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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
1 October 2015 (01.10.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/147840 A1

(51) International Patent Classification:
H04L 12/931 (2013.01)

(21) International Application Number:
PCT/US2014/032066

(22) International Filing Date:
27 March 2014 (27.03.2014)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.** [US/US]; 11445 Compaq Center Drive W., Houston, Texas 77070 (US).

(72) Inventors: **MORRIS, Terrel**; 714 S. 9th Street, Garland, Texas 75040 (US). **PURCELL, Brian T.**; 11445 Compaq Center Dr W, Houston, Texas 77070 (US). **CHALMERS, F. Steven**; 8000 Foothills Blvd., Roseville, California 95747 (US). **HANDGEN, Erin A.**; 3404 E Harmony Rd., Ft. Collins, Colorado 80528-9544 (US). **GOODRUM, Alan L.**; 11445 Compaq Center Dr W, Houston, Texas 77070 (US).

(74) Agents: **KINCAID, David K.** et al.; Hewlett-Packard Company, Intellectual Property Administration, 3404 E. Harmony Road, Mail Stop 35, Fort Collins, Colorado 80528 (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, IR, IS, JP, KE, KG, 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, 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, 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))*

[Continued on next page]

(54) Title: MODULAR INPUT/OUTPUT AGGREGATION ZONE

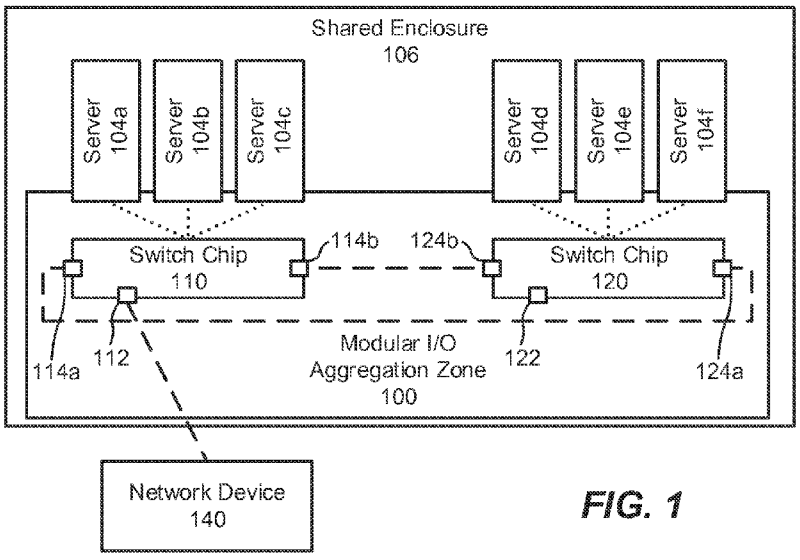

FIG. 1

(57) Abstract: In one example implementation according to aspects of the present disclosure, a system is disclosed having a modular input/output aggregation zone to directly communicatively couple together a plurality of servers within an enclosure shared by the modular input/output aggregation zone and the plurality of servers. The example modular input/output aggregation zone includes a first switch chip having link ports configurable as uplink ports and crosslink ports, the uplink ports being communicatively couple-able to a network device and the crosslink ports being communicatively coupleable to a crosslink port of a second switch chip.

WO 2015/147840 A1

Published:

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

MODULAR INPUT/OUTPUT AGGREGATION ZONE

BACKGROUND

[0001] The amount and size of electronic data consumers and companies generate and use continues to grow in size and complexity, as does the size and complexity of related applications. In response, data centers housing the growing and complex data and related applications have begun to implement a variety of networking and server configurations to provide access to the data and applications.

BRIEF DESCRIPTION OF THE DRAWINGS

[0002] The following detailed description references the drawings, in which:

[0003] FIG. 1 illustrates a block diagram of a system utilizing a modular input/output aggregation zone having a switch chip according to examples of the present disclosure;

[0004] FIG. 2 illustrates a block diagram of a system utilizing two modular input/output aggregation zones each having a switch chip according to examples of the present disclosure;

[0005] FIG. 3 illustrates a block diagram of a modular input/output aggregation zone having a first switch chip and a second switch chip according to examples of the present disclosure; and

[0006] FIG. 4 illustrates a block diagram of a modular input/output aggregation zone having a first switch chip, a second switch chip, and a third switch chip according to examples of the present disclosure.

DETAILED DESCRIPTION

[0007] Data centers store growing amounts of data and host increasingly complex applications. The data and applications may be distributed across numerous servers networked together in a traditional hierarchical network topology. Server application architecture, particularly those employing heavy use of virtualization technology and data spread across multiple scale-out servers, may not be well-served by traditional hierarchical network switching topologies.

Problems associated with a traditional hierarchical network approach may include cost, latency, and management complexity.

[0008] Cost is typically measured in terms of cost per connected server. Thus each layer of networking adds to the total solution cost, affecting the cost per connected server. This situation is particularly aggravated by high-density, low-cost servers, as many individual servers connect to a top-of-rack (TOR) switch, which then must connect to the next level network. These connections typically utilize a network interface controller for each port of each server for purposes of redundancy.

[0009] Switches with many ports (*e.g.*, 24 ports, 48 ports, or more) are disproportionally expensive, on a cost-per-port basis, relative to switches with fewer ports (*e.g.*, less than 24 ports). The cost disparity is at least in part due to the increased connectivity and bandwidth between the switch chips in the switch chassis. As more ports are added, more internal connections are utilized, switch chip sizes are increased, silicon area is increased, and cost thus is increased. Furthermore, as chip-to-chip distances expand due to the number of switch chips utilized, signal loading degrades the switch bit rate and aggregate bandwidth.

[0010] The cost is further aggravated by the number of cables to connect the servers to the switches. In a typical redundant connection topology, an enclosure containing ten servers, with an "A" link and a "B" link for each server will implement twenty cables. A rack containing four such enclosures will have eighty cables. The top-of-rack (TOR) switch (or switches) then handle eighty downlinks as well as an appropriate number of uplinks. Though switch over-subscription is often employed in order to mitigate costs, it is an insufficient remedy.

