedicated direct interconnect ASIC.

[00019] **FIGURE 2b** is an overview of a gateway device in accordance with one embodiment of the present invention (related to that shown in Figure 2), wherein each direct interconnect ASIC can be connected to more than one direct interconnect port, and can be connected to more than one traditional network switch ASIC port.

[00020] **FIGURE 2c** is an overview of a gateway device in accordance with one embodiment of the present invention, comprising a single switch and direct interconnect ASIC that combines the functions of a traditional network switch ASIC and one or more direct interconnect ASICs.

[00021] **FIGURE 2d** is an overview of a gateway device in accordance with one embodiment of the present invention, comprising a host interface card containing a traditional network port and a direct interconnect port.

[00022] **FIGURE 2e** is an overview of a gateway device in accordance with one embodiment of the present invention, comprising a host interface card where the functions of a direct interconnect ASIC and traditional network controller ASIC are combined into a single direct interconnect and traditional ASIC.

[00023] **FIGURE 2f** is an overview of a gateway device in accordance with one embodiment of the present invention, wherein the direct interconnect ports contain one or more links per port and are combined into groups by the gateway device, such

that each group is logically associated by the gateway device to act as a single node within the direct interconnect network.

[00024] **FIGURE 3** is an overview of a gateway device in accordance with one embodiment of the present invention wherein direct interconnect ports are connected to a passive patch panel/hub.

[00025] **FIGURE 4** is an overview of a gateway device in accordance with one embodiment of the present invention wherein a direct interconnect port replaces a server within a direct interconnect topology.

[00026] **FIGURE 4b** displays an embodiment as shown in Figure 4, wherein the links of the direct interconnect ports are connected to different nodes.

[00027] **FIGURE 5** is an overview of one embodiment of the present invention wherein the direct interconnect ports on the gateway are linked to different direct interconnects to allow bridging, switching or routing between multiple direct interconnects.

[00028] **FIGURE 6** is an overview of one embodiment of the present invention wherein all or a majority of the nodes within the direct interconnect are composed of gateway ports, and wherein all or a majority of the other devices would be connected directly to one or more gateways.

[00029] **FIGURE 7** provides an example of how to minimize the average distance (in hops) from each device in the direct interconnect to the nearest gateway port.

[00030] **FIGURE 8** provides an example of how the gateway may coordinate which direct interconnect port should respond to an ARP request received at more than one port in the same direct interconnect.

[00031] **FIGURE 8b** provides an example of an embodiment of a coordination mechanism (as per Figure 8) wherein the coordination function is distributed across the direct interconnect ASICs and each direct interconnect ASIC is connected to a coordination bus used to communicate coordination information and decisions.

[00032] **FIGURE 9** provides one embodiment of a logic tree explaining how, when more than one gateway is connected to the same torus, the gateways may coordinate their knowledge of the torus topology and response to ARP, broadcast, multicast and anycast traffic.

[00033] **FIGURE 10** provides an example of how an intermediate gateway node may process a packet instead of routing the packet to another gateway node.

[00034] **FIGURES 11** and **12** describe the operation of a preferred embodiment of a direct interconnect ASIC.

DETAILED DESCRIPTION OF THE INVENTION

[00035] The present invention provides for a dedicated device, namely a gateway device, that is capable of bridging, switching or routing between traditional and direct interconnect networks. By employing such a dedicated device, the resources on the direct interconnect nodes do not have to be burdened by bridging, switching or routing between a direct interconnect and traditional network, thereby minimizing impacts on I/O performance. In addition, as opposed to the prior art use of gateway nodes, the present gateway device is a highly manageable device that can be managed by network management systems. Moreover, the gateway device of the present invention allows for the coordination of MAC tables and ARP, broadcast, multicast and anycast responses between multiple direct interconnect ports.

[00036] **Figure 1** shows a high-level overview of a gateway device **50** in accordance with one embodiment of the present invention, comprising two sets of ports. The first set of ports (in this example, the left-most twelve ports) are standard, traditional network ports **100** with a single link per port (e.g. SFP+, QSFP, QSFP+ connectors) that are connected to the existing switches and/or devices that form the traditional network. In a Clos topology (a multi-stage circuit switching network), for example, these ports **100** would most likely be connected to spine or super-spine ports. The second set of ports (the right-most twelve ports) are direct interconnect ports **102** with a high number of links (two or more) per port (e.g. with MXC/MTP/MTO connectors). If a passive patch panel/hub **60** similar to that disclosed in PCT Patent Application Publication No. WO 2015/027320 A1 is being used to support the direct interconnect topology, for example, then the direct interconnect ports **102** are connected to the passive patch panel/hub **60** (see **Figure 3**). The gateway device **50** may be in the form of a 1 rack unit (RU) rack-mountable device, or even ½ RU (or otherwise as desired) for efficient rack space savings. **Figure 2** shows an embodiment comprised of a traditional network switch application-specific integrated circuit (ASIC) **106**, wherein the first set of traditional network ports **100** are each connected to traditional network connectors, and wherein the second set of direct interconnect ports **102** are each connected to their own dedicated direct interconnect ASIC **104**. These

dedicated direct interconnect ASICs **104** are preferably each capable of acting as a direct interconnect node with locally destined/sourced traffic sent over a traditional network interface line (e.g. 100 Gbps Ethernet). The connections between the ports of switch ASIC **106** and direct interconnect ASICs **104** can be implemented using any standard component interconnect such as 100GBase-KR4, 100GBase-KP4, 40GBase-KR4, 25GBase-KR, 10GBase-KR or any other similar standard.

[00037] Traditional network switch ASIC **106** contains standard network traffic forwarding functionality, including learning the devices reachable through its ports, sending received traffic through the appropriate egress port, network filtering, traffic inspection and other functionality typically found in layer 2, layer 3 and layer 4 and above network switches, routers and bridges. Forwarding decisions may be based on one or more factors, including but not limited to source and destination layer 2 (MAC) addresses, source port, source and destination layer 3 (IPv4, IPv6, etc.) addresses, source and destination layer 4 ports, and layer 5 and above headers and data payloads.

[00038] Network traffic received from the direct interconnect at a direct interconnect ASIC **104** that has an ultimate destination that is reachable through the standard ports of switch ASIC **106** will be sent from direct interconnect ASIC **104** to switch ASIC **106** where switch ASIC **106** standard traffic forwarding functionality will transmit the traffic through the appropriate standard port.

[00039] Similarly, network traffic received by switch ASIC **106** from a standard port that has an ultimate destination that is reachable through a direct interconnect ASIC **104** will be forwarded by switch ASIC **106** to direct interconnect ASIC **104**.

[00040] In another embodiment (not shown), and as applicable for every possible embodiment, it would be understood that switch ASIC **106** (and all similar ASICs discussed herein) may be replaced by a field-programmable gate array (FPGA), general purpose processor, network processor or any other device capable of performing network traffic forwarding.

[00041] In another embodiment (not shown), and as applicable for every possible embodiment, it would be understood that direct interconnect ASIC **104** (and all similar ASICs discussed herein) may be replaced by a field-programmable gate array (FPGA), general purpose processor, network processor or any other device capable of acting as a node in a direct interconnect network.

[00042] **Figure 2b** shows another embodiment where each direct interconnect ASIC **104** can be connected to one or more direct interconnect ports **102** and can be connected to one or more traditional network switch ASIC **106** ports.

[00043] **Figure 2c** shows another embodiment where a single switch and direct interconnect ASIC **105** combines the functions of the traditional network switch ASIC **106** and one or more direct interconnect ASICs **104**. Once again, it would be understood that single switch and direct interconnect ASIC **105** (and all similar ASICs discussed herein) may be replaced by a field-programmable gate array (FPGA), general purpose processor, network processor or any other device capable of performing network traffic forwarding and/or acting as a node in a direct interconnect network.

[00044] **Figure 2d** shows another embodiment of a gateway device in the form of a host interface card **110** containing a traditional network port and direct interconnect port. Host interface card **110** is designed to be connected to a host device via a communication bus such as PCIe, Gen-Z or CCIX. The host device could be a server, computer, hard drive, solid state drive, drive cabinet or any other device capable of sending and receiving data. The traditional network port is connected to standard network controller ASIC **108** (or similarly a field-programmable gate array (FPGA), general purpose processor, network processor or any other device capable of performing network traffic forwarding, etc.) and the direct interconnect port is connected to direct interconnect ASIC **104**. Direct interconnect ASIC **104** is also connected to network controller ASIC **108**. In this form, network controller ASIC **108** would be capable of switching traffic between the traditional network and direct interconnect network without intervention by the host.

[00045] **Figure 2e** shows an embodiment of another host interface card **111** where the functions of the direct interconnect ASIC **104** and network control ASIC **108** are combined into a single direct interconnect and traditional ASIC **107**. Once again, it would be understood that single direct interconnect and traditional ASIC **107** (and all similar ASICs discussed herein) may be replaced by a field-programmable gate array (FPGA), general purpose processor, network processor or any other device capable of performing network traffic forwarding and/or acting as a node in a direct interconnect network.

[00046] **Figure 2f** shows another embodiment where the direct interconnect ports **102** contain one or more links per port and are combined into groups **55** by the

gateway device **50**. Each group **55** is logically associated by the gateway device **50** to act as a single node within the direct interconnect network.

[00047] As shown in **Figure 4**, the direct interconnect ports **102** and groups **55** would each take the place of a device within the direct interconnect topology.

[00048] In yet another embodiment, if a passive patch panel/hub **60** is not utilized in the direct interconnect, then the individual links of each gateway port (i.e. the direct interconnect ports **102**) may be connected to devices that are part of the direct interconnect. In this respect, **Figure 4b** shows an embodiment wherein the links of one of the direct interconnect ports **102** are individually connected to neighbor nodes within the direct interconnect network. In **Figure 4b**, the direct interconnect takes the form of a two-dimensional torus and the 4 links that comprise the direct interconnect port **102** are numbered 1, 2, 3 and 4. In order to form the two-dimensional torus, link 1 is connected to device A, link 2 is connected to device B, link 3 is connected to device C and link 4 is connected to device D. Similarly, for other topologies, each link of the direct interconnect port **102** should be connected to the appropriate neighbor device required by that topology.

