



(51) International Patent Classification:
G06F 13/40 (2006.01)

(21) International Application Number:

PCT/US2018/022181

(22) International Filing Date:

13 March 2018 (13.03.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

15/490,490

18 April 2017 (18.04.2017)

US

(71) Applicant: **WESTERN DIGITAL TECHNOLOGIES INC.** [US/US]; 5601 Great Oaks Parkway, San Jose, California 95119 (US).

(72) Inventors: **HESS, Randall**; c/o Western Digital Technologies Inc., 5601 Great Oaks Parkway, San Jose, California 95119 (US). **NASH, Berck**; c/o Western Digital Technologies Inc., 5601 Great Oaks Parkway, San Jose, Califor-

nia 95119 (US). **WALKER, Michael**; c/o Western Digital Technologies Inc., 5601 Great Oaks Parkway, San Jose, California 95119 (US).

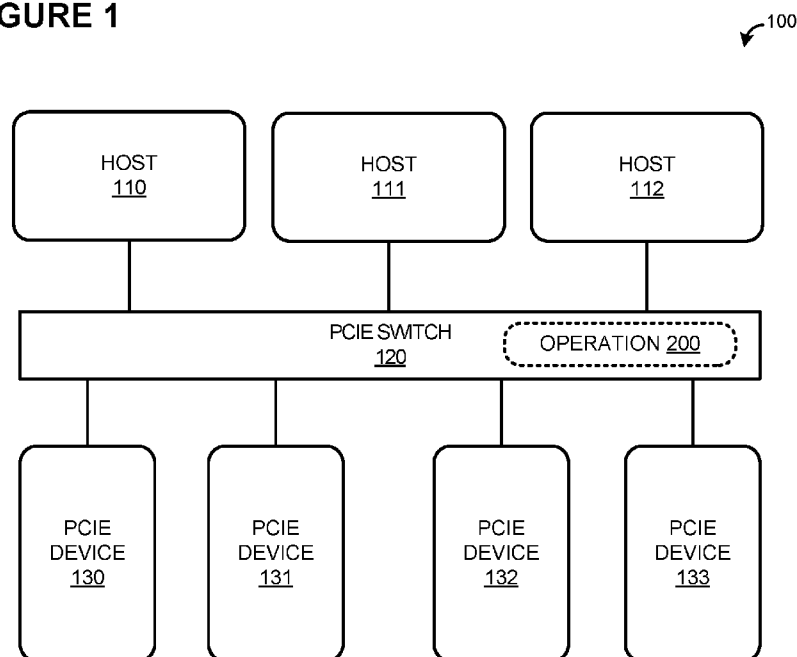
(74) Agent: **BAKER, Kent L.** et al.; Setter Roche LLP, 14694 Orchard Pkwy, Building A, Suite 200, Westminster, Colorado 80023 (US).

(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, 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,

(54) Title: MULTI-HOST PERIPHERAL COMPONENT INTERCONNECT EXPRESS (PCIe) SWITCHING

FIGURE 1



(57) Abstract: Described herein are enhancements for managing multi-host Peripheral Component Interconnect Express (PCIe) switching. In one implementation, a system includes one or more PCIe devices and a PCIe switch configured to receive a first interrupt corresponding to a first interrupt vector from a PCIe device, wherein the first interrupt vector comprises at least a virtual address and a first data value. The switch is further configured to translate the first interrupt vector into a second interrupt vector, wherein the second interrupt vector comprises a second address and a second data value, and transfer a second interrupt using the second interrupt vector to a host of a plurality of hosts that corresponds to the second interrupt vector.

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 applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

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

MULTI-HOST PERIPHERAL COMPONENT INTERCONNECT EXPRESS (PCIe) SWITCHING

5 BACKGROUND

[0001] Peripheral Component Interconnect Express (PCIe) is a high-speed serial computer expansion bus standard that is used to communicate between a host processing system and an expansion device. These expansion devices may comprise networking devices, storage devices, graphics processing devices, among other possible devices. To
10 provide the communications, packets are transferred to and from the host processing system to provide the desired functionality. For example, when data is requested from a PCIe storage device, one or more packets are transferred to the host computing system over the PCIe lanes to the host processing system.

[0002] In some implementations, in signaling the host processing system of a
15 particular event, the PCIe devices may be configured with interrupt vectors that are used in notifying the host when a communication is required. These interrupt vectors often include a data portion and an address to be written to. Accordingly, when an interrupt vector is received, the host may identify the address and the corresponding data and determine the source, and in some examples type, of the interrupt before providing any required
20 operations associated with the interrupt. However, although interrupt vectors are useful when a single host is connected to a PCIe device, issues can arise when multiple hosts are desired to be connected to the same PCIe device.

OVERVIEW

[0003] Examples disclosed herein provide enhancements for managing multi-host Peripheral Component Interconnect Express (PCIe) switching. In one implementation, a computing apparatus comprises a computing apparatus with computer readable storage media and a processing system. The computing apparatus further includes program

5 instructions stored on the computer readable storage media that provide PCIe switching functionality that, when executed by the processing system, direct the processing system to receive a first interrupt corresponding to a first interrupt vector from a PCIe device, wherein the first interrupt vector comprises at least a virtual address and a data value. The program instructions further direct the processing system to translate the first interrupt vector into a

10 second interrupt vector, wherein the second interrupt vector comprises a second address and a second data value, and transfer a second interrupt using the second interrupt vector to a host of a plurality of hosts that corresponds to the second interrupt vector.

BRIEF DESCRIPTION OF THE DRAWINGS

15 [0004] The following description and associated figures teach the best mode of the invention. For the purpose of teaching inventive principles, some conventional aspects of the best mode may be simplified or omitted. The following claims specify the scope of the invention. Note that some aspects of the best mode may not fall within the scope of the invention as specified by the claims. Thus, those skilled in the art will appreciate variations

20 from the best mode that fall within the scope of the invention. Those skilled in the art will appreciate that the features described below can be combined in various ways to form multiple variations of the invention. As a result, the invention is not limited to the specific examples described below, but only to the claims and their equivalents.

[0005] Figure 1 illustrates a computing environment to provide Peripheral Component Interconnect Express (PCIe) switching for a plurality of hosts according to an implementation.

[0006] Figure 2 illustrates an operation of a PCIe switch to manage interrupts from
5 connected PCIe devices according to an implementation.

[0007] Figure 3 illustrates a configuration of a PCIe switch and PCIe devices according to an implementation.

[0008] Figure 4 illustrates an operational scenario of managing interrupts from PCIe devices according to an implementation.

10 [0009] Figure 5 illustrates an operational scenario of configuring data structures to manage interrupts from PCIe devices according to an implementation.

[0010] Figure 6 illustrates a PCIe switch computing system to manage interrupts for PCIe devices according to an implementation.