[0011] Latency remains another issue. For example, for a server to communicate with a peer server in a system complex, the central processing unit communicates with the NIC, which then communicates via a cable to the TOR switch. The TOR switch forwards the information to the next level (L2) switch, then to the next level (L3) switch, then down to the next level (L2) which for purposes of this example will forward the information to the next TOR switch, then down the cable to the appropriate NIC then onward to the appropriate central processing unit. Each of these transactions will accumulate a switching delay. For this

example, servers in the same row would experience a 5-hop path with switch latency at each hop.

[0012] The latency problem is exacerbated in architectures with a high degree of east-west traffic—that is traffic between peer servers in the same enclosure. Since each packet traverses cables up to the TOR and back down, the aggregate switch bandwidth should be high to enable the servers to effectively communicate.

[0013] In terms of management, the more complex a switch is, the more difficult it is to manage. When multiple complex switches are implemented to facilitate communications between servers, the management problem is compounded.

[0014] Various implementations are described below by referring to several examples of a modular input/output aggregation zone having a switch chip. For example, a system is disclosed having a modular input/output aggregation zone to directly communicatively couple together a plurality of servers within an enclosure shared by the modular input/output aggregation zone and the plurality of servers. The example modular input/output aggregation zone includes a first switch chip having link ports configurable as uplink ports and crosslink ports, the uplink ports being communicatively coupleable to a network device and the crosslink ports being communicatively coupleable to a crosslink port of a second switch chip. Additional examples are described below.

[0015] In some implementations, cost can be significantly decreased by reducing the number of TOR switches and cables used to connect a plurality of servers. For example, efficient communications for workloads with significant amounts of east-west traffic is provided by communicating at a lower level in the switching hierarchy, reducing latency while also reducing the port count for expensive L2 and L3 switches. Also, a path with fewer hops provides improvements in latency. Moreover, management of the network topology is simplified. These and other advantages will be apparent from the description that follows.

[0016] FIG. 1 illustrates a block diagram of a system utilizing a modular input/output (I/O) aggregation zone 100 having a switch chip 110 according to examples of the present disclosure. It should be understood that FIG. 1 includes particular components, modules, etc. according to various examples. However, in

different embodiments, more, fewer, and/or other components, modules, arrangements of components/modules, etc. may be used according to the teachings described herein. In addition, various components, modules, etc. described herein may be implemented as one or more software modules, hardware modules, special-purpose hardware (e.g., application specific hardware, application specific integrated circuits (ASICs), embedded controllers, hardwired circuitry, etc.), or some combination of these.

[0017] The example system shown utilizes the modular I/O aggregation zone 100 to directly communicatively couple together a plurality of servers 104a–f within an enclosure such as shared enclosure 106 that is shared by the modular I/O aggregation zone 100 and the plurality of servers 104a–f. That is, the modular I/O aggregation zone 100 and the plurality of servers 104a–f are contained within the shared enclosure 106. The shared enclosure 106 may be made of a suitable material and of a suitable size to contain both the modular I/O aggregation zone 100 and the plurality of servers 104a–f.

[0018] The modular I/O aggregation zone 100 may contain a switch chip, such as switch chip 110, in one example. However, in another example, such as that shown in FIG. 1, the modular I/O aggregation zone 100 also includes a second switch chip 120. The modular I/O aggregation zone 100 is also configured to be directly communicatively coupled to the plurality of servers 104a–f. The switch chips described herein may include various switch chips from different manufacturers including, for example, Intel's® Red Rock Canyon switch chip.

[0019] The plurality of servers 104a–f may include servers of a similar or identical configuration, or the servers may be of a variety of types and configurations. It should be appreciated that the servers 104a–f may be blade servers, modular servers, or servers of a similar type, and may include hardware components such as processing resources, memory resources, storage resources, and other appropriate components. For example, the plurality of servers 104a–f may include a processing resource that represents generally any suitable type or form of processing unit or units capable of processing data or interpreting and executing instructions. The instructions may be stored on a non-transitory tangible computer-readable storage medium, such as a memory resource, or on a separate

device (not shown), or on any other type of volatile or non-volatile memory that stores instructions. Alternatively or additionally, the plurality of servers 104a–f may include dedicated hardware, such as one or more integrated circuits, Application Specific Integrated Circuits (ASICs), Application Specific Special Processors (ASSPs), Field Programmable Gate Arrays (FPGAs), or any combination of the foregoing examples of dedicated hardware. In some implementations, multiple processors may be used, as appropriate, along with multiple memories and/or types of memory.

[0020] The plurality of servers 104a–f are directly communicatively coupled to the modular I/O aggregation zone 100, and more specifically are directly communicatively coupled to the switch chips 110 and 120 of the modular I/O aggregation zone 100. The direct coupling may include a Peripheral Component Interconnect Express (PCIe) or similar connection between the servers 104a–f and the modular I/O aggregation zone 100. These connections are depicted by the dotted lines in FIG 1. Once directly communicatively coupled to the modular I/O aggregation zone 100, the servers may transmit and receive data among one another and with other network connected devices via the direct communicatively coupled connection to the switch chips 110 and 120.

[0021] The switch chips 110 and 120 may each include link ports, which are configurable as uplink ports and crosslink ports. In the example shown, switch chip 110 includes an uplink port 112 and two crosslink ports 114a,b. Similarly, switch chip 120 includes an uplink port 122 and two crosslink ports 124a,b. In other examples, the switch chips may include additional ports in a variety of configurations. It should be understood that the link ports may be configured (and re-configured) as either uplink ports or crosslink ports, either automatically by the nature of the connections created to the uplink ports or manually by an administrator when the chips are installed or when the modular I/O aggregation zone 100 is set up. Because the link ports are configurable, the bandwidth for the connections between the servers, the switch chips, and the network devices may be variable such that some connections may support only minimal bandwidth while other connections support much greater bandwidth. Each switch chip may be

configured individually, thus increasing the flexibility and bandwidth possibilities for each modular I/O aggregation zone.

[0022] The crosslink ports (*e.g.*, crosslink ports 114a,b and 124a,b) are communicatively coupleable to one another (or to additional switch chips) using any suitable network connection, including Ethernet, optical, or other electrical connection. In the example shown in FIG. 1, the crosslink port 114a of switch chip 110 is communicatively coupled to the crosslink port 124a of switch chip 120, and the crosslink port 114b of switch chip 110 is communicatively coupled to the crosslink port 124b of switch chip 120. These connections are depicted by the dashed lines in FIG 1. Additional, either or both of the uplink ports 112 and 122 may be configured as crosslink ports, and any or all of the crosslink ports 114a,b and 124a,b may be configured as uplink ports, as appropriate. Additional ports may also be implemented.