[00049] In a further embodiment, as shown in **Figure 5**, the gateway device **50** could also be used to bridge, switch or route between multiple direct interconnects. In this case, the direct interconnect ports **102** on the gateway device **50** will be divided between the different direct interconnects (shown in **Figure 5** as A and B). All of the devices in direct interconnect A would then be reachable from direct interconnect A through the gateway device **50** and vice versa. In the example provided, traffic from a device in Direct Interconnect A destined for a device in Direct Interconnect B would first traverse Direct Interconnect A to the direct interconnect port **102** shown as **A4** on gateway device **50**. Gateway device **50** would then forward this traffic through the direct interconnect port **102** shown as **B2** where it would then be forwarded through Direct Interconnect B to the destination node.

[00050] In yet another embodiment of the present invention, the gateways could be used as access switches and the direct interconnect would form the backbone (see **Figure 6**). In this case, all or the majority of the nodes within the direct interconnect would be composed of gateway ports, and all or the majority of the other devices would be connected directly to the traditional network ports **100** of gateway device **50**. In the example provided in **Figure 6**, traffic from Device group A destined for Device group B would first be forward to gateway device **50A**. The

gateway **50A** forwarding function would recognize that the destination device is reachable through the direct interconnect and would forward the traffic through one of gateway device **50A**'s direct interconnect ports **102**. The direct interconnect would then forward the traffic to one of the direct interconnect ports of gateway device **50B**. The gateway **50B** forwarding function would recognize that the destination device is reachable through one of its standard network ports and would forward the traffic appropriately.

[00051] In order to maximize I/O traffic efficiencies, **Figure 7** provides an example of how, in a single direct interconnect deployment, the direct interconnect ports **102** could be chosen to minimize the average distance (number of hops) from each gateway device **50** in the direct interconnect to the nearest gateway port. In a 4x4 2D torus with the nodes numbered as per **Figure 7,** having the gateway act as nodes 1, 6, 11 and 16 would minimize the distance to the other nodes in the direct interconnect. However, a person skilled in the art would also understand that if a subset of the nodes tends to generate a higher amount of I/O than the average, then the deployment of the gateway nodes could be biased to be closer to these higher I/O nodes. Multiple algorithms are known that can be used to determine the optimum location for these I/O nodes (see, for instance, Bae, M., Bose, B.: "Resource Placement in Torus-Based Networks" in: Proc. IEEE International Parallel Processing Symposium, pp. 327-331. IEEE Computer Society Press, Los Alamitos (1996); Dillow, David A. et al.: "I/O Congestion Avoidance via Routing and Object Placement" United States: N. p., 2011. Print. Proceedings of the Cray User Group conference (CUG 2011), Fairbanks, AK, USA; Almohammad, B., Bose, B.: "Resource Placements in 2D Tori" in: Proc. IPPS 1998 Proceedings of the 12th International Parallel Processing Symposium on International Parallel Processing Symposium, p. 431. IEEE Computer Society, Washington, DC (1998); Dillow, David A. et al.: "Enhancing I/O Throughput via Efficient Routing and Placement for Large-scale Parallel File Systems" in: Conference Proceedings of the IEEE International Performance, Computing, and Communications Conference (IPCCC), 2011 IEEE 30th International; Ezell, M. et al.: "I/O Router Placement and Fine-Grained Routing on Titan to Support Spider II" in: Proceedings of the Cray User Group Conference (CUG), May 2014; and Babatunde, A. et al.: "I/O Node Placement for Performance and Reliability in Torus Networks", Department of Computer Science, Texas A&M University, Jan 2006).

[00052] In a preferred embodiment, the direct interconnect ports **102** will act as standard direct interconnect ports and autonomously forward traffic remaining in the direct interconnect (i.e. forwarding FLITs). They will also recombine FLITs into network packets for traffic destined for devices not in the direct interconnect (see PCT Patent Application Publication No. WO 2015/120539 A1 for an optimal method to route packets in a distributed direct interconnect network). The gateway device **50** should preferably also have the capability to transmit/receive network packets to/from each of the traditional network ports **100** and direct interconnect ports **102**, and also be able to interpret and forward this traffic based on layer 2, 3 or as per the above.

[00053] In a preferred embodiment, standard northbound network management interfaces would be exposed (e.g. CLI, OpenFlow, SNMP, REST, etc.) to allow a network management system to manage the gateway device **50**.

[00054] In one embodiment, when multiple gateway ports are connected to the same direct interconnect, all packets from a given flow should preferably egress on the same gateway port to aid in guaranteeing in-order packet delivery.

[00055] The gateway device **50** should preferably be configured to aggregate MAC forwarding tables between the direct interconnect ports **102** connected to the same direct interconnect (i.e. when a direct interconnect port learns of a VLAN/MAC address/node_id tuple, this tuple should preferably be shared with the other direct interconnect ports **102** connected to the same direct interconnect).

[00056] In a preferred embodiment, when an ARP request is received at one or more of the direct interconnect ports **102** connected to the same direct interconnect, the decision of which direct interconnect port should respond should be coordinated by the gateway **50** to ensure only a single response is transmitted (see e.g. at **Figure 8**). This could be done, for example, by choosing the direct interconnect port closest to the source of the ARP request, the first direct interconnect port to receive the ARP, through a round-robin selection scheme, through a hash of some portion of the ARP request (IP address, source MAC, etc.), or another algorithm known to persons skilled in the art. **Figure 8b** shows one embodiment of this coordination mechanism wherein the coordination function is distributed across the direct interconnect ASICs **104** and each direct interconnect ASIC **104** is connected to coordination bus **112** used to communicate coordination information and decisions.

In another embodiment, this coordination function may be centralized in a dedicated coordination ASIC and each direct interconnect ASIC **104** is connected to the coordination ASIC via a dedicated or shared coordination bus **112**.

[00057] When more than one gateway is connected to the same torus, the gateway devices **50** should preferably coordinate their knowledge of the torus topology and response to ARP requests in a similar fashion to the single gateway case discussed above (see **Figure 9** for a logic tree re such coordination). In one embodiment, the gateways could discover each other through a broadcast protocol. In another embodiment, the gateways could be configured to know the location of the other gateways. In yet another embodiment, a consensus algorithm such as Raft could be used to ensure coordination of the gateways and consistency of the traffic – gateway/port associations (see e.g. Ongaro, D., Ousterhout, J.: "In Search of an Understandable Consensus Algorithm (Extended Version)" in: 2014 USENIX Annual Technical Conference, June 19, 2014, Philadelphia, PA.; Woos, D. et al.: "Planning for Change in a Formal Verification of the Raft Consensus Protocol" in: Certified Programs and Proofs (CPP), Jan. 2016).

[00058] In general, whenever a torus node would like to communicate with a resource that is accessible through one or more gateway devices **50**, the gateway(s) should preferably coordinate which gateway port is chosen to provide access to that resource in a similar manner to the ARP example described above. Examples of this include anycast, broadcast and multicast traffic, node and service discovery protocols and IPv6 neighbor discovery.

[00059] As a further consideration, it is important to note that, in many cases, non-minimal routing is used within a direct interconnect. Since the gateway(s) within a direct interconnect have gateway ports in multiple locations within the topology, it is possible for traffic destined for one gateway port to traverse one of the other gateway ports first. It would therefore be preferable to increase efficiencies by having a single, first gateway port process the traffic instead of allowing the traffic to traverse to a more distant gateway port. An example of this is provided in **Figure 10**, where node 5 is sending traffic to node 1, but due to non-minimal routing, the packet will traverse node 6. In one embodiment, to improve efficiencies, the gateway with the port located at node 6 will recognize that the packet is destined for another gateway port and will process the packet as if it is destined for node 6 instead of forwarding it through the direct interconnect via node 2.

[00060] As noted above, the Direct Interconnect ASIC **104** provides connectivity between Switch ASIC **106** and a direct interconnect. In order to ensure that a person skilled in the art would be able to make and work a network gateway device of the present invention, **Figures 11** and **12** describe the operation of a preferred embodiment of Direct Interconnect ASIC **104**. This operation is also applicable, with any modifications as necessary, for other ASICs within the scope of this invention, including ASICs **105**, **107** and **108**, as appropriate.

[00061] As is well-known in the art, Switch ASIC **106** transmits and receives Ethernet frames. **Figure 11** describes how Ethernet Frame **200, generated by Switch ASIC** 106, is processed by Direct Interconnect ASIC **104**. MAC Address Database **201** contains a list of Ethernet MAC addresses and the direct interconnect nodes associated with each Ethernet MAC address. The Ethernet Frame **200** received from Switch ASIC **106** is examined and the Source and Destination MAC addresses are retrieved from its header. These MAC Addresses are used as indices to MAC Address Database **201**. The source MAC address of the Ethernet Frame **301** is combined with the node number of the current to create or update an association between this node number and source MAC address in MAC Address Database **201**.

[00062] If the Destination MAC address of Ethernet Frame **200** is in MAC Address Database **201**, then the Node Number **202** associated with this MAC Address is retrieved from Mac Address Database **201**. The Node Number **202** is then used as an index into Source Route Database **206** and source route **203** associated with Node Number **202** is retrieved from the Source Route Database **206**. As is well-known in the art, source route databases contain a list of network destinations and one or more paths through the network to reach each destination. A source route database may be manually populated or may rely on well-known automated topology discovery and route computation algorithms. Ethernet Frame **200** is then converted into FLITs **204**. A FLIT is a specialized frame type used in direct interconnects and can be of either fixed or variable size. In a preferred embodiment, FLITs will be of a fixed size. In another embodiment, FLITs will be of a variable size within a minimum and maximum size. In yet another embodiment, FLITs will be sized so that the Ethernet Frame **200** exactly fits into the FLIT payload.