15 DETAILED DESCRIPTION

[0011] Figure 1 illustrates a computing environment 100 to provide Peripheral Component Interconnect Express (PCIe) switching for a plurality of hosts according to an implementation. Computing environment 100 includes hosts 110-112, PCIe switch 120, and PCIe devices 130-131. PCIe switch 120 is configured to provide operation 200 to
20 manage interrupts from PCIe devices 130-133 to hosts 110-112.

[0012] In operation, hosts 110-112 may execute various applications and processes that require operations with PCIe devices 130-133. These operations may include

transferring data as packets to the devices, receiving data as packets from the devices, or some other similar functionality. PCIe devices 130-133 may comprise data storage devices, such as Non-Volatile Memory express (NVMe) devices, networking devices, such as a network interface card (NIC), a graphics processing unit (GPU), or some other PCIe device.

5 **[0013]** To provide the operations with the PCIe devices, PCIe switch 120 is provided that is used to ensure that each PCIe device of PCIe devices 130-133 is capable of providing interrupts to the required host. For example, when PCIe device 133 requires an interrupt of host 110, PCIe switch 120 may identify the request, identify the appropriate host for the communication, and forward the interrupt to the appropriate host. In providing this
10 forwarding mechanism, PCIe switch 120 may maintain one or more data structures that can be used to store interrupt vectors provided from each host in hosts 110-112, wherein the interrupt vectors are used to identify interrupts from each PCIe device. For example, when coupled to PCIe switch 120, host 110 may provide interrupt vectors for each PCIe device of PCIe devices 130-133 to communicate with host 110. These interrupt vectors each include
15 an address (memory address) on the host and a data value, such that when an interrupt is required a PCIe device may write to the address with the specified data value.

[0014] In addition to maintaining the data structure for the interrupt vectors provided from hosts 110-112, PCIe switch 120 may further provide, for each interrupt vector in the plurality of interrupt vectors configured by the plurality of hosts, a virtual interrupt vector to
20 a corresponding PCIe device. These virtual interrupt vectors, similar to the interrupts from hosts 110-112, may include an address portion and a data portion, wherein the address does not reflect the address provided from the individual hosts. For example, if host 110 provided an address “FFEE0000” with a data value of “4A00” to be used by PCIe device 130, then PCIe switch 120 may identify a virtual interrupt vector with a virtual address,

such as “BBB80000,” and provide the virtual interrupt vector to the PCIe device.

Accordingly, rather than requiring the individual PCIe device to manage interrupt vectors for each of the hosts, PCIe switch 120 may be used as an intermediary to provide a set of virtual interrupt vectors that may then be translated to the interrupt vector associated with
5 the destination host, and forwarded to the destination host.

[0015] To further demonstrate the operations of virtual switch 120, Figure 2 is provided. Figure 2 illustrates an operation 200 of a PCIe switch to manage interrupts from connected PCIe devices according to an implementation. The processes of operation 200 are referenced parenthetically in the paragraphs that follow with reference to systems and
10 elements of computing environment 100 of Figure 1.

[0016] As described in Figure 1, hosts 110-112 provide interrupt vectors to PCIe switch 120, wherein each interrupt vector includes an address portion and a data portion. As an example, host 110, may provide sixteen different interrupt vectors, wherein four of vectors may correspond to each of the PCIe devices. Thus, when an interrupt is required
15 from one of the devices, host 110 will expect to receive an interrupt using one of the defined vectors. Here, instead of forwarding the defined vectors from the hosts to the PCIe devices, PCIe switch 120 will define virtual interrupt vectors to be supplied to the PCIe devices, wherein each virtual interrupt vector corresponds to an interrupt vector provided from hosts 110-112. These virtual vectors may each include a virtual address, wherein the virtual
20 address is different than the corresponding address in the vector provided from one of hosts 110-112. For example, PCIe device 130 may receive virtual interrupt vectors that correspond to the interrupt vectors provided from hosts 110-112 for PCIe device 130. Consequently, rather than using the interrupt vectors from the hosts, each of the PCIe devices may use the single set of virtual interrupt vectors provided from the switch.

Additionally, in some implementations, switch 120 may ensure that each of the virtual interrupt vectors are unique when they are provided to the corresponding PCIe devices.

[0017] As the vectors are maintained on the switch and the PCIe devices, operation 200 directs PCIe switch 120 to receive (201) a first interrupt using a first interrupt vector from a PCIe device, wherein the first interrupt vector comprises at least a virtual address and a data value. In response to receiving the first interrupt vector, PCIe switch 120 translates (202) the first interrupt vector into a second interrupt vector, wherein the second interrupt vector comprises a second address and a second data value, which may be equivalent or different from the original data value. In particular, because PCIe switch 120 maintains interrupt vectors received from hosts 110-112, PCIe switch 120 may be required to translate the virtual interrupt vectors received from the PCIe devices to the appropriate interrupt vectors associated with hosts 110-112. In some implementations, the virtual interrupt vectors may appear to the PCIe device as though the PCIe switch is the host for the device. Consequently, when a virtual interrupt vector is received at the PCIe switch, the virtual vector will require translation (often into a different address) into the interrupt vector associated with a host in hosts 110-112. Once translated, operation 200 includes transferring (203) a second interrupt using the second interrupt vector to a host of hosts 110-112 that corresponds to the second interrupt vector.

[0018] As an illustrative example, PCIe device 132 may deliver a first interrupt with a first address to PCIe switch 120. In response to the request, PCIe switch 120 may, based at least on the first address provided in the vector, translate the first interrupt vector into a second interrupt vector, with a different address, that was provided from one of hosts 110-112. Once translated, a second interrupt may be transferred using the second interrupt vector to the appropriate host communicatively coupled to PCIe switch 120. In some

implementations, in making the translation of the interrupt vectors, PCIe switch 120 may use an offset to determine the interrupt vector associated with the host. For example, if a device provided an address “BBB80000,” PCIe switch 120 may apply an offset to the first portion (“BBB8”) and use the value after the offset to determine interrupt vector address for the host. Once the address is calculated using the offset, the host interrupt vector may be provided to the associated host.

[0019] In other implementations, the one or more data structures on PCIe switch 120 may match the virtual interrupt vectors to the corresponding interrupt vectors from the host. Accordingly, when a virtual interrupt vector is received from a PCIe device, at least one data structure may be searched to determine the associated host interrupt vector. Once identified, the interrupt vector may be provided to the associated host.

[0020] Figure 3 illustrates a configuration 300 of a PCIe switch and PCIe devices according to an implementation. Configuration 300 uses systems and elements from computing environment 100 of Figure 1, and further includes data structures 310-312 to be located on PCIe switch 120 and data structures 320-323 to be located on PCIe devices 130-133.