[0023] Data or network traffic transmitted from one of the plurality of servers to another of the plurality of servers is transmitted through at least one crosslink port of at least one of the first switch chip and the second switch chip to the other of the plurality of servers. For example, data transmitted from the server 104a to the server 104d is transmitted through crosslink port 114b of switch chip 110 and the crosslink port 124b of switch chip 120 to the server 104d. In another example, the data could be transmitted through crosslink port 114a of switch chip 110 and the crosslink port 124a of switch chip 120 to the server 104d. By transmitting the data between the switch chips within the modular I/O aggregation zone 100, the data need not be transmitted up to the network device 140 and back down to the server, thus reducing the latency, cost, and management concerns discussed above.

[0024] In this configuration, switch chips 110 and 120 are said to have redundant connections. That is, the switch chips 110 and 120 are connected to each other along two separate paths, such that if one path fails, the switch chips 110 and 120 may communicate via the second path. In other examples, such as illustrated in FIG. 4, additional switch chips may be implemented in the modular I/O aggregation zone 100, enabling the modular I/O aggregation zone 100 to provide a mesh network, ring network, star network, fully connected network, linear

network, tree network, bus network, dragonfly network, and any other suitable network topology or combinations of network topologies.

[0025] The uplink ports (*e.g.*, uplink ports 112 and 122) are communicatively coupleable to a network device, such as network device 140 using any suitable network connection, including Ethernet, optical, or other electrical connection. In the example shown in FIG. 1, the uplink port 112 of switch chip 110 is communicatively coupled to the network device 140. In other examples, the uplink port 122 of switch chip 120 may also be communicatively coupled to the network device 140 or another network device as depicted by the dashed line between the crosslink port 112 of switch chip 110 and the network device 140.

[0026] The network device 140 may be any suitable network device, including at least a switch, a hub, and a router. The network device 140 may be part of a larger network, the network representing generally hardware components and computers interconnected by communications channels that allow sharing of resources and information. The network may include one or more of a cable, wireless, fiber optic, or remote connection via a telecommunication link, an infrared link, a radio frequency link, or any other connectors or systems that provide electronic communication. The network may include, at least in part, an intranet, the Internet, or a combination of both. The network may also include intermediate proxies, routers, switches, load balancers, and the like, including the network device 140 and the modular I/O aggregation zone 100 via the switch chips 110 and 120. The paths followed by the network between switch chip 110 and network device 140 as depicted in FIG. 1 represent the logical communication paths between these devices, not necessarily the physical paths between the devices.

[0027] In other examples, as discussed below, the modular I/O aggregation zone 100 may be communicatively coupled to another modular I/O aggregation zone to expand or scale the number of servers serviced by the functionality that the modular I/O aggregation zone 100 provides. For example, the additional modular input/output aggregation zone directly communicatively couple together additional pluralities of servers within shared enclosures, and each of the plurality of additional modular input/output aggregation include at least one switch chip having link ports configurable as uplink ports and crosslink ports. The additional

modular I/O aggregation zones may be arranged in a variety of network topologies. For instance, the modular I/O aggregation zones may be arranged in a mesh network, ring network, star network, fully connected network, linear network, tree network, bus network, dragonfly network, and any other suitable network topology or combinations of network topologies.

[0028] FIG. 2 illustrates a block diagram of a system utilizing two modular input/output aggregation zones each having a switch chip according to examples of the present disclosure. It should be understood that FIG. 2 includes particular components, modules, etc. according to various examples. However, in different embodiments, more, fewer, and/or other components, modules, arrangements of components/modules, etc. may be used according to the teachings described herein. In addition, various components, modules, etc. described herein may be implemented as one or more software modules, hardware modules, special-purpose hardware (e.g., application specific hardware, application specific integrated circuits (ASICs), embedded controllers, hardwired circuitry, etc.), or some combination of these.

[0029] Like FIG. 1, FIG. 2 illustrates a modular I/O aggregation zone 200 to directly communicatively couple together a plurality of servers 204a–c within an enclosure such as shared enclosure 206 that is shared by the modular I/O aggregation zone 200 and the plurality of servers 204a–c. The modular I/O aggregation zone 200 may contain a switch chip, such as switch chip 210. The modular I/O aggregation zone 200 is also configured to be directly communicatively coupled to the plurality of servers 204a–c.

[0030] Additionally, FIG. 2 illustrates a second modular I/O aggregation zone 201 to directly communicatively couple together a second plurality of servers 205d–f within a second enclosure such as shared enclosure 207 that is shared by the second modular I/O aggregation zone 201 and the second plurality of servers 205d–f. The second modular I/O aggregation zone 201 may contain a second switch chip, such as switch chip 211. The second modular I/O aggregation zone 201 is also configured to be directly communicatively coupled to the plurality of servers 205d–f.

[0031] The switch chips 210 and 211 each include link ports configurable as uplink ports and crosslink ports. For example, switch chip 210 includes uplink port 212 and crosslink ports 214a,b while switch chip 211 includes uplink port 213 and crosslink ports 215a,b. In the example illustrated, the crosslink port 214b of switch chip 210 is communicatively coupled to the crosslink port 215a of the second switch chip 211. Thus, data may be transmitted between the first plurality of servers 204a–c and the second plurality of servers 204d–f via the first switch chip 210 and the second switch chip 211 without having to transmit the data up to a higher level network device (not shown).

[0032] In examples, additional crosslink ports of switch chips 210 and 211 may be communicatively coupled to additional switch chips (not shown) within the respective modular I/O aggregation zones 200 and 201. For example, either of the illustrated modular I/O aggregation zones 200 and 201 may include additional switch chips, which may be communicatively coupled via optical or electrical links such as Ethernet links.