[00063] If the Ethernet Frame **200** is larger than the payload of a single FLIT, multiple FLITs **204** will be created. If the Ethernet Frame **200** fits into the payload of

a single FLIT, then a single FLIT will be created. Source Route **203** is then inserted into the header of the first of the FLITs **204** along with the node number of the current node. FLITs **204** are then transmitted from the Egress Port **205** specified in Source Route **203.**

[00064] If the Destination MAC address of Ethernet Frame **200** is not in MAC Address Database **201** or if the Destination MAC Address of Ethernet Frame **200** indicates that it is a broadcast Ethernet packet, then Ethernet Frame **200** is converted into FLITs **204** as in the case described above although a source route will not be included. Once FLITs **204** have been created, a flag in the header of the first FLIT is set to indicate that these FLITs should be broadcast to every node in the direct interconnect. A time-to-live (TTL) value is also set in the header of the first FLIT. The TTL determines the maximum number of times broadcast FLITs can be forwarded through the direct interconnect. In one embodiment, anycast and multicast Ethernet frames are treated as if they are broadcast frames, as above.

[00065] **Figure 12** describes how FLITs **204** are processed by Direct Interconnect ASIC **104** in a preferred embodiment of Direct Interconnect ASIC **104**. FLITs **204** from Direct Interconnect **301** are received by Direct Interconnect ASIC **104** and the header of the first of these FLITs is examined to see if the broadcast flag is set. If the broadcast flag is not set, Source Route **203** is retrieved from the first FLIT and it is determined whether the source route indicates that the current node is the destination node for the FLITs **204**. In another embodiment, the FLITs contain the node number of the destination node and it is the node number that is used to determine if the current node is the destination node for the FLITs **204**. If the current node is the destination node for the FLITs **204**, the FLITs **204** are combined to form Ethernet Frame **301**. The source MAC address of the Ethernet Frame **301** is combined with the node number in the header of the first FLIT to create or update an association between said node number and source MAC address in MAC Address Database **201**. Ethernet frame **301** is then transmitted to Switch ASIC **106.**

[00066] If it is determined that this is not the destination node, the source route is used to determine the egress port **302** for FLITs **204**. FLITs **204** are then transmitted out the egress port **302**.

[00067] If the broadcast flag is set in the first FLIT header, the FLITs **204** are combined to form Ethernet Frame **301**. The source MAC address of the Ethernet Frame **301** is combined with the node number in the header of the first FLIT to

create or update an association between said node number and source MAC address in MAC Address Database **301**. Ethernet Frame **301** is transmitted to Switch ASIC **106**.

[00068] The TTL in the header of the first FLIT is then decremented by one. If the TTL is now equal to zero, then FLITs **204** are discarded. If the TTL is greater than zero, the FLITs **204** are transmitted out all egress ports except for the ingress port from which FLITs **204** were originally received.

[00069] In other embodiments of Direct Interconnect ASIC **104,** source routing may not be used. In one embodiment, the destination MAC address of Ethernet Frame **200** will be used by each node to perform a local next-hop route lookup. In another embodiment, destination node information in the FLIT header will be used by each node to perform a local next-hop route lookup.

[00070] It will be obvious to those well-versed in the art that other embodiments of Direct Interconnect **104** and Switch ASIC **106** may be designed to work with protocols other than Ethernet. In one embodiment, these elements will be designed to work with Gen-Z. In this case, Direct Interconnect **104** would expect to received Gen-Z Core64 packets instead of Ethernet frames. Instead of Ethernet MAC addresses, Gen-Z GCIDs (Global Component IDs) would be used and associated with direct interconnect node numbers.

[00071] Although specific embodiments of the invention have been described, it will be apparent to one skilled in the art that variations and modifications to the embodiments may be made within the scope of the following claims.

CLAIMS:

We claim:

1. A dedicated network gateway device (50) that is capable of bridging, switching or routing network traffic between traditional and direct interconnect networks, comprising:

a first set of at least one traditional network ports (100) with a single link per port and

a second set of at least one direct interconnect ports (102) with at least two links per port;

wherein the first set of traditional network ports (100) are connected to one of switches and devices that form a traditional network; and

wherein the second set of at least one direct interconnect ports (102) are connected to a direct interconnect network.

2. The device (50) of claim 1 wherein the first set of traditional network ports (100) comprise at least one of SFP+, QSFP, and QSFP+ connectors.

3. The device (50) of claim 1 wherein the second set of at least one direct interconnect ports (102) comprise at least one of MXC, MTP, and MTO connectors.

4. The device (50) of claim 1 wherein the traditional network ports (100) are connected to at least one of a switch port and router port in the traditional network.

5. The device (50) of claim 1 wherein the second set of direct interconnect ports (102) are connected to a passive patch panel/hub (60) used in the implementation of the direct interconnect network.

6. The device (50) of claim1 wherein the second set of direct interconnect ports (102) are each connected to at least one of their own dedicated direct interconnect application-specific integrated circuit (ASIC) 104, field-

programmable gate array (FPGA), general purpose processor, network processor, and device capable of acting as a node in a direct interconnect network.

7. The device (50) of claim 1 wherein the second set of direct interconnect ports (102) are each connected to at least one of shared application-specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), general purpose processors, network processors, and devices capable of acting as a node in a direct interconnect network.

8. The device (50) of claim 1 wherein at least one of the bridging, switching and routing function is performed by at least one of network switch ASIC, field-programmable gate array (FPGA), general purpose processor, network processor, and a device capable of performing network traffic forwarding.

9. The device (50) of claim 1 wherein at least one of the bridging, switching and routing function is performed by at least one of a network controller ASIC, field-programmable gate array (FPGA), general purpose processor, network processor, and device capable of performing network traffic forwarding.

10. The device (50) of claim 9 wherein at least one of the ASICs, FPGAs, general purpose processors, network processors, and nodes are capable of acting as a direct interconnect node with locally destined/sourced traffic sent over a traditional network interface line.

11. The device (50) of claim 1 wherein the direct interconnect ports take the place of a device within the direct interconnect network.

12. The device (50) of claim 1 wherein the second set of direct interconnect ports are connected to multiple direct interconnect networks.

13. A dedicated network gateway device (50) capable of at least one of bridging, switching and routing network traffic between traditional and direct interconnect networks comprising:

a first port that is a direct interconnect port configured to be connected to a direct interconnect network, and

a second port that is a standard network port configured to be connected to at least one of switches and devices that form a traditional network.

14. The device (50) of claim 13, wherein the first port comprises at least one of a MXC, MTP, and MTO connector.

15. The device (50) of claim 13, wherein the second port comprises at least one of a SFP+, QSFP, and QSFP+ connector.

16. The device (50) of claim 13 wherein the first port is connected to a passive patch panel/hub used in the implementation of the direct interconnect network.

17. The device (50) of claim 13 wherein the second port is connected to at least one of switch and router ports in the traditional network.

18. A dedicated network gateway device (50) for at least one of bridging and routing network traffic between a traditional network and a direct interconnect network, comprising:

a first set of traditional network ports (100) with a single link per port, such ports being connected to end devices that form a first traditional network; and

a second set of direct interconnect ports (102) with two or more links per port, such ports being connected to the direct interconnect network,

wherein said direct interconnect network acts as a backbone that allows network traffic to route from the dedicated network gateway device (50) to another dedicated network gateway device (50), said another dedicated network gateway device (50) comprising:

a first set of traditional network ports (100) with a single link per port, such ports being connected to end devices that form a second traditional network; and

a second set of direct interconnect ports (102) with two or more links per port, such ports being connected to the direct interconnect network.

19. A dedicated network gateway device (50) for at least one of bridging, switching and routing network traffic between traditional network and direct interconnect networks, comprising:

a first set of at least one traditional network ports (100) with a single link per port, such ports being connected to at least one of switches and devices that form a traditional network;

a second set of at least one direct interconnect ports (102) with one or more links per port, such ports being connected to a direct interconnect network; and

a plurality of direct interconnect ports that are logically associated to act as a single direct interconnect node.

20. A computer-implemented method of at least one of bridging, switching and routing network traffic between a traditional network and a direct interconnect network, comprising the steps of:

connecting the dedicated gateway device (50) of claim 1 to the traditional network and the direct interconnect network, said dedicated gateway device (50) acting as one or more nodes within the direct interconnect network; and

forwarding the network traffic by means of the gateway device (50) between the traditional network and the direct interconnect network based on headers or content of the network traffic.

21. A computer-implemented method of coordinating which gateway device (50) having a capability to provide access to a resource located in a traditional network when said resource is accessible by more than one gateway device (50), the method comprising the steps of:

receiving an ARP, broadcast, multicast, or anycast traffic at a direct interconnect port, wherein said traffic is requesting access to the resource located in the traditional network, and wherein said direct interconnect port is linked via one or more hops to the more than one gateway devices (50), each of which is capable of providing access to the resource;

calculating an optimum gateway device port out of the more than one gateway devices that should provide access to the resource;

creating an association between the traffic, the direct interconnect node, and the calculated optimum gateway device port; and

communicating the association with each of the more than one gateway devices to ensure that the calculated optimum gateway device port provides access to the resource.

22. The computer-implemented method of claim 21, wherein the step of calculating the optimum gateway device port that should provide access to the resource comprises determining which of the more than one gateway device ports is closest to the direct interconnect port.

23. The computer-implemented method of claim 21, wherein the step of calculating the optimum gateway device port that should provide access to the resource comprises employing a consensus algorithm to ensure consistency of the traffic.

24. The computer-implemented method of claim 21, wherein the step of communicating the association is handled by a dedicated or shared coordination bus.