[0021] In operation, as hosts are connected to PCIe switch 120 and/or new PCIe devices are coupled to PCIe switch 120, data structures 310-312 are generated and updated to reflect the currently connected PCIe devices. In the particular example, of configuration 300, host 112 provides information to maintain data structure 312. This information includes interrupt vectors represented by individual rows in data structure 312, wherein the interrupt vectors comprise a data portion consisting of a first value and an address portion composing an upper and a lower value. As depicted, when the interrupt values are provided, the values are associated with particular PCIe devices. Thus, one or more vectors

may be provided for each PCIe device in PCIe devices 130-133, wherein host 112 may identify the PCIe device associated with the interrupt vector based on the data value and address value for the vector.

[0022] As the interrupt vector values are provided from host 112, PCIe switch 120 further identifies virtual interrupt vectors and provides the virtual interrupt vectors to the PCIe devices. By providing the virtual interrupt vectors, PCIe switch 120 may appear as a “host” to the connected PCIe devices, and not require the individual PCIe devices to direct interrupts to a particular host of hosts 110-112. Here, data structure 320 is provided as an example, wherein virtual interrupt vectors are generated that correspond to the host interrupt vectors provided from host 112. In the present example, the first address portion “FFEE” is modified to “BBB8,” to generate the virtual interrupt vectors, however, it should be understood that any portion of the address may be modified to generate the virtual interrupt vectors. In some implementations, in generating the address for the virtual interrupt vectors, the PCIe switch 120 may apply an offset to the address provided from the host. However, it should be understood that the address for the virtual interrupt vector may be generated in any manner, such as allocating predefined addresses, generating new addresses as required, or some other operation. Additionally, while the data portion provided from host 112 is supplied to data structure 320, it should be understood that PCIe switch 120 may modify data value in some examples. As a result, when an interrupt is received from the PCIe device that corresponds to a virtual interrupt vector, the address and the data value of the virtual interrupt vector may be translated into the values expected by the associated host.

[0023] As an illustrative example, when PCIe device 130 requires an interrupt of host 112, an interrupt using a first interrupt vector may be provided to PCIe switch 120. For example, a first interrupt vector with a data value of “4A00” may be used with an address of

“BBB80000”. In response to receiving the interrupt, PCIe switch 120 will determine a second interrupt vector associated with a host of hosts 110-112. Once determined, the second interrupt vector may be used in providing an interrupt to the appropriate host of hosts 110-112. In translating the first interrupt vector, PCIe switch 120 may translate the address portion of the vector from “BBB80000” to “FFEE0000,” and provide an interrupt to host 112 using the translated address.

[0024] Although illustrated in the example of configuration 300 as configuring data structures 312 and data structure 320, it should be understood that similar operations may be applied in the creation and modification of data structures 310-311 and data structures 321-323. Further, while demonstrated in the example of Figure 3 as using table data structures to manage the interrupt vectors, it should be understood that one or more data structure, including trees, linked lists, or some other data structure may be used in managing the interrupt vectors.

[0025] Figure 4 illustrates an operational scenario 400 of managing interrupts from PCIe devices according to an implementation. Operational scenario 400 includes systems and elements from computing environment 100 of Figure 1. Although not illustrated in the example of operational scenario 400, it should be understood that one or more data structures may be maintained by PCIe switch 120 and each PCIe device in PCIe devices 130-133.

[0026] As depicted in operational scenario 400, PCIe device 133 transfers, at step 1, a first interrupt corresponding to a first interrupt vector to PCIe switch 120, wherein the interrupt vector may include at least a data portion and an address portion. When the interrupt is received, PCIe device 120, translates, at step 2, the first interrupt vector into second interrupt vector. This second interrupt vector may include a different addressing

portion, and may further include a different data portion in some examples. In some implementations, to translate the first interrupt vector into the second interrupt vector, PCIe switch 120 may maintain at least one table or some other data structure to identify the second interrupt vector associated with the first interrupt vector. In other implementations, an offset may be applied to the address in the first interrupt vector to generate the second interrupt vector. Once the second address is determined for the second interrupt vector, a second interrupt using the second interrupt vector may be communicated, at step 3, to the host associated with the second interrupt vector. Here, the second interrupt vector identified by virtual switch 120 is associated with host 111. Consequently, the second interrupt vector is forwarded to host 111, wherein the interrupt may indicate a completion command or some other similar interrupt operation with respect to the PCIe device and the corresponding host.

[0027] Figure 5 illustrates an operational scenario 500 of configuring data structures to manage interrupts from PCIe devices according to an implementation. Operational scenario 500 includes systems and elements of computing environment 100 of Figure 1, and further includes data structures 510-512, data structures 520-523. Operational scenario 500 is an example of generating interrupt vectors as required for completion commands with respect to data storage devices, such as NVMe devices.

[0028] In operational scenario 500, host 112 may generate, at step 1, a create completion queue command with a defined interrupt vector available for the host. In the present example, rather than writing the interrupt vector to the storage device immediately when the host provides the interrupt vectors to generate data structure 512, PCIe switch 120 may delay the allocation of virtual interrupt vectors until they are required by the hosts. Accordingly, in response to the request from host 112 with the vector with data value

“4A01” and address values of “FFEE” and “0001” for PCIe device 130, PCIe switch may translate the interrupt vector into the appropriate virtual interrupt vector, and forward, at step 2, the vector to be implemented in data structure 520 on the device, wherein the virtual interrupt vector includes a different addressing portion than the addressing portion on PCIe switch 120. Thus, rather than writing the table on the device immediately, the tables on the individual PCIe devices may be updated as they are required for the hosts. This would permit host 112 to transfer a second queue creation request associated with a second interrupt vector, and add a second virtual interrupt vector as required for communication with PCIe device 130 or any other PCIe device of PCIe devices 131-133.

10 [0029] Once the virtual interrupt vectors are provided to the PCIe devices, the devices may then communicate the interrupts using the provided interrupt vectors when required to the virtual switch, which can then be forwarded to the appropriate host. Using the example provided in Figure 5, PCIe device 130 may provide an interrupt corresponding to a vector with “4A01” and addressing portions “BBB8” and “0001.” Responsive to the interrupt, PCIe switch 120 may translate the interrupt vector into a second interrupt vector corresponding to the host. This may be accomplished using the offset and/or data structure provided on the PCIe switch. Once translated the second interrupt vector may be provided to a corresponding host. Again, using the example in Figure 5, a data value of “4A01” may be written to an address “FFEE0001.” In some examples, by using an offset with the addressing portion, the switch may cause minimal delay in forwarding the interrupt to the desired host. In particular, software or logic of the switch may identify addressing attributes of the first vector, apply logic to the addressing attributes of the first vector to generate second addressing attributes, and forward a second vector with the second addressing attributes to the corresponding host.