[0033] Moreover, additional crosslink ports of switch chips 210 and 211 may be communicatively coupled to the switch chips of additional modular I/O aggregation zones (not shown). For example, a third modular input/output aggregation zone may communicatively couple together a third plurality of servers. The third modular input/output aggregation zone may include a third switch chip having link ports configurable as uplink ports and crosslink ports. Then, the first, second, and third modular input/output aggregation zones may be communicatively coupled in any number or combinations of appropriate network topologies such as mesh network, ring network, star network, fully connected network, linear network, tree network, bus network, dragonfly network, and any other suitable network topology. In this way, multiple modular I/O aggregation zones can be linked together in a variety of network topologies to enable servers such as servers 204a–c, servers 205d–f, and additional servers to transmit and receive network traffic and data without having to transmit the network traffic and data up to a higher level network device (not shown).

[0034] The crosslink port 214a of switch chip 210 may be communicatively coupled to the crosslink port 215b of switch chip 211 to create two discrete network

paths between the modular I/O aggregation zone 200 and the second modular I/O aggregation zone 201.

[0035] In another example, the uplink port 212 of the switch chip 210 and/or the uplink port 213 of the second switch chip 211 may be communicatively coupled to a network device, such as a switch, hub, router, or other appropriate network device, using optical or electrical networking connections. Additional, either or both of the uplink ports 212 and 213 may be configured as crosslink ports, and any or all of the crosslink ports 214a,b and 215a,b may be configured as uplink ports, as appropriate. Additional ports may also be implemented.

[0036] Because the link ports are configurable, the bandwidth for the connections between the servers, the switch chips, and the network devices may be variable such that some connections may support only minimal bandwidth while other connections support much greater bandwidth. Each switch chip may be configured individually, thus increasing the flexibility and bandwidth possibilities for each modular I/O aggregation zone.

[0037] FIG. 3 illustrates a block diagram of a modular input/output aggregation zone 300 having a first switch chip 310 and a second switch chip 320 according to examples of the present disclosure. It should be understood that FIG. 3 includes particular components, modules, etc. according to various examples. However, in different embodiments, more, fewer, and/or other components, modules, arrangements of components/modules, etc. may be used according to the teachings described herein. In addition, various components, modules, etc. described herein may be implemented as one or more software modules, hardware modules, special-purpose hardware (e.g., application specific hardware, application specific integrated circuits (ASICs), embedded controllers, hardwired circuitry, etc.), or some combination of these.

[0038] The modular input/output aggregation zone 300 shown directly communicatively couples together a plurality of servers 304a–f within an enclosure such as shared enclosure 306 that is shared by the modular I/O aggregation zone 300 and the plurality of servers 304a–f. That is, the modular I/O aggregation zone 300 and the plurality of servers 304a–f are contained within the shared enclosure 306. The shared enclosure 306 may be made of a suitable material and of a

suitable size to contain both the modular I/O aggregation zone 300 and the plurality of servers 304a–f.

[0039] The plurality of servers 304a–f may include servers of a similar or identical configuration, or the servers may be of a variety of types and configurations. It should be appreciated that the servers 304a–f may be blade servers, modular servers, or servers of a similar type, and may include hardware components such as processing resources, memory resources, storage resources, and other appropriate components. For example, the plurality of servers 304a–f may include a processing resource that represents generally any suitable type or form of processing unit or units capable of processing data or interpreting and executing instructions. The instructions may be stored on a non-transitory tangible computer-readable storage medium, such as a memory resource, or on a separate device (not shown), or on any other type of volatile or non-volatile memory that stores instructions. Alternatively or additionally, the plurality of servers 304a–f may include dedicated hardware, such as one or more integrated circuits, Application Specific Integrated Circuits (ASICs), Application Specific Special Processors (ASSPs), Field Programmable Gate Arrays (FPGAs), or any combination of the foregoing examples of dedicated hardware. In some implementations, multiple processors may be used, as appropriate, along with multiple memories and/or types of memory.

[0040] The plurality of servers 304a–f are directly communicatively coupled to the modular I/O aggregation zone 300, and more specifically are directly communicatively coupled to the switch chips 310 and 320 of the modular I/O aggregation zone 300 via a Peripheral Component Interconnect Express (PCIe) connection 306 (shown as dotted lines in FIG. 3). Once directly communicatively coupled to the modular I/O aggregation zone 300, the servers may transmit and receive data among one another and with other network connected devices via the PCIe connection to the switch chips 310 and 320.

[0041] The modular I/O aggregation zone 300 may contain a first switch chip 310 and a second switch chip 320. The switch chips 310 and 320 may each include link ports, which are configurable as uplink ports and crosslink ports. In the example shown, switch chip 310 includes an uplink port 312 and two crosslink ports 314a,b.

Similarly, switch chip 320 includes an uplink port 322 and two crosslink ports 324a,b. In other examples, the switch chips may include additional ports in a variety of configurations. It should be understood that the link ports may be configured (and re-configured) as either uplink ports or crosslink ports, either automatically by the nature of the connections created to the uplink ports or manually by an administrator when the chips are installed or when the modular I/O aggregation zone 300 is set up. Because the link ports are configurable, the bandwidth for the connections between the servers, the switch chips, and the network devices may be variable such some connections may support only minimal bandwidth while other connections support much greater bandwidth. Each switch chip may be configured individually, thus increasing the flexibility and bandwidth possibilities for each modular I/O aggregation zone.

[0042] The crosslink ports (*e.g.*, crosslink ports 314a,b and 324a,b) are communicatively coupleable to one another (or to additional switch chips) using any suitable network connection, including Ethernet, optical, or other electrical connection. In the example shown in FIG. 3, the crosslink port 314a of switch chip 310 is communicatively coupled to the crosslink port 324a of switch chip 320, and the crosslink port 314b of switch chip 310 is communicatively coupled to the crosslink port 324b of switch chip 320. These connections are depicted by the dashed lines in FIG 3. Additional, either or both of the uplink ports 312 and 322 may be configured as crosslink ports, and any or all of the crosslink ports 314a,b and 324a,b may be configured as uplink ports, as appropriate. In an example, at least one of the uplink ports 312 and 322 of switch chips 310 and 320 respectively may be communicatively coupled to a network device (not shown) such as a switch, router, hub, or other suitable networking device. The connection between the uplink ports and the networking device may be an optical network connection in one example or may be an electrical connection in another example. It is also possible that multiple connections between the switch chips and the network device are implemented using a combination of different network connection types. For example, the uplink port 312 of the switch chip 310 may be connected to a network device via an electrical connection while the uplink port 322 of the switch chip 320 may be connected to the same network device via an optical connection. Of course,

in examples, the connections between the network device and the uplink ports may be of the same type in any suitable number. Additional ports may also be implemented.