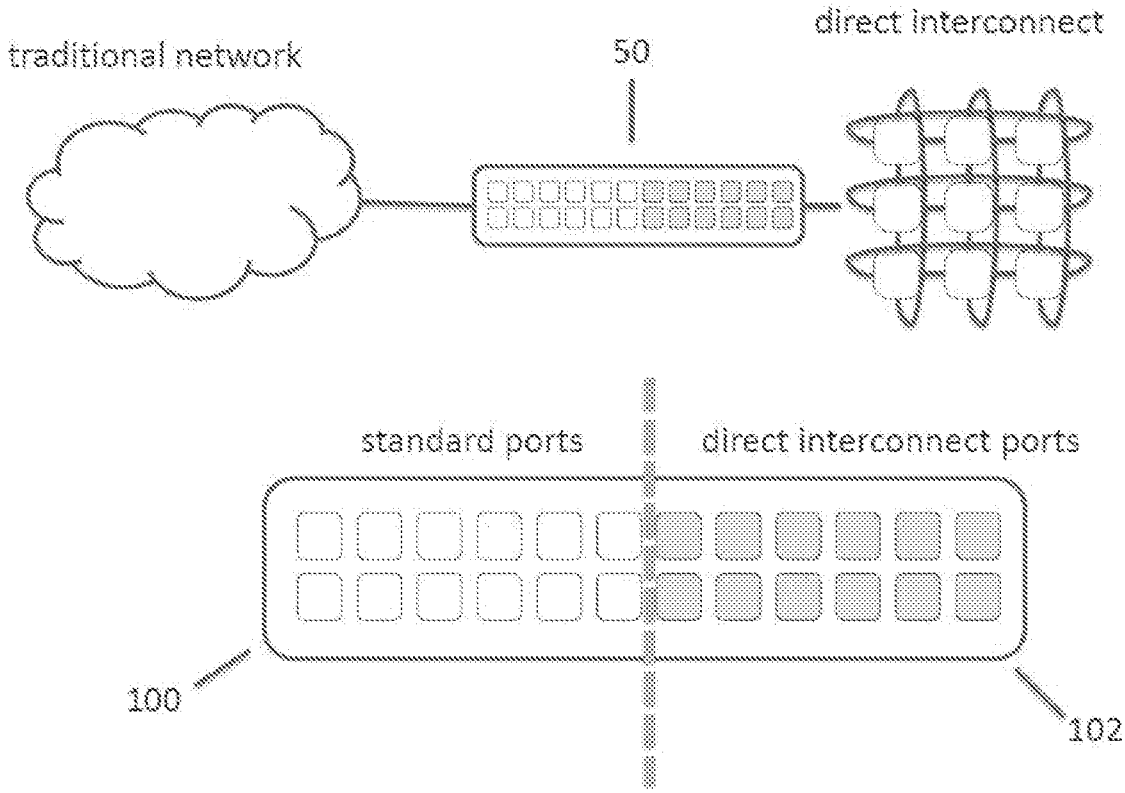
traditional network
50
direct interconnect
standard ports
direct interconnect ports
100
102

FIGURE 1

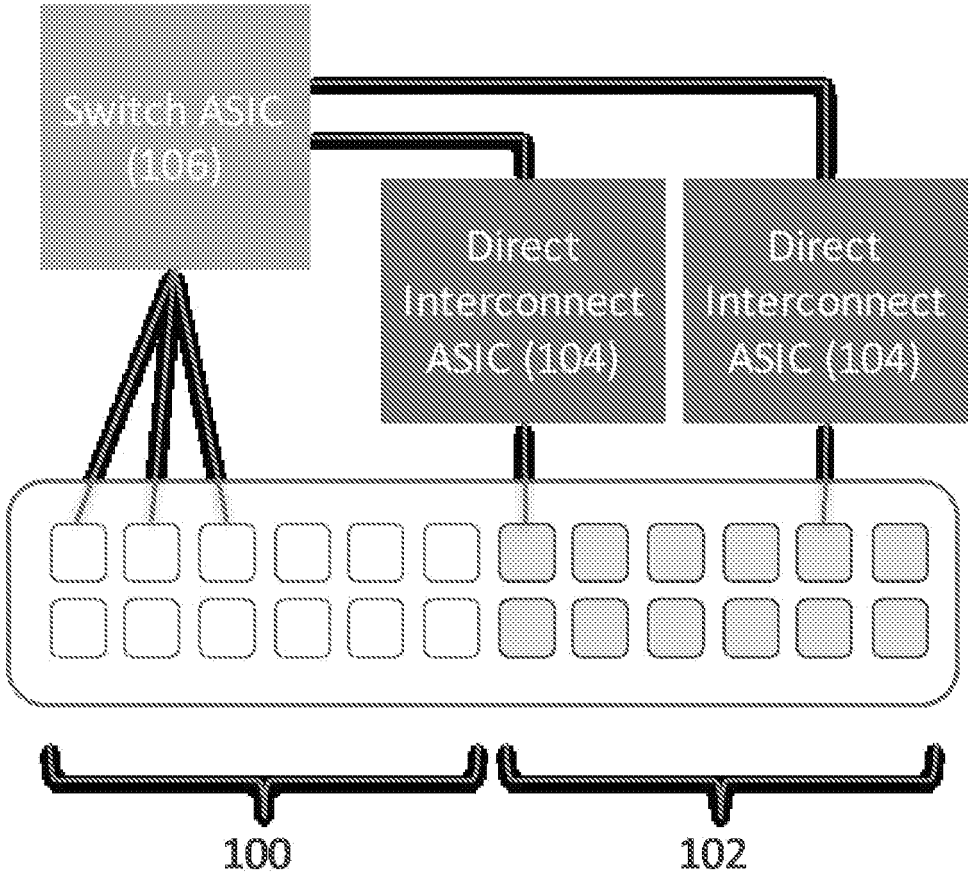
Switch ASIC (106)
Direct Interconnect ASIC (104)
Direct Interconnect ASIC (104)
100
102

FIGURE 2

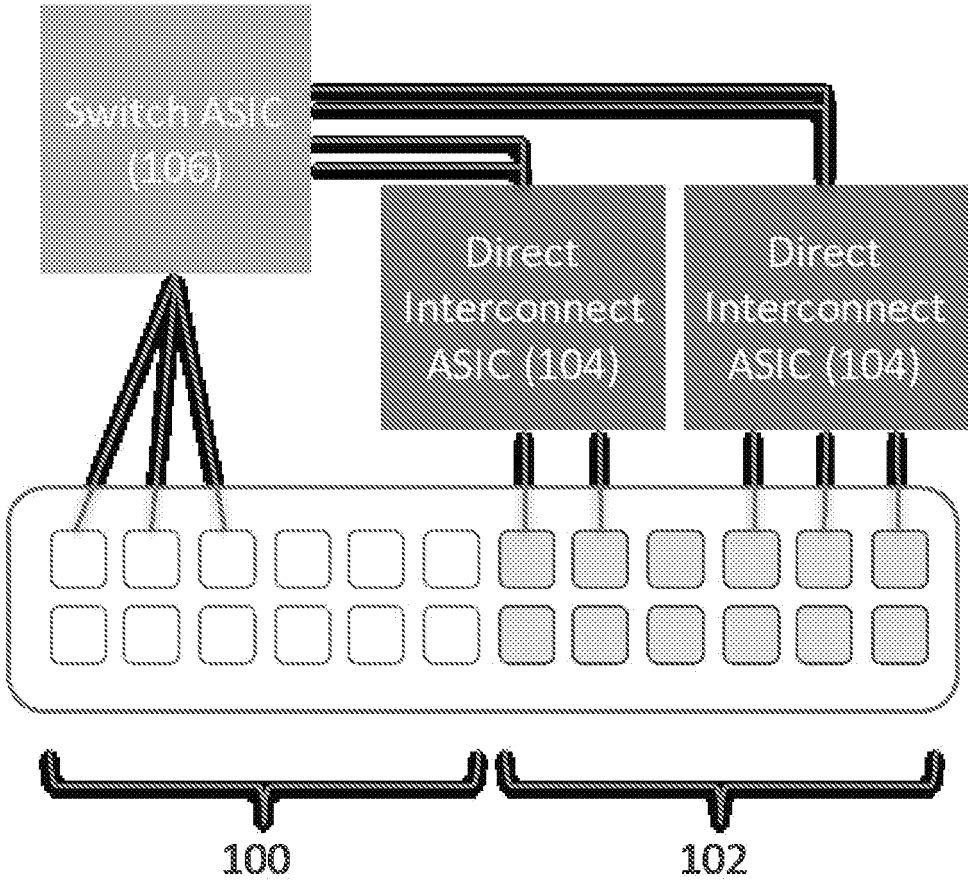
Switch ASIC (106)
Direct Interconnect ASIC (104)
Direct Interconnect ASIC (104)
100
102