[0030] In some implementations, PCIe switch 120 may be responsible for limiting the virtual interrupt vectors that are provided to each of the devices. In particular, to ensure that interrupt vectors are available on each of the devices, PCIe switch 120 may be configured with limits to the number of vectors available to each of the hosts. In some
5 implementations, the same number of vectors may be available to each of the hosts. In other implementations, each of the hosts may be provided with various limits to the number of vectors available. For example, the hosts may be provided different numbers of interrupt vectors based on a quality of service associated with the host, which may be determined based on the application on the host, the user of the host, or some other similar quality of
10 service determination.

[0031] Figure 6 illustrates a PCIe switch computing system 600 to manage interrupts for PCIe devices according to an implementation. Computing system 600 is representative of any computing system or systems with which the various operational architectures, processes, scenarios, and sequences disclosed herein for a PCIe switch can be implemented.
15 Computing system 600 is an example of PCIe switch 120, although other examples may exist. Computing system 600 comprises communication interface 601, user interface 602, and processing system 603. Processing system 603 is linked to communication interface 601 and user interface 602. Processing system 603 includes processing circuitry 605 and memory device 606 that stores operating software 607. Computing system 600 may include
20 other well-known components such as a battery and enclosure that are not shown for clarity.

[0032] Communication interface 601 comprises components that communicate over communication links, such as network cards, ports, radio frequency (RF), processing circuitry and software, or some other communication devices. Communication interface 601 may be configured to communicate over metallic, wireless, or optical links.

Communication interface 601 may be configured to use Time Division Multiplex (TDM), Internet Protocol (IP), Ethernet, PCIe, optical networking, wireless protocols, communication signaling, or some other communication format – including combinations thereof. Communication interface 601 is configured to communicate with a plurality of
5 host systems, such as serving computers, and a plurality of end PCIe devices, such as storage devices, co-processing units, graphical processing units, or some other similar PCIe device.

[0033] User interface 602 comprises components that interact with a user to receive user inputs and to present media and/or information. User interface 602 may include a
10 speaker, microphone, buttons, lights, display screen, touch screen, touch pad, scroll wheel, communication port, or some other user input/output apparatus – including combinations thereof. User interface 602 may be omitted in some examples.

[0034] Processing circuitry 605 comprises microprocessor and other circuitry that retrieves and executes operating software 607 from memory device 606. Memory device
15 606 may include volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information, such as computer readable instructions, data structures, program modules, or other data. Memory device 606 may be implemented as a single storage device, but may also be implemented across multiple storage devices or sub-systems. Memory device 606 may comprise additional
20 elements, such as a controller to read operating software 607. Examples of storage media include random access memory, read only memory, magnetic disks, optical disks, and flash memory, as well as any combination or variation thereof, or any other type of storage media. In some implementations, the storage media may be a non-transitory storage media.

In some instances, at least a portion of the storage media may be transitory. It should be understood that in no case is the storage media a propagated signal.

[0035] Processing circuitry 605 is typically mounted on a circuit board that may also hold memory device 606 and portions of communication interface 601 and user interface 5 602. Operating software 607 comprises computer programs, firmware, or some other form of machine-readable program instructions. Operating software 607 includes maintain module 608, identify module 609, and translate module 610, although any number of software modules may provide similar operations. Operating software 607 may further include an operating system, utilities, drivers, network interfaces, applications, or some 10 other type of software. When executed by processing circuitry 605, operating software 607 directs processing system 603 to operate computing system 600 as described herein.

[0036] In at least one implementation, maintain module 608 is used to maintain one or more data structures for interrupt vectors from hosts communicatively coupled to computing system 600 using communication interface 601. In particular, when a host or a 15 new PCIe device is coupled to computing system 600, the hosts may be responsible for providing interrupt vectors, which are used by a PCIe device when an interrupt is required, such as a completion command interrupt. In addition to receiving the interrupt vectors from the hosts, maintain module 608 is further responsible for translating the interrupt vectors into virtual interrupt vectors that can be provided to individual PCIe devices. For example, 20 rather than directly providing the interrupt vectors from the hosts, which may include duplicates when received from multiple hosts, each of the interrupt vectors may be translated into a unique virtual interrupt vector. This translation may be accomplished via an offset that can be applied to at least a portion of the addressing for the interrupt vector, may be replacement values that are used to replace a portion of the interrupt vectors (such

as the address of the vectors), or may comprise any other mechanism to provide a unique virtual interrupt vector to the corresponding PCIe device. Once provided, the PCIe device may cache the virtual interrupt vectors and use the virtual interrupt vectors when service is required with a host.

5 **[0037]** As the interrupt vectors and the virtual interrupt vectors are maintained, identify module 609 directs processing system 603 to identify or receive a first interrupt corresponding to a first interrupt vector from a PCIe device, wherein the interrupt vector comprises at least a virtual address and a first data value. In response to receiving the first interrupt via communication interface 601, translate module 610 directs processing system
10 603 to translate the first interrupt vector into a second interrupt vector, wherein the second interrupt vector comprises a second address and a second data value. In some implementations, computing system 600 may maintain one or more data structures that can be used to match attributes of the first interrupt vector (virtual interrupt vector) with attributes of the second interrupt vector. In other implementations, computing system 600
15 may apply an offset to at least a portion of the virtual address to derive the second address for the second interrupt vector. In some examples, the second data value may be an equal value to that supplied by the first interrupt vector. However, it should be understood that the second data value may be a different value than the first value in some examples.

20 **[0038]** Once the second interrupt vector is determined via the translation process, translation module 610 further directs processing system 603 to transfer a second interrupt using the second interrupt vector to a host that corresponds to the second interrupt vector. In at least one implementation, at least one table (or some other data structure) may be used to translate the first address and the first value from the first interrupt vector, into a second address and a second value for the second interrupt vector. Additionally, based on the

identified values for the second interrupt vector, computing system 600 may identify a relevant host from a plurality of hosts coupled to the switch, and transfer the required interrupt to the identified host.

[0039] Although demonstrated in the examples herein as translating a first address in the first interrupt vector to a second address in the second interrupt vector, it should be understood that the address may not necessarily be the value that is translated by the PCIe switch. In particular, rather than translating the address fields for the interrupt vector, the data value may be modified in a way to distinctly identify interrupts for each of the hosts. Consequently, when a first interrupt vector is received from a PCIe device, the translation mechanism of the PCIe switch may translate the data value from the first interrupt vector into a second data value for a second interrupt vector. Once translated, the second interrupt vector may be forwarded to the associated host.

[0040] Returning to the elements of Figure 1, hosts 110-112 may each comprise communication interfaces, network interfaces, processing systems, computer systems, microprocessors, storage systems, storage media, or some other processing devices or software systems, and can be distributed among multiple devices. Hosts 110-112 can each include software such as an operating system, logs, databases, utilities, drivers, networking software, and other software stored on a computer-readable medium. Hosts 110-112 may each comprise a serving computing system, a desktop computing system, or some other similar computing system capable of PCIe communication with supported devices.