[0043] Data or network traffic transmitted from one of the plurality of servers to another of the plurality of servers is transmitted through at least one crosslink port of at least one of the first switch chip and the second switch chip to the other of the plurality of servers. For example, data transmitted from the server 304a to the server 304d is transmitted through crosslink port 314b of switch chip 310 and the crosslink port 324b of switch chip 320 to the server 304d. In another example, the data could be transmitted through crosslink port 314a of switch chip 310 and the crosslink port 324a of switch chip 320 to the server 304d. By transmitting the data between the switch chips within the modular I/O aggregation zone 300, the data need not be transmitted up to a network device (not shown) and back down to the server, thus reducing the latency, cost, and management concerns discussed above.

[0044] In this configuration, switch chips 310 and 320 are said to have redundant connections. That is, the switch chips 310 and 320 are connected to each other along two separate paths, such that if one path fails, the switch chips 310 and 320 may communicate via the second path. In other examples, such as illustrated in FIG. 4, additional switch chips may be implemented in the modular I/O aggregation zone 300, enabling the modular I/O aggregation zone 300 to provide a mesh network, star network, or other appropriate network topology among the switch chips.

[0045] FIG. 4 illustrates a block diagram of a modular input/output aggregation zone 400 having a first switch chip 410, a second switch chip 420, and a third switch chip 430 according to examples of the present disclosure. It should be understood that FIG. 4 includes particular components, modules, etc. according to various examples. However, in different embodiments, more, fewer, and/or other components, modules, arrangements of components/modules, etc. may be used according to the teachings described herein. In addition, various components, modules, etc. described herein may be implemented as one or more software modules, hardware modules, special-purpose hardware (e.g., application specific

hardware, application specific integrated circuits (ASICs), embedded controllers, hardwired circuitry, etc.), or some combination of these.

[0046] The modular input/output aggregation zone 400 shown directly communicatively couples together a plurality of servers 404a–i within an enclosure such as shared enclosure 406 that is shared by the modular I/O aggregation zone 400 and the plurality of servers 404a–i. That is, the modular I/O aggregation zone 400 and the plurality of servers 404a–i are contained within the shared enclosure 406. The shared enclosure 406 may be made of a suitable material and of a suitable size to contain both the modular I/O aggregation zone 400 and the plurality of servers 404a–i.

[0047] The plurality of servers 404a–i may include servers of a similar or identical configuration, or the servers may be of a variety of types and configurations. It should be appreciated that the servers 404a–i may be blade servers, modular servers, or servers of a similar type, and may include hardware components such as processing resources, memory resources, storage resources, and other appropriate components. For example, the plurality of servers 404a–i may include a processing resource that represents generally any suitable type or form of processing unit or units capable of processing data or interpreting and executing instructions. The instructions may be stored on a non-transitory tangible computer-readable storage medium, such as a memory resource, or on a separate device (not shown), or on any other type of volatile or non-volatile memory that stores instructions. Alternatively or additionally, the plurality of servers 404a–i may include dedicated hardware, such as one or more integrated circuits, Application Specific Integrated Circuits (ASICs), Application Specific Special Processors (ASSPs), Field Programmable Gate Arrays (FPGAs), or any combination of the foregoing examples of dedicated hardware. In some implementations, multiple processors may be used, as appropriate, along with multiple memories and/or types of memory.

[0048] The plurality of servers 404a–i are directly communicatively coupled to the modular I/O aggregation zone 400, and more specifically are directly communicatively coupled to the switch chips 410, 420, and 430 of the modular I/O aggregation zone 400 via a direct connection such as a Peripheral Component

Interconnect Express (PCIe) connection or other suitable connection (shown as dotted lines in FIG. 4). Once directly communicatively coupled to the modular I/O aggregation zone 400, the servers may transmit and receive data among one another and with other network connected devices via the direct connection to the switch chips 410, 420, and 430.

[0049] The modular I/O aggregation zone 400 may contain a first switch chip 410, a second switch chip 420, and a third switch chip 430. The switch chips 410, 420, and 430 may each include link ports, which are configurable as uplink ports and crosslink ports. It should be understood that the link ports may be configured (and re-configured) as either uplink ports or crosslink ports, either automatically by the nature of the connections created to the uplink ports or manually by an administrator when the chips are installed or when the modular I/O aggregation zone 400 is set up. Because the link ports are configurable, the bandwidth for the connections between the servers, the switch chips, and the network devices may be variable such some connections may support only minimal bandwidth while other connections support much greater bandwidth. Each switch chip may be configured individually, thus increasing the flexibility and bandwidth possibilities for each modular I/O aggregation zone.

[0050] The crosslink ports are communicatively coupleable to one another (or to additional switch chips) using any suitable network connection, including Ethernet, optical, or other electrical connection. In the example shown in FIG. 4, a crosslink port of switch chip 410 is communicatively coupled to a crosslink port switch chip 420 and to a crosslink port of switch chip 430. Similarly, a crosslink port of switch chip 420 is communicatively coupled to a crosslink port of switch chip 430. In this configuration, each switch chip is communicatively coupled to each of the other two switch chips, forming a redundant network topology such that if any one connection fails, the switch chips may still communicate via the remaining connections. The configuration shown is only one possible network topology, and other network topologies may include a mesh network, ring network, star network, fully connected network, linear network, tree network, bus network, dragonfly network, and any other suitable network topology or combinations of network topologies.

[0051] It should be emphasized that the above-described examples are merely possible examples of implementations and set forth for a clear understanding of the present disclosure. Many variations and modifications may be made to the above-described examples without departing substantially from the spirit and principles of the present disclosure. Further, the scope of the present disclosure is intended to cover any and all appropriate combinations and sub-combinations of all elements, features, and aspects discussed above. All such appropriate modifications and variations are intended to be included within the scope of the present disclosure, and all possible claims to individual aspects or combinations of elements or steps are intended to be supported by the present disclosure.

CLAIMS

WHAT IS CLAIMED IS:

1. A system comprising:

a modular input/output aggregation zone to directly communicatively couple together a plurality of servers within an enclosure shared by the modular input/output aggregation zone and the plurality of servers, the modular input/output aggregation zone comprising a first switch chip having link ports configurable as uplink ports and crosslink ports, the uplink ports being communicatively coupleable to a network device and the crosslink ports being communicatively coupleable to a crosslink port of a second switch chip.