FIGURE 2b

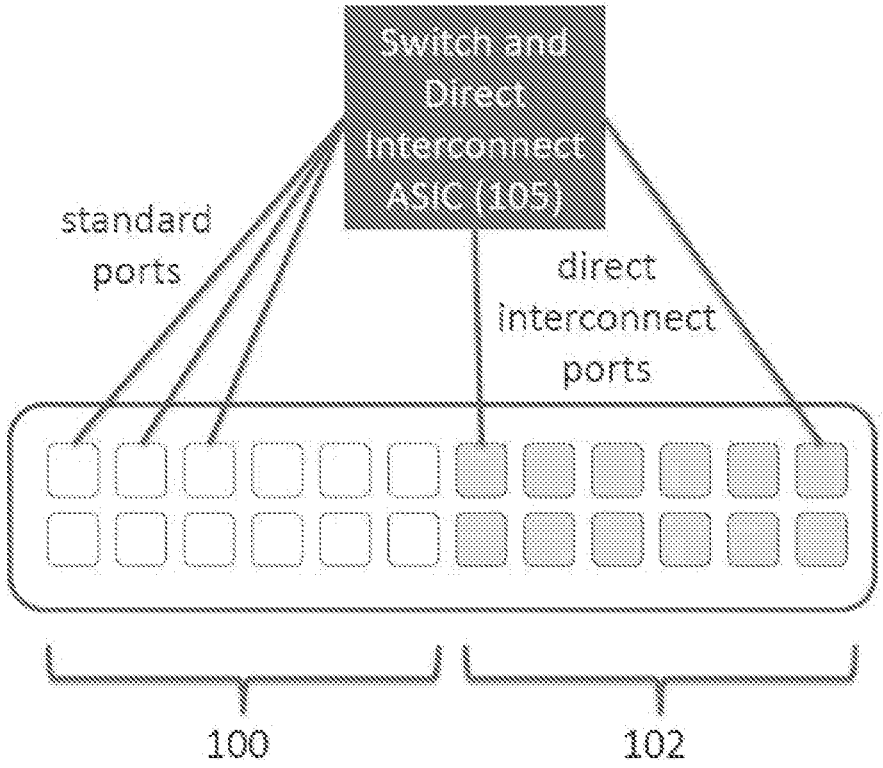
Switch and Direct Interconnect ASIC (105)
standard ports
direct interconnect ports
100
102

FIGURE 2c

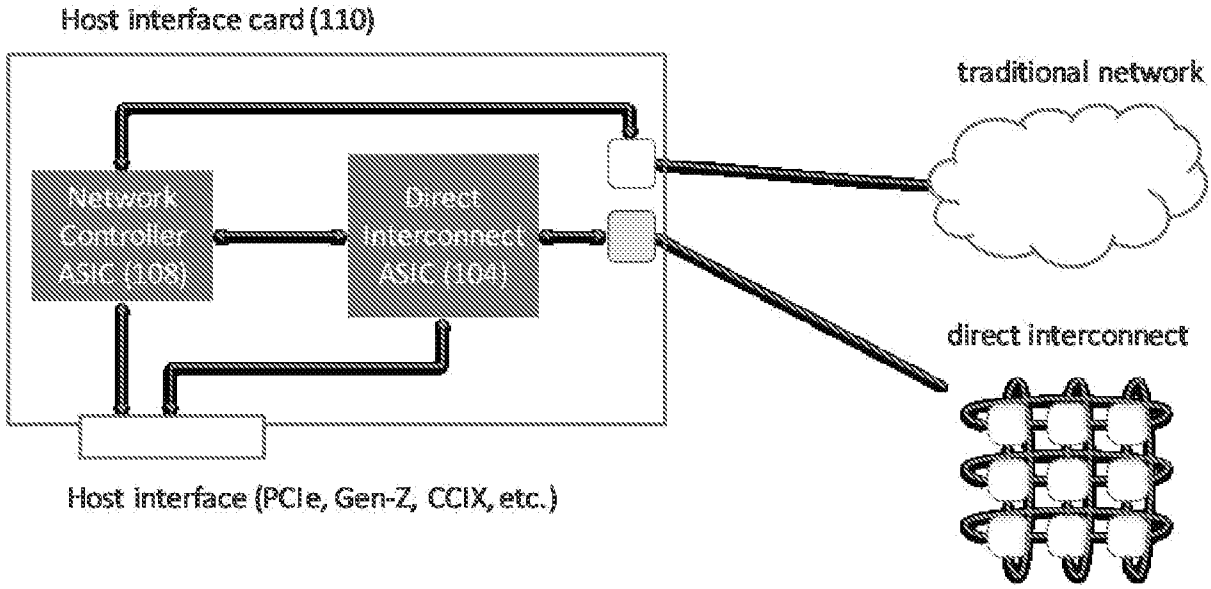
Host interface card (110)
traditional network
Network Controller ASIC (103)
Direct Interconnect ASIC (104)
direct interconnect
Host interface (PCIe, Gen-Z, CCIX, etc.)

FIGURE 2d

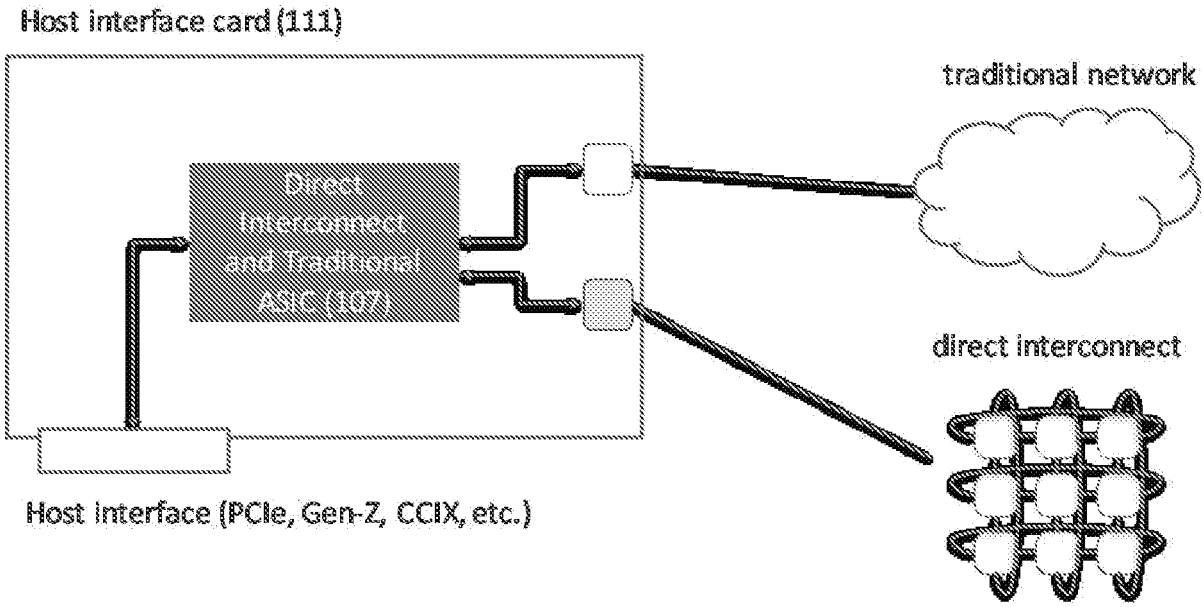
Host interface card (111)
traditional network
Direct Interconnect and Traditional ASIC (107)
direct interconnect
Host interface (PCIe, Gen-Z, CCIX, etc.)

FIGURE 2e

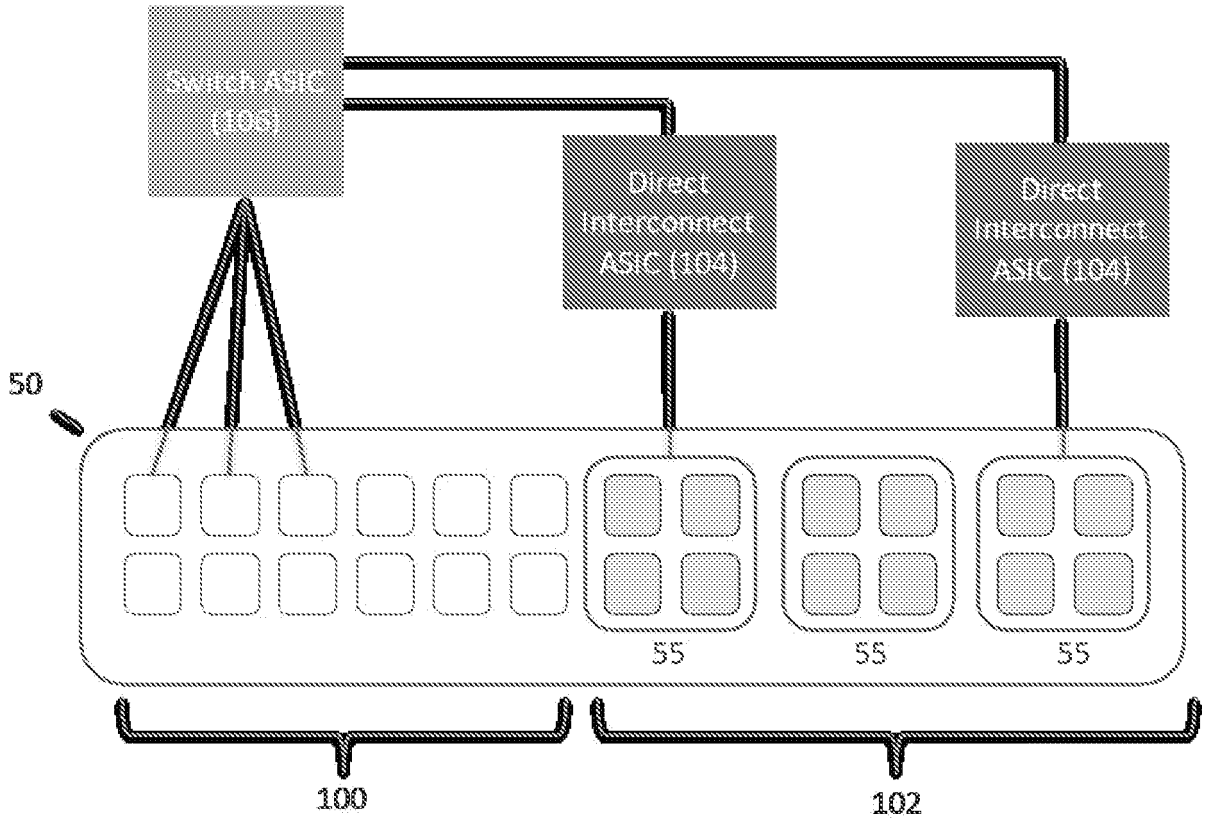
Switch ASIC (105)
Direct Interconnect ASIC (104)
Direct Interconnect ASIC (104)
50
55
55
55
100
102