[0041] PCIe switch 120 may comprise communication interfaces, network interfaces, processing systems, computer systems, microprocessors, storage systems, storage media, or some other processing devices or software systems, and can be distributed among multiple devices. Examples of virtual computing service 150 can include software such as an

operating system, logs, databases, utilities, drivers, networking software, and other software stored on a computer-readable medium.

[0042] PCIe devices 130-133 may comprise storage devices, such as NVMe storage devices, graphical processing units, co-processing units, networking card, or some other
5 similar end PCIe device. PCIe devices 130-133 may each comprise one or more communication interfaces, network interfaces, processing systems, microprocessors, storage systems, storage media, or some other processing devices or software systems.

[0043] PCIe switch 120 provides communication between hosts 110-112 and PCIe devices 130-133 using PCIe bus lanes, wherein each host of hosts 110-112 is
10 communicatively coupled to the switch using a PCIe interface, and each PCIe device of PCIe devices 130-133 is communicatively coupled to the switch using a PCIe interface.

[0044] The above description and associated figures teach the best mode of the invention. The following claims specify the scope of the invention. Note that some aspects of the best mode may not fall within the scope of the invention as specified by the claims.
15 Those skilled in the art will appreciate that the features described above can be combined in various ways to form multiple variations of the invention. As a result, the invention is not limited to the specific embodiments described above, but only by the following claims and their equivalents.

CLAIMS

What is claimed is:

1. A computing apparatus comprising:
 - one or more computer readable storage media;
 - 5 a processing system operatively coupled with the one or more computer readable storage media; and
 - program instructions stored on the one or more computer readable storage media to operate a Peripheral Component Interconnect Express (PCIe) switch that, when read and executed by the processing system, direct the processing system to at least:
 - 10 receive a first interrupt corresponding to a first interrupt vector from a PCIe device, wherein the first interrupt vector comprises at least a virtual address and a first data value;
 - translate the first interrupt vector into a second interrupt vector, wherein the second interrupt vector comprises a second address and a second data value; and
 - 15 transfer the second interrupt using the second interrupt vector to a host of a plurality of hosts that corresponds to the second interrupt vector.
2. The computing apparatus of claim 1, wherein the program instructions further direct the processing system to:
 - 20 maintain at least one data structure for the plurality of hosts, wherein the at least one data structure comprises a plurality of interrupt vectors received from the plurality of hosts for one or more PCIe devices coupled to the PCIe switch to provide interrupts to the plurality of hosts.

3. The computing apparatus of claim 2, wherein the program instructions further direct the processing system to:

for each interrupt vector in the plurality of interrupt vectors received from the plurality of hosts, provide a virtual interrupt vector to a corresponding PCIe device in the one or more PCIe devices coupled to the PCIe switch, wherein the virtual interrupt vector comprises at least a different address than the interrupt vector.

4. The computing apparatus of claim 1, wherein the PCIe device comprises a data storage device.

10

5. The computing apparatus of claim 1, wherein the PCIe device comprises a Non-Volatile Memory express (NVMe) device.

6. The computing apparatus of claim 1, wherein the program instructions to translate the first interrupt vector into a second interrupt vector direct the processing system to apply an addressing offset to the first interrupt vector to generate the second interrupt vector.

15

7. The computing apparatus of claim 1, wherein the first data value is equivalent to the second data value.

20

8. A system comprising:
one or more Peripheral Component Interconnect Express (PCIe) devices; and
a PCIe switch configured to:

receive a first interrupt corresponding to a first interrupt vector from a PCIe device, wherein the first interrupt vector comprises at least a virtual address and a first data value;

translate the first interrupt vector into a second interrupt vector, wherein the
5 second interrupt vector comprises a second address and a second data value; and

transfer a second interrupt using the second interrupt vector to a host of a plurality of hosts that corresponds to the second interrupt vector.

9. The system of claim 8, wherein the PCIe switch is further configured to maintain at
10 least one data structure for the plurality of hosts, wherein the at least one data structure comprises a plurality of interrupt vectors received from the plurality of hosts for one or more PCIe devices coupled to the PCIe switch to provide interrupts to the plurality of hosts.

10. The system of claim 9, wherein the PCIe switch is further configured to, for each
15 interrupt vector in the plurality of interrupt vectors received from the plurality of hosts, provide a virtual interrupt vector to a corresponding PCIe device in the one or more PCIe devices coupled to the PCIe switch, wherein the virtual interrupt vector comprises at least a different address than the interrupt vector.

20 11. The system of claim 8, wherein the PCIe device comprises a data storage device.

12. The system of claim 8, wherein the PCIe switch configured to translate the first interrupt vector into a second interrupt vector is configured to apply an addressing offset to the first interrupt vector to generate the second interrupt vector.

25

13. The system of claim 8, wherein the first data value is different than the second data value.
14. An apparatus comprising:
- 5 one or more computer readable storage media; and
- program instructions stored on the one or more computer readable storage media to operate a Peripheral Component Interconnect Express (PCIe) switch that, when read and executed by a processing system, direct the processing system to at least:
- receive a first interrupt corresponding to a first interrupt vector from a PCIe
- 10 device, wherein the first interrupt vector comprises at least a virtual address and a first data value;
- translate the first interrupt vector into a second interrupt vector, wherein the second interrupt vector comprises a second address and a second data value; and
- transfer a second interrupt using the second interrupt vector to a host of a
- 15 plurality of hosts that corresponds to the second interrupt vector.
15. The apparatus of claim 14, wherein the program instructions further direct the processing system to:
- maintain at least one data structure for the plurality of hosts, wherein the at least one
- 20 data structure comprises a plurality of interrupt vectors received from the plurality of hosts for one or more PCIe devices coupled to the PCIe switch to provide interrupts to the plurality of hosts; and
- for each interrupt vector in the plurality of interrupt vectors received from the plurality of hosts, provide a virtual interrupt vector to a corresponding PCIe device in the

one or more PCIe devices coupled to the PCIe switch, wherein the virtual interrupt vector comprises at least a different address than the interrupt vector.