2. The system of claim 1, wherein two crosslink ports of the first switch chip are communicatively coupled to two crosslink ports of the second switch chip.

3. The system of claim 2, wherein the two crosslink ports of the first and second switch chips are communicatively coupled via at least one of the group consisting of an optical cable and an electrical cable.

4. The system of claim 1, wherein the plurality of servers are directly communicatively coupled together using Peripheral Component Interconnect Express connections in the modular input/output aggregation zone.

5. The system of claim 1, wherein the modular input/output aggregation zone is communicatively coupleable to a plurality of additional modular input/output aggregation zones, the plurality of additional modular input/output aggregation zone to directly communicatively couple together additional pluralities of servers within shared enclosures, wherein each of the plurality of additional modular input/output aggregation zones comprises a switch chip having link ports configurable as uplink ports and crosslink ports.

6. The system of claim 5, wherein the modular input/output aggregation zone and the plurality of additional modular input/output aggregation zones are arranged in a mesh network topology.

7. The system of claim 1,

wherein an uplink port of the second switch chip is communicatively coupled to the networking device, and

wherein a second uplink port of the first switch chip and a second uplink port of the second switch chip are communicatively coupled to the networking device via at least one of the group consisting of an optical cable and an electrical cable.

8. The system of claim 1, wherein data transmitted from one of the plurality of servers to another of the plurality of servers is transmitted through at least one crosslink port of at least one of the first switch chip and the second switch chip to the other of the plurality of servers.

9. The system of claim 8, wherein the data transmitted from one of the plurality of servers to another of the plurality of servers is not transmitted to the network device.

10. A system comprising:

a first modular input/output aggregation zone to directly communicatively couple together a first plurality of servers within a first enclosure shared by the first modular input/output aggregation zone and the first plurality of servers, the first modular input/output aggregation zone comprising a first switch chip having link ports configurable as uplink ports and crosslink ports; and

a second modular input/output aggregation zone to directly communicatively couple together a second plurality of servers within a second enclosure shared by the second modular input/output aggregation zone and the second plurality of servers, the second modular input/output aggregation zone comprising a second switch chip having link ports configurable as uplink ports and crosslink ports,

wherein at least one of the crosslink ports of the first switch chip of the first modular input/output aggregation zone is communicatively coupled to at least one of the crosslink ports of the second switch chip of the second modular input/output aggregation zone, and

wherein a total bandwidth of the crosslink ports and the uplink ports of the first and second switch chips is greater than a total bandwidth of the direct communicative coupling together of first and second plurality of servers to the respective first and second modular input/output aggregation zones.

11. The system of claim 10, wherein at least one of the uplink ports of the first switch chip and at least one of the uplink ports of the second switch chip are communicatively coupleable to a network device.

12. The system of claim 10, further comprising:

a third modular input/output aggregation zone to directly communicatively couple together a third plurality of servers, the third modular input/output aggregation zone comprising a third switch chip having link ports configurable as uplink ports and crosslink ports,

wherein the first, second, and third modular input/output aggregation zones are communicatively coupled in a networking topology selected from the group consisting of a mesh network, a ring network, a star network, a fully connected network, a linear network, a tree network, a bus network, and a dragonfly network.

13. A modular input/output aggregation zone comprising:

a first switch chip having first link ports configurable as first uplink ports and first crosslink ports; and

a second switch chip having second link ports configurable as second uplink ports and second crosslink ports,

wherein at least one of the first crosslink ports of the first switch chip is communicatively coupleable to at least one of the second crosslink ports of the second switch chip, and

wherein the modular input/output aggregation zone is directly communicatively coupled to a plurality of servers within an enclosure shared by the plurality of servers and the modular input/output aggregation zone via Peripheral Connect Interconnect Express connections.

14. The modular input/output aggregation zone of claim 13, wherein at least one of the first uplink port and at least one of the second uplink port is communicatively coupleable to a network device by an optical network connection.

15. The modular input/output aggregation zone of claim 13, wherein the at least one of the first crosslink ports of the first switch chip is communicatively coupleable to the at least one of the second crosslink ports of the second switch chip via an electrical network connection.

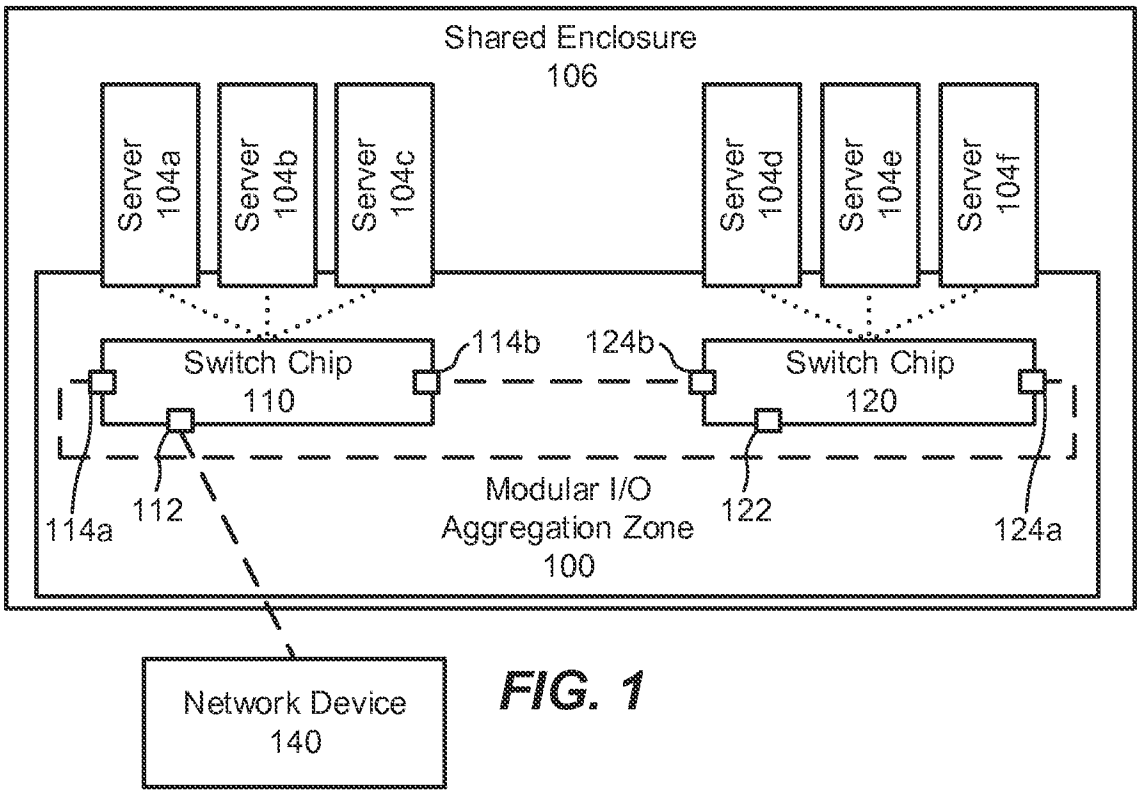
Shared Enclosure
106
Server 104a
Server 104b
Server 104c
Server 104d
Server 104e
Server 104f
Switch Chip
110
114b
124b
Switch Chip
120
114a
112
Modular I/O
Aggregation Zone
100
122
124a
Network Device
140