FIGURE 2f

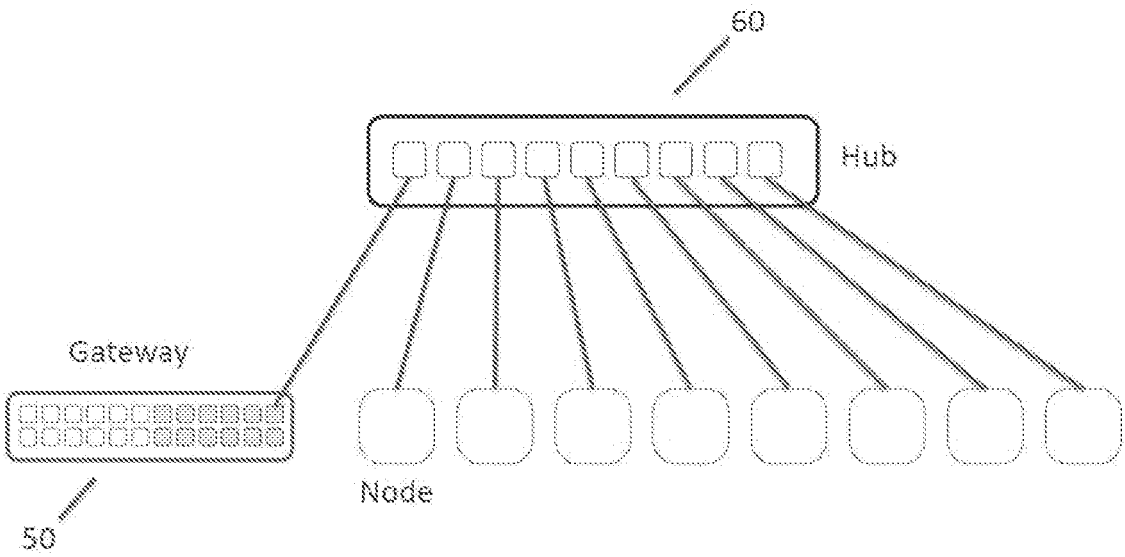

FIGURE 3

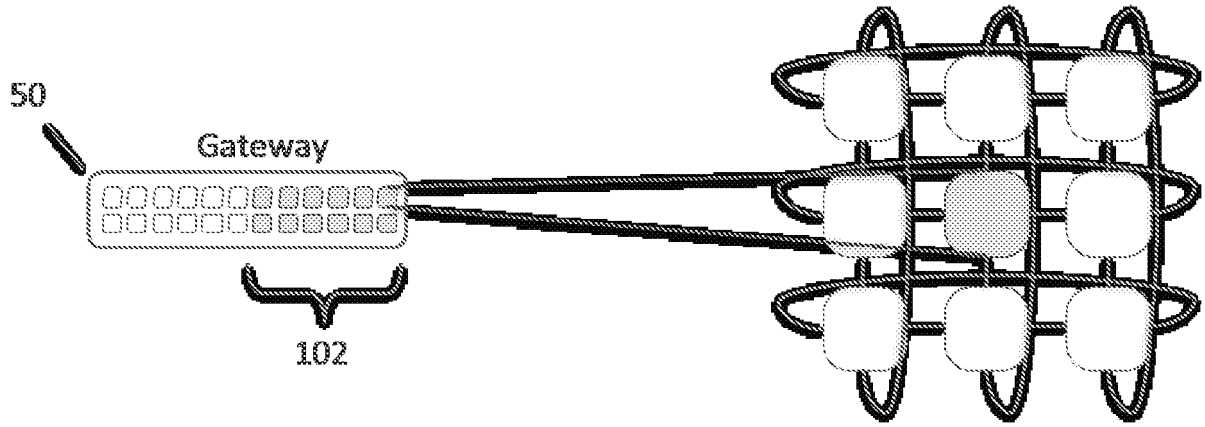

FIGURE 4

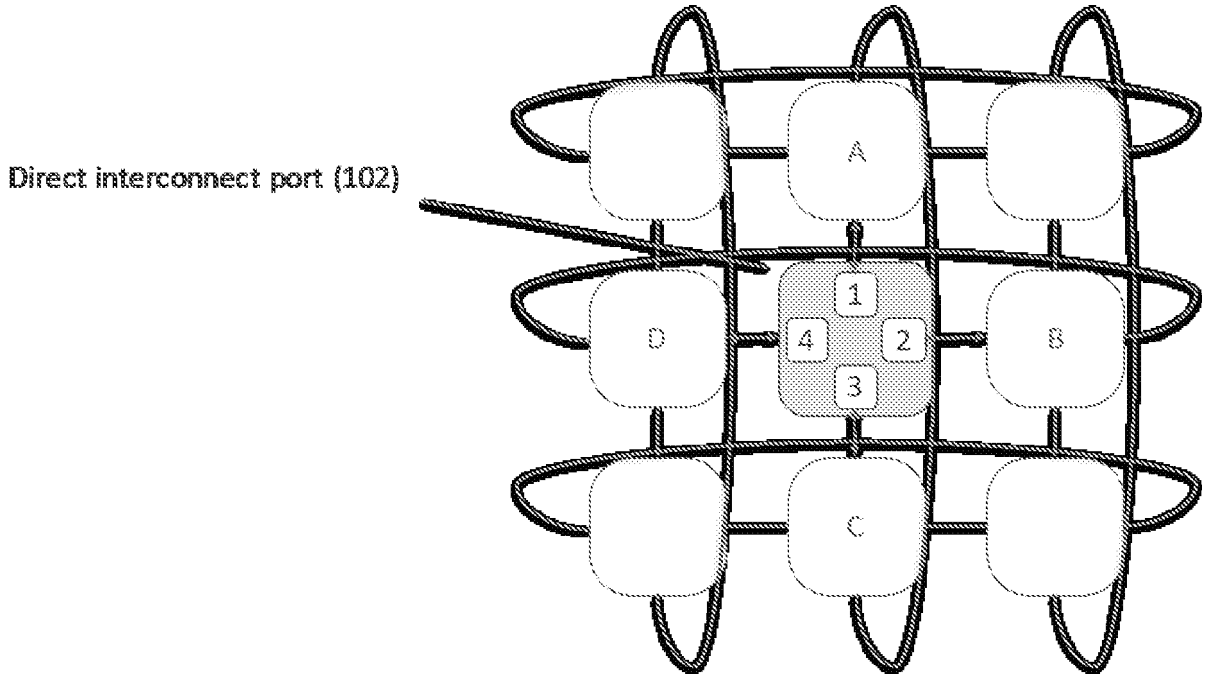

FIGURE 4b

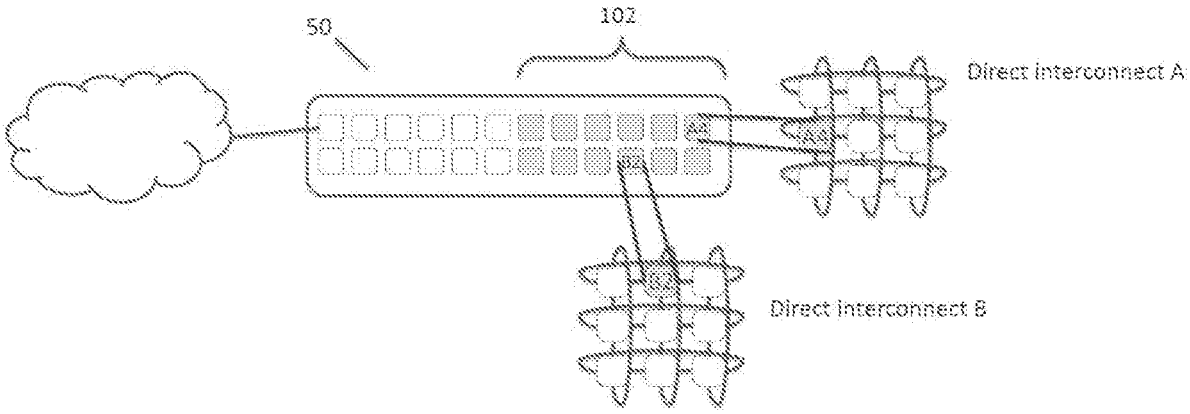

FIGURE 5

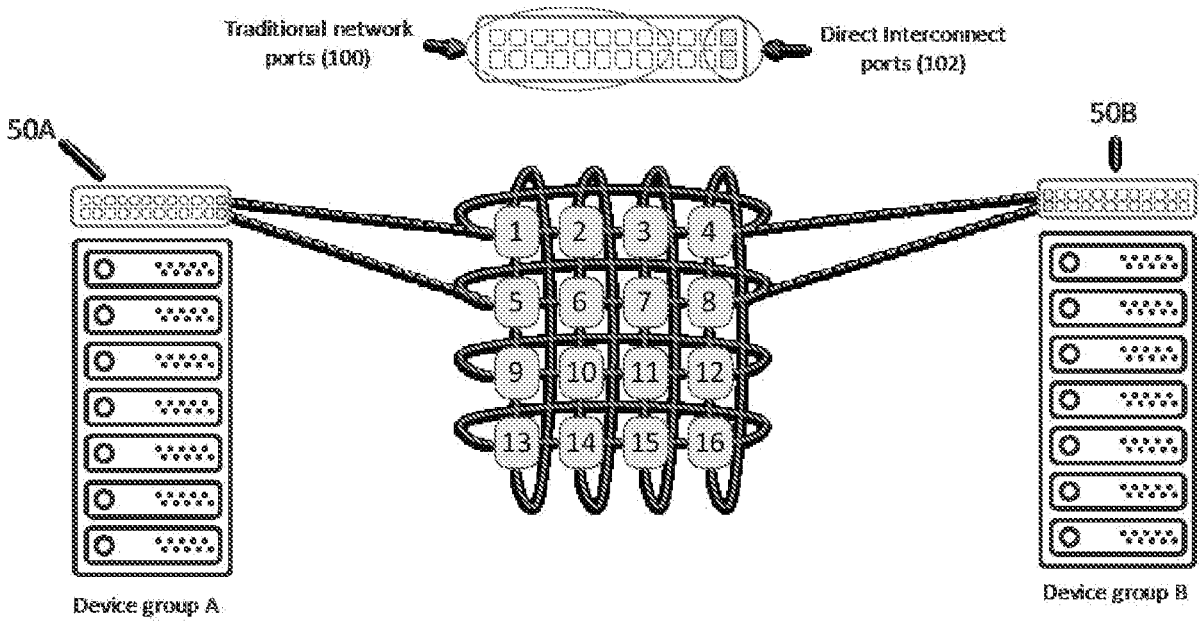
Traditional network ports (100)
Direct Interconnect ports (102)
50A
50B
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
Device group A
Device group B