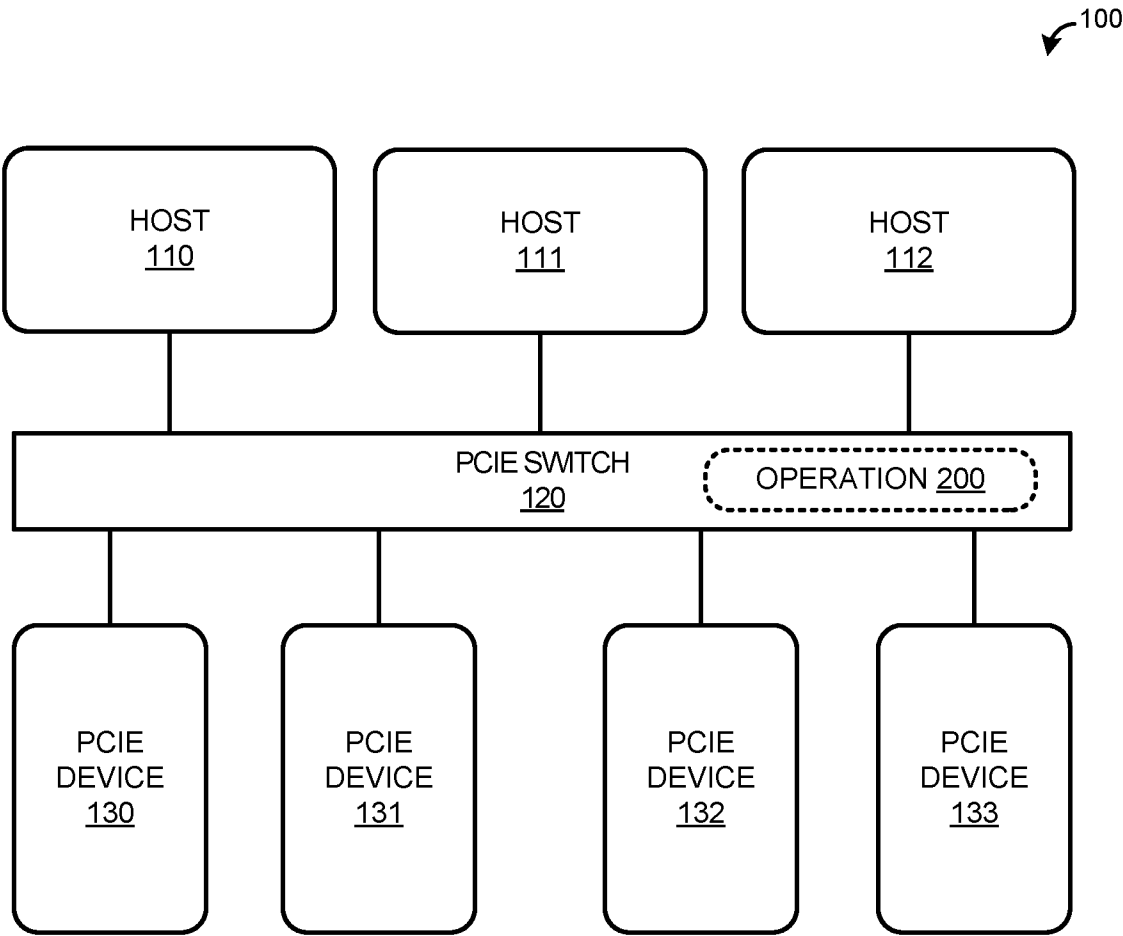
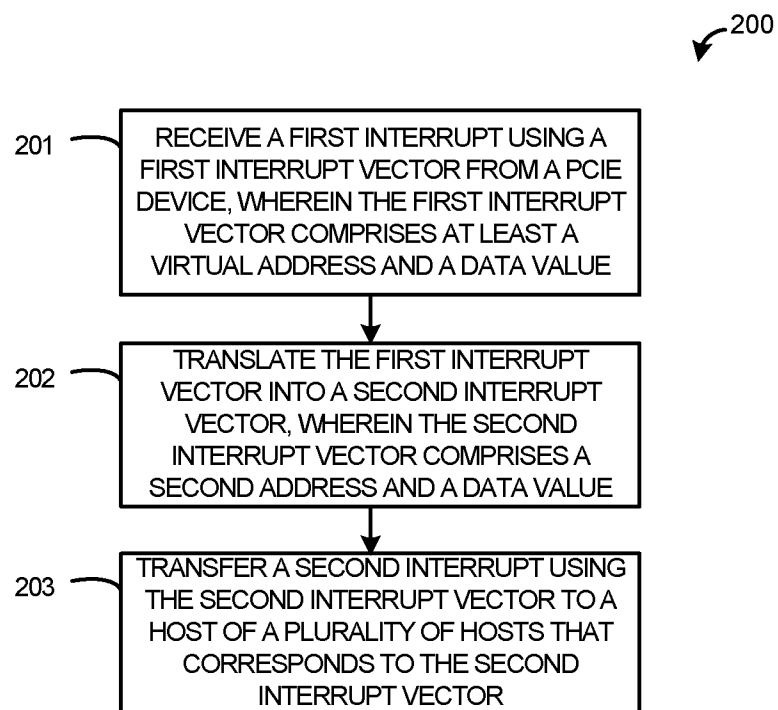