FIG. 1

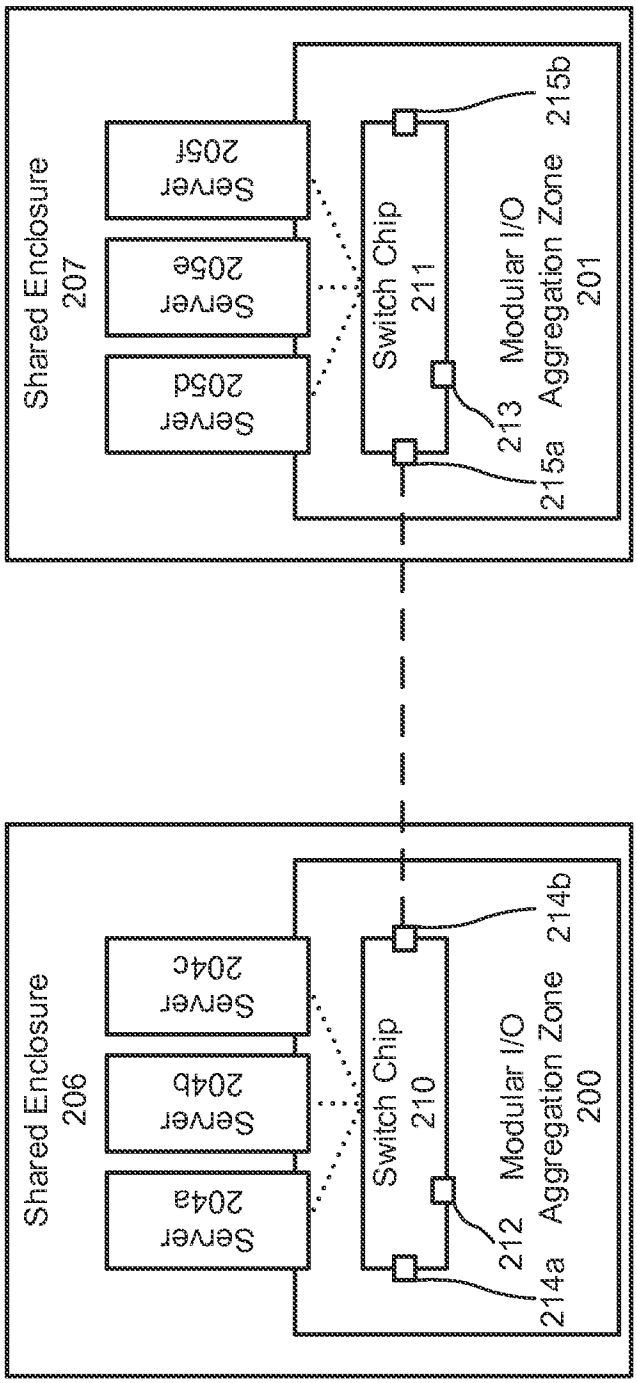
Shared Enclosure
206
Server
204a
Server
204b
Server
204c
Switch Chip
210
212
214a
Modular I/O
Aggregation Zone
200
214b
Shared Enclosure
207
Server
205d
Server
205e
Server
205f
Switch Chip
211
213
215a
Modular I/O
Aggregation Zone
201
215b

FIG. 2

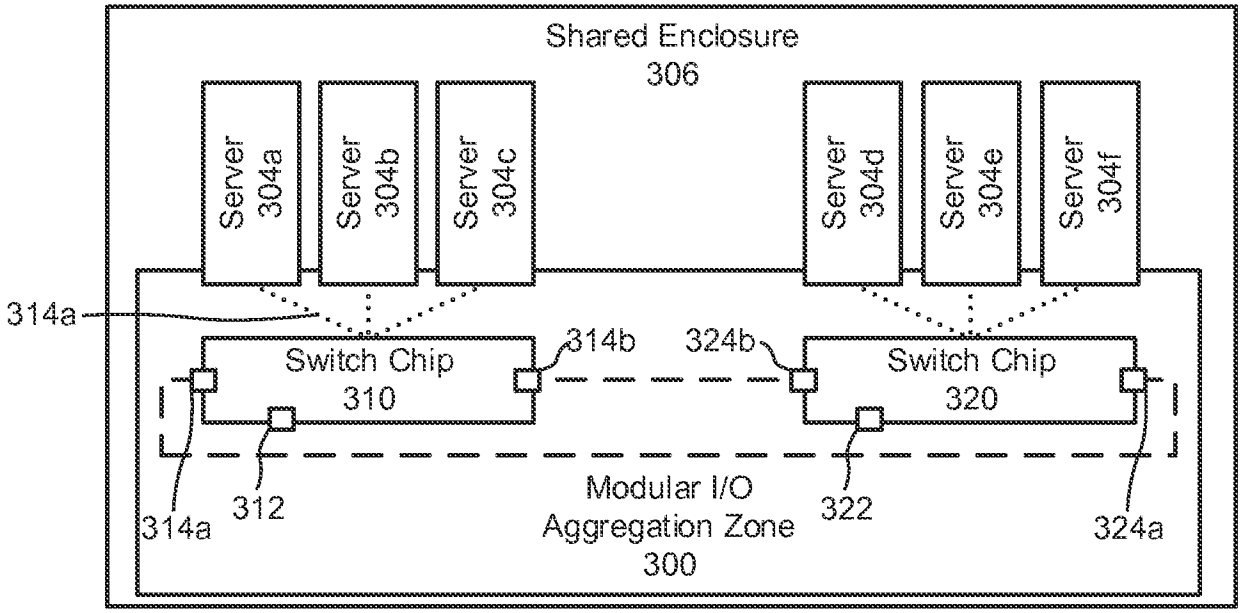
Shared Enclosure
306
Server 304a
Server 304b
Server 304c
Server 304d
Server 304e
Server 304f
314a
Switch Chip
310
314b
324b
Switch Chip
320
314a
312
Modular I/O
Aggregation Zone
300
322
324a