FIGURE 6

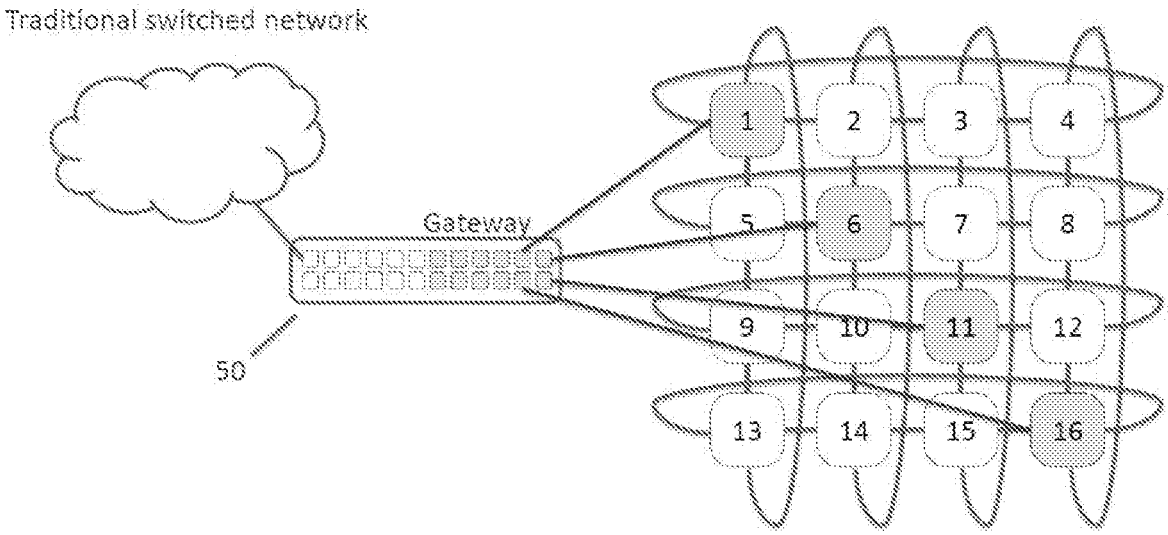
Traditional switched network
Gateway
50
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

FIGURE 7

- Nodes 1, 6, 11 and 16 are ports on the same gateway device
- Node 5 sends broadcast packet
- Nodes 1, 6, 11 and 16 receive broadcast packet
- Gateway determines that nodes 1 and 6 are the closest nodes to node 5
- Gateway chooses node 1 as optimum node
- Node 1 is associated with this traffic and will handle future traffic related to the broadcast packet

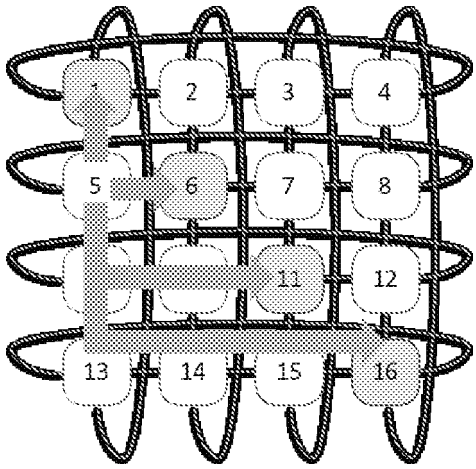

FIGURE 8

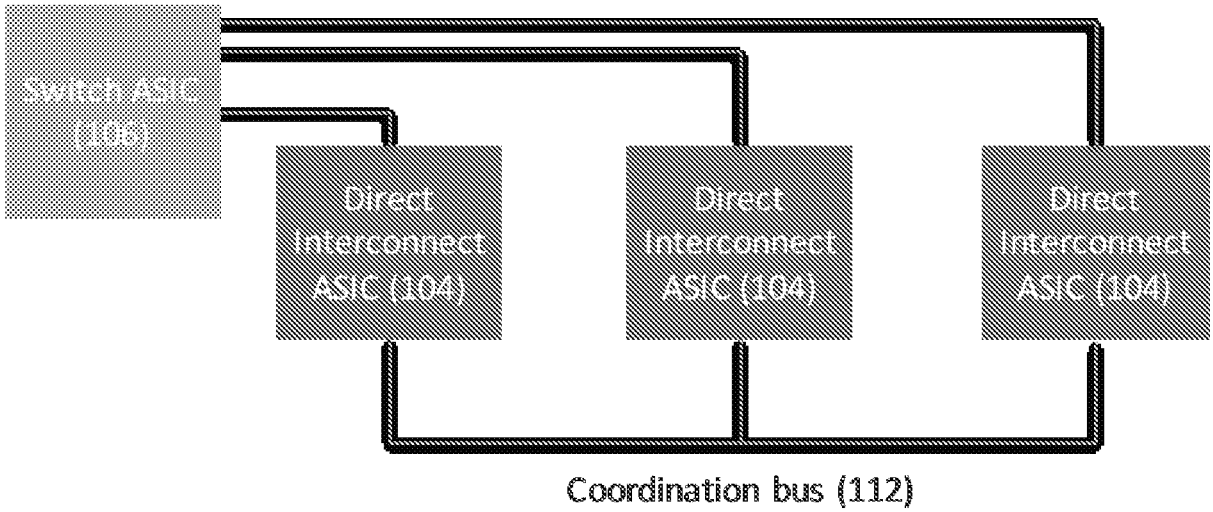
Switch ASIC (108)
Direct Interconnect ASIC (104)
Direct Interconnect ASIC (104)
Direct Interconnect ASIC (104)
Coordination bus (112)

FIGURE 8b

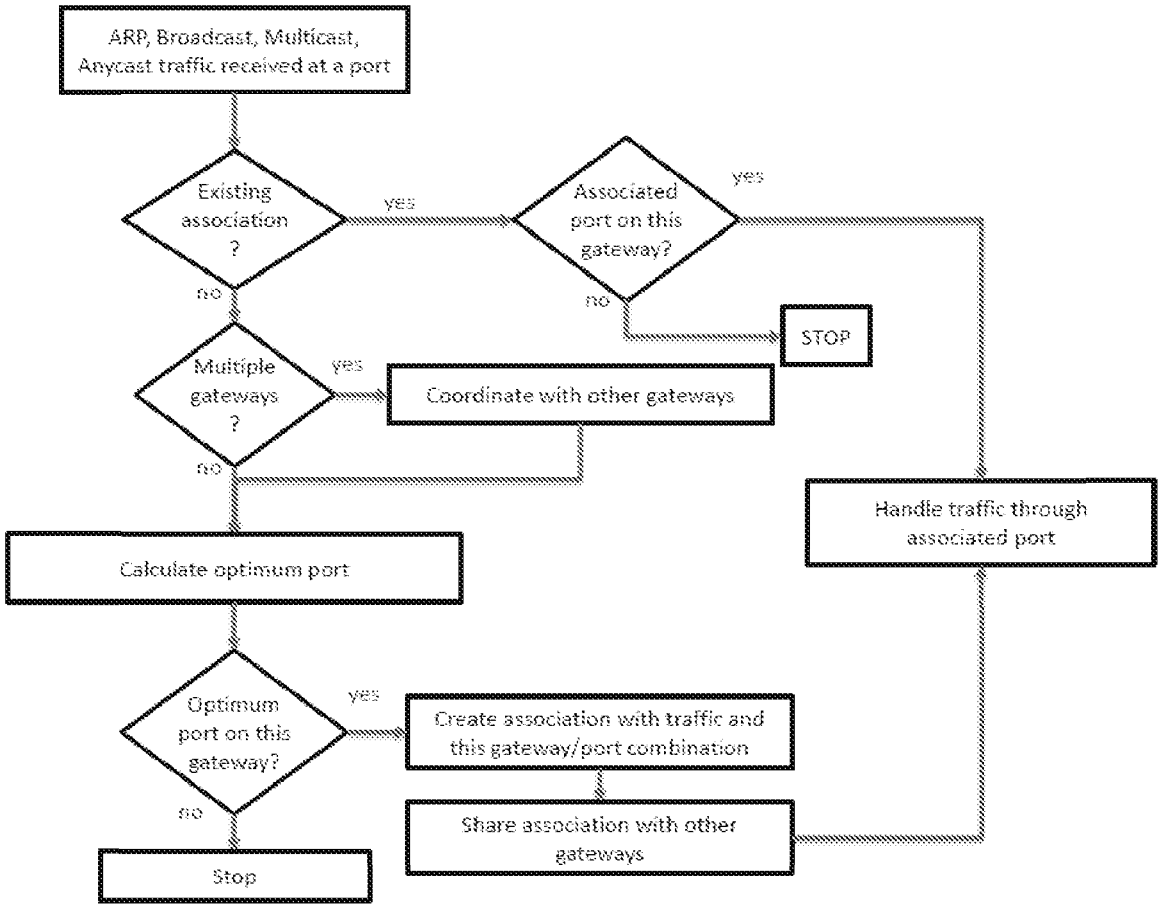
ARP, Broadcast, Multicast, Anycast traffic received at a port
Existing association ?
yes
Associated port on this gateway?
yes
no
STOP
no
Multiple gateways ?
yes
Coordinate with other gateways
no
Handle traffic through associated port
Calculate optimum port
Optimum port on this gateway?
yes
Create association with traffic and this gateway/port combination
no
Share association with other gateways
Stop