FIGURE 1

**FIGURE 2**

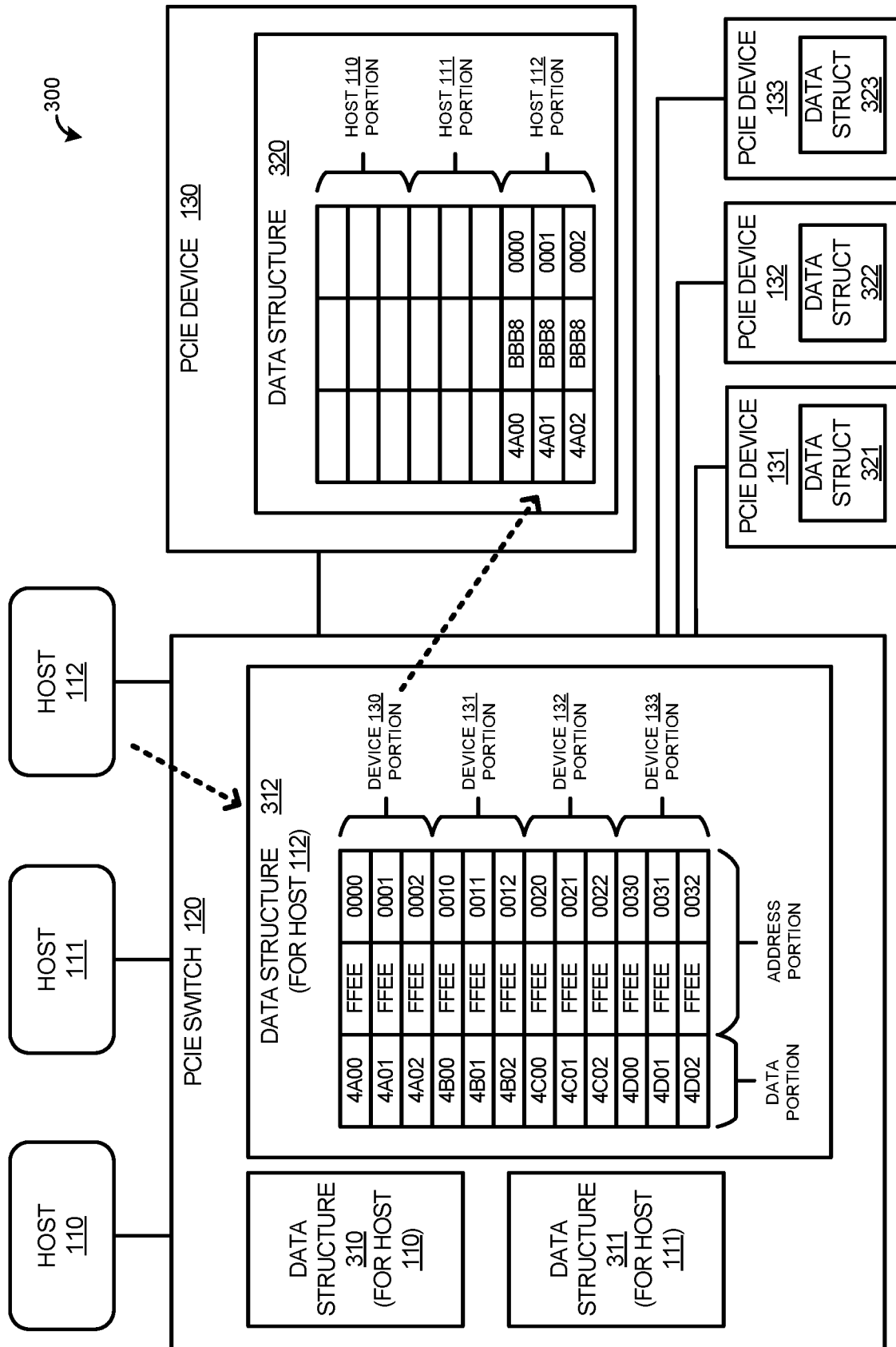


FIGURE 3

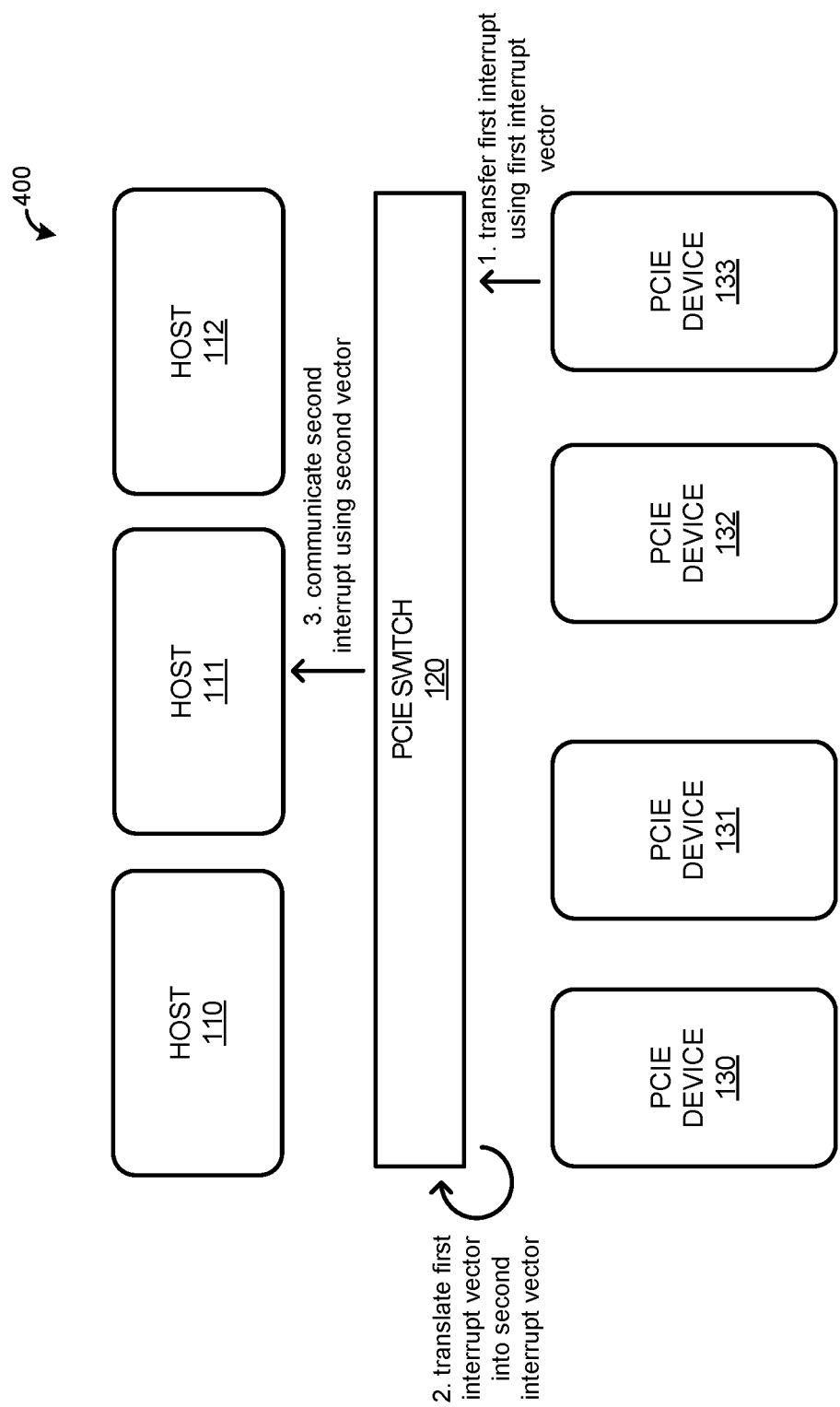


FIGURE 4

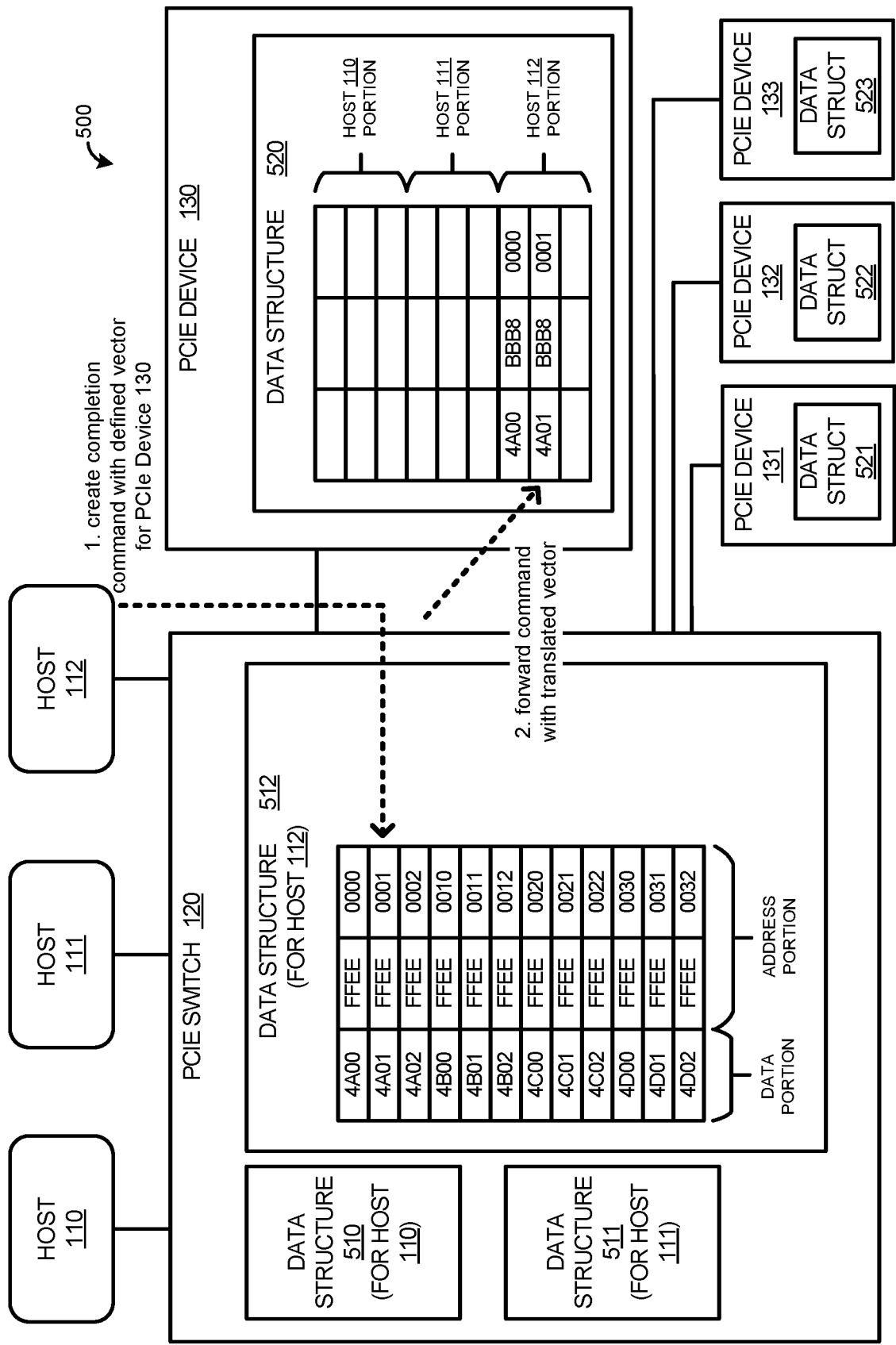


FIGURE 5

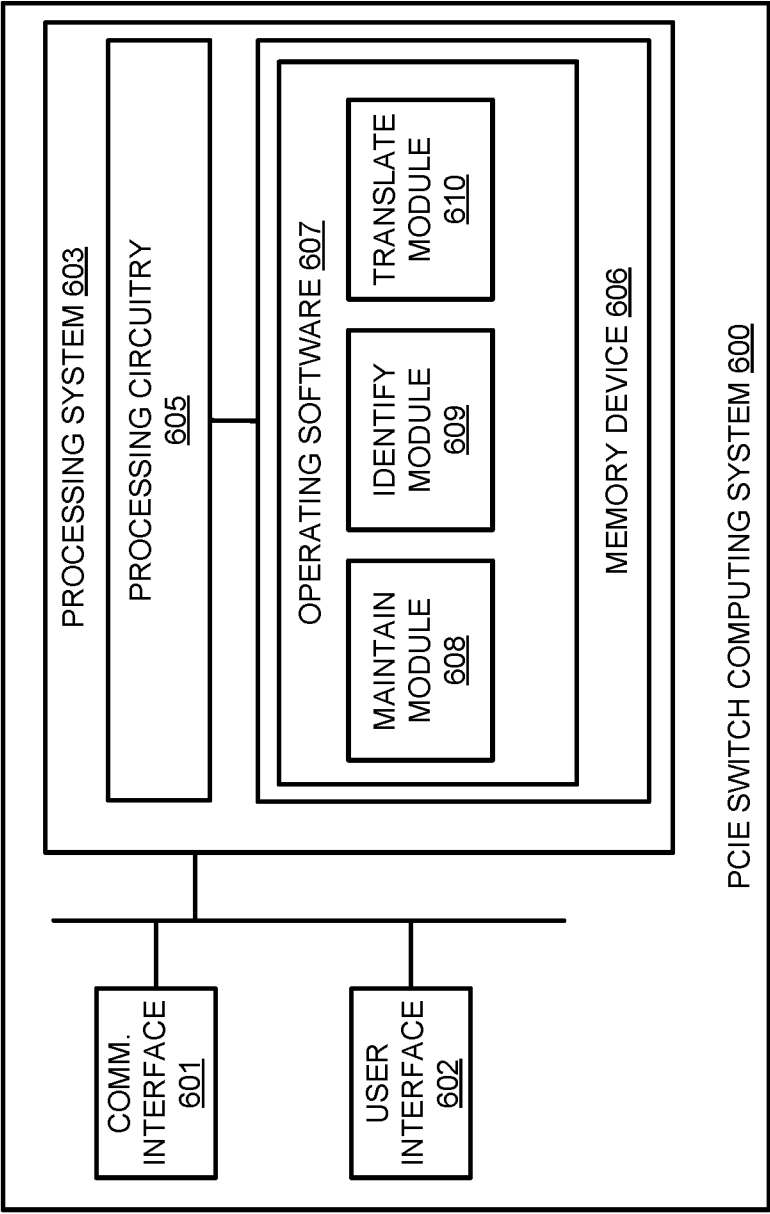


FIGURE 6

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2018/022181

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F13/40
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

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

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

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2014/181354 A1 (YI CHANG [CN] ET AL) 26 June 2014 (2014-06-26) abstract paragraphs [0003] - [0006]; figure 1 -----	1-15
A	US 2015/356038 A1 (FEEHRER JOHN R [US] ET AL) 10 December 2015 (2015-12-10) abstract; figure 4 paragraph [0007] paragraphs [0066] - [0069] -----	1-15
A	US 2009/248947 A1 (MALWANKAR KIRON [US] ET AL) 1 October 2009 (2009-10-01) abstract; figures 4,5 paragraphs [0059] - [0069] -----	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

11 June 2018

Date of mailing of the international search report

29/06/2018

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Jünger, Bernhard

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2018/022181

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2014181354 A1	26-06-2014	CN 103023824 A	03-04-2013
		EP 2768188 A1	20-08-2014
		US 2014181354 A1	26-06-2014
		WO 2014089980 A1	19-06-2014

US 2015356038 A1	10-12-2015	NONE	

US 2009248947 A1	01-10-2009	US 2009248947 A1	01-10-2009
		WO 2009120798 A2	01-10-2009