FIG. 3

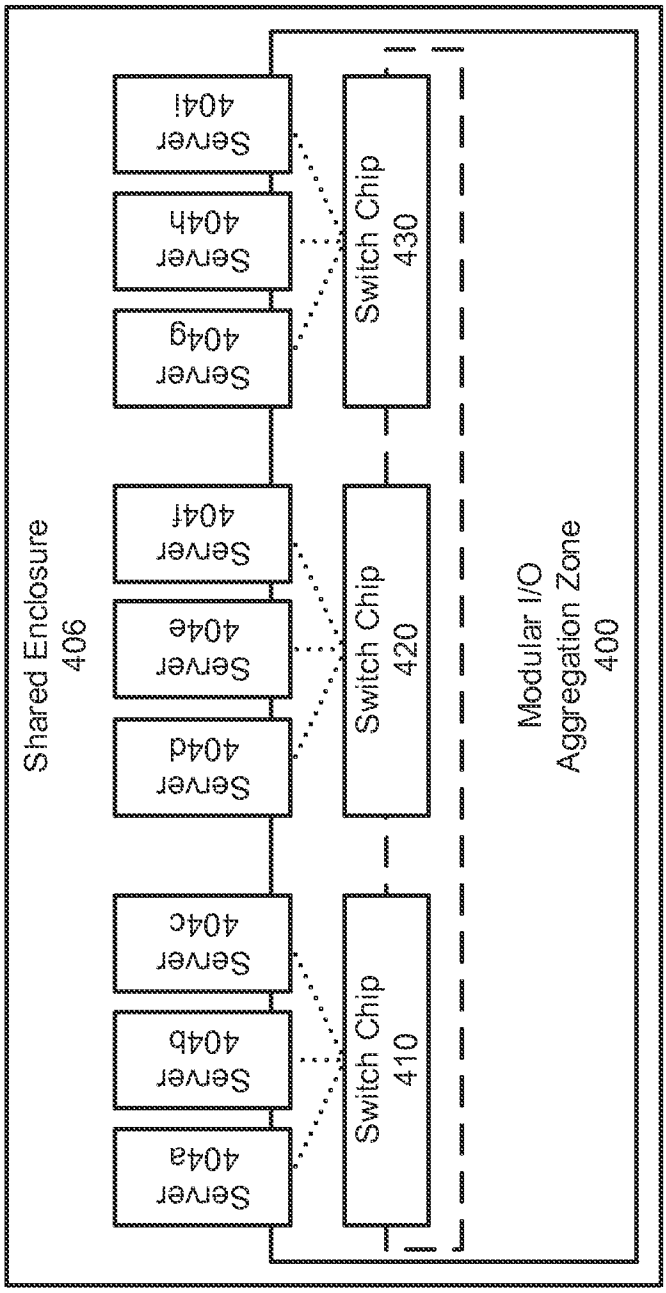
Shared Enclosure
406
Server
404a
Server
404b
Server
404c
Server
404d
Server
404e
Server
404f
Server
404g
Server
404h
Server
404i
Switch Chip
410
Switch Chip
420
Switch Chip
430
Modular I/O
Aggregation Zone
400

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/032066

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/931(2013.01)i

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 12/931; H04L 12/50; H04L 12/28; H04L 12/16; G06F 13/00; H04L 12/40; H04L 12/56

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: modular, input/output, aggregation zone, switch, crosslink, uplink, port.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 8031722 B1 (ALEX SANVILLE et al.) 04 October 2011 See See column 4, line 9 - column 5, line 30, column 7, lines 43-45, column 9, lines 42-47; and figures 1-4.	1-3,5-8
Y		4,13-15
A		9-12
Y	US 2007-0097948 A1 (WILLAM T. BOYD. et al.) 03 May 2007 See paragraphs [0030]-[0045]; claim 9; and figures 1-3.	4,13-15
A	US 2013-0156028 A1 (HENDRICH M. HERNANDEZ et al.) 20 June 2013 See paragraphs [0019]-[0028]; and figures 1-2.	1-15
A	US 2013-0322434 A1 (MICHAEL ARMBRUSTER et al.) 05 December 2013 See paragraphs [0112]-[0120], [0142]-[0145]; claim 1; and figures 1-9.	1-15
A	US 2012-0201253 A1 (OMAR CARDONA et al.) 09 August 2012 See paragraphs [0060]-[0067]; and figures 5-6.	1-15

☐ Further documents are listed in the continuation of Box C.

☒ 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
10 December 2014 (10.12.2014)	**11 December 2014 (11.12.2014)**

Name and mailing address of the ISA/KR
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701,
Republic of Korea
Facsimile No. +82-42-472-7140

Authorized officer
KIM, Seong Woo
Telephone No. +82-42-481-3348

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/US2014/032066

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 8031722 B1	04/10/2011	None	
US 2007-0097948 A1	03/05/2007	CN 100488147 C	13/05/2009
		CN 1968170 A	23/05/2007
		US 2008-0140839 A1	12/06/2008
		US 7363404 B2	22/04/2008
		US 7549003 B2	16/06/2009
US 2013-0156028 A1	20/06/2013	None	
US 2013-0322434 A1	05/12/2013	CN 103454992 A	18/12/2013
		DE 102012209108 A1	05/12/2013
		DE 102012209108 B4	15/05/2014
		EP 2670087 A1	04/12/2013
US 2012-0201253 A1	09/08/2012	US 2012-102217 A1	26/04/2012
		US 8819235 B2	26/08/2014
		US 8856340 B2	07/10/2014