FIGURE 9

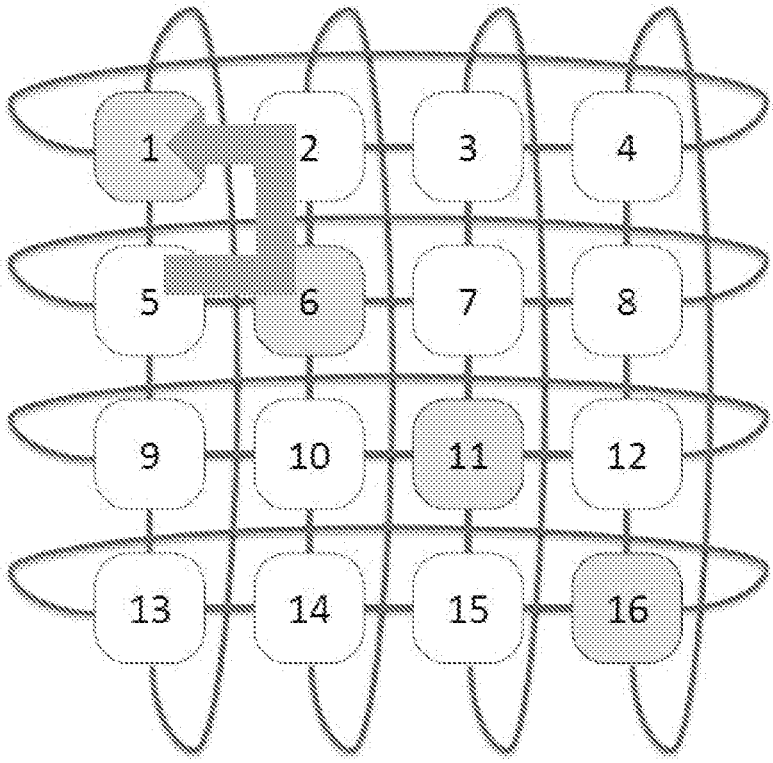

FIGURE 10

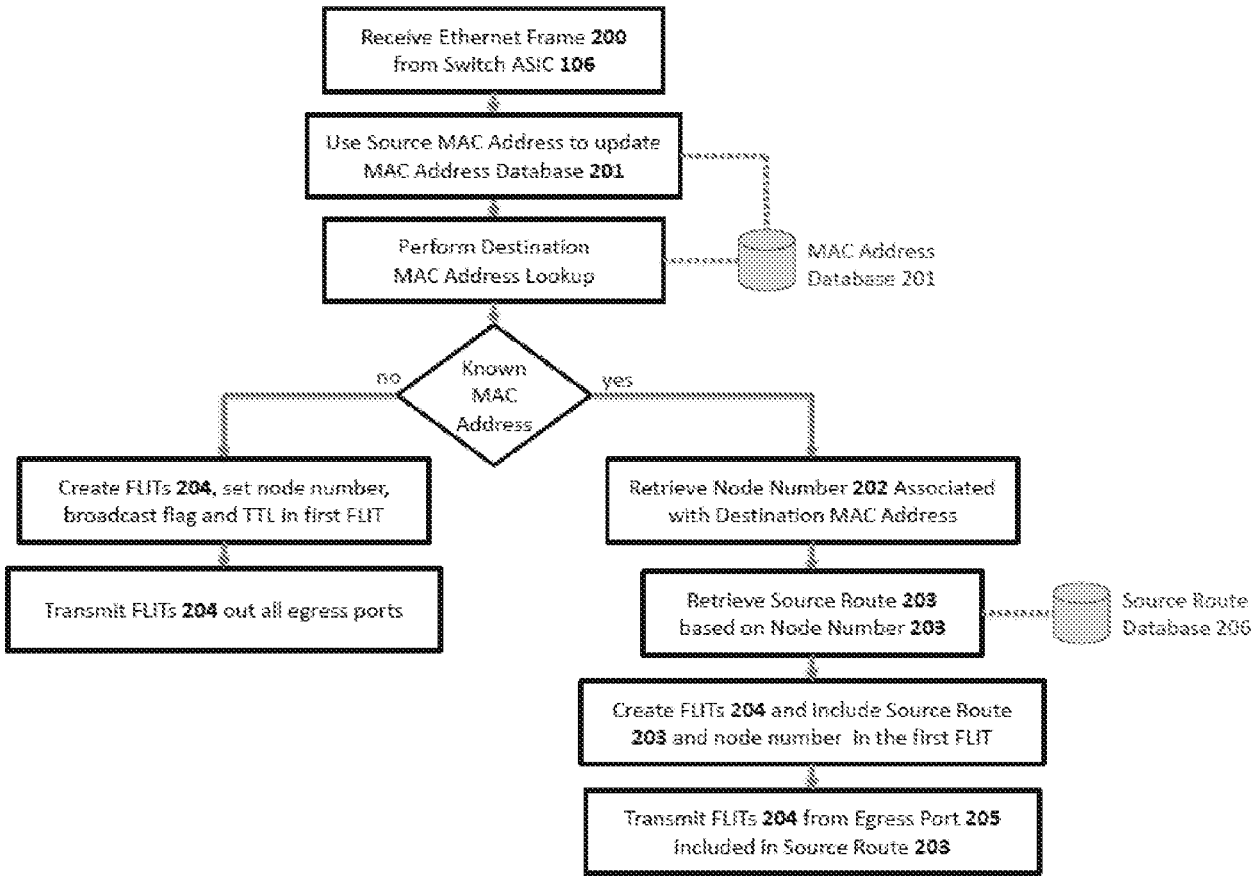
Receive Ethernet Frame 200 from Switch ASIC 106
Use Source MAC Address to update MAC Address Database 201
Perform Destination MAC Address Lookup
MAC Address Database 201
Known MAC Address
no
yes
Create FLITs 204, set node number, broadcast flag and TTL in first FLIT
Transmit FLITs 204 out all egress ports
Retrieve Node Number 202 Associated with Destination MAC Address
Retrieve Source Route 203 based on Node Number 203
Source Route Database 206
Create FLITs 204 and include Source Route 203 and node number in the first FLIT
Transmit FLITs 204 from Egress Port 205 included in Source Route 203

FIGURE 11

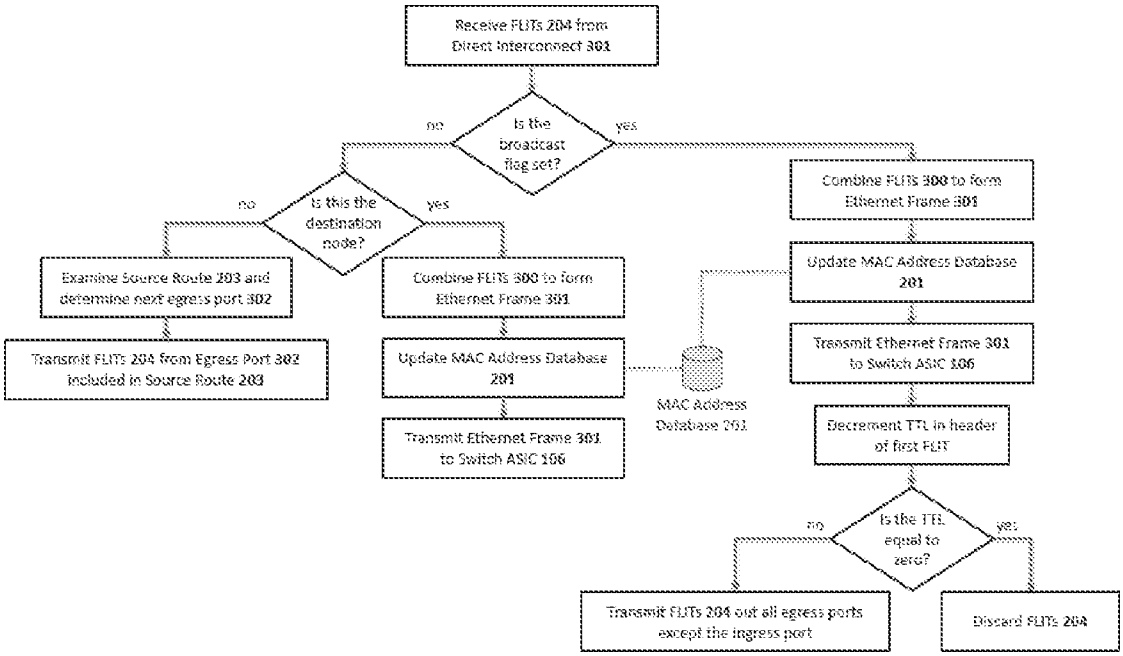
Receive FLITs 204 from Direct Interconnect 301
Is the broadcast flag set?
no
yes
Is this the destination node?
no
yes
Examine Source Route 203 and determine next egress port 302
Transmit FLITs 204 from Egress Port 302 included in Source Route 203
Combine FLITs 300 to form Ethernet Frame 301
Update MAC Address Database 201
Transmit Ethernet Frame 301 to Switch ASIC 156
MAC Address Database 201
Combine FLITs 300 to form Ethernet Frame 301
Update MAC Address Database 201
Transmit Ethernet Frame 301 to Switch ASIC 156
Decrement TTL in header of first FLIT
Is the TTL equal to zero?
no
yes
Transmit FLITs 204 out all egress ports except the ingress port
Discard FLITs 204

FIGURE 12