(12) UK Patent Application (19) GB (11) 2492620 (13) A

(43) Date of A Publication 09.01.2013

(21) Application No: 1208064.4

(22) Date of Filing: 09.05.2012

(30) Priority Data:
(31) 13177680 (32) 07.07.2011 (33) US

(51) INT CL:
*G06F 13/40* (2006.01)

(56) Documents Cited:
WO 2007/064466 A2 US 20100165983 A1
US 20070083707 A1 US 20020124114 A1

(58) Field of Search:
INT CL G06F, H04L, H05K
Other: WPI, EPODOC.

(71) Applicant(s):
**International Business Machines Corporation (Incorporated in USA - New York)**
**New Orchard Road, Armonk, New York 10504, United States of America**

(72) Inventor(s):
**Steven Michael Tri**
**James O'Connor**
**William Gavin Holland**
**David Engebretsen**
**James Edward Hughes**
**William Joseph Armstrong**
**Martin Joseph Crippen**
**Dhruv Manmohandas Desai**
**John Michael Borkenhagen**
**Philip Rogers Hillier III**
**Pravin Patel**

(74) Agent and/or Address for Service:
**IBM United Kingdom Limited**
**Intellectual Property Law, Hursley Park, WINCHESTER, Hampshire, SO21 2JN, United Kingdom**

(54) Title of the Invention: **Switch fabric management**
Abstract Title: **Midplane for blade server management**

(57) In a switching system (fig. 1) comprising server card "blades" (104, fig. 1 & 202, fig. 6), a midplane 123 is connected to switch modules $420_{1-3}$ on different server cards thus providing a fabric interconnect for a switch fabric under the operation of a management controller (1604, fig. 16) which updates and monitors the switch modules, and informs others when one fails. Switch modules may be replaced ("hot swapped") without restarting the system or fabric.

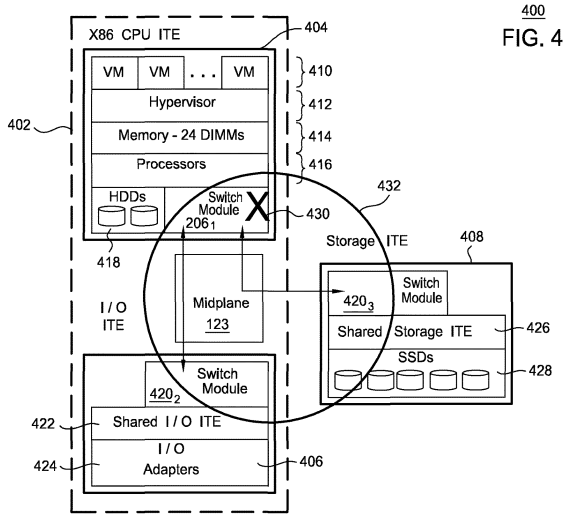



FIG. 4

GB 2492620 A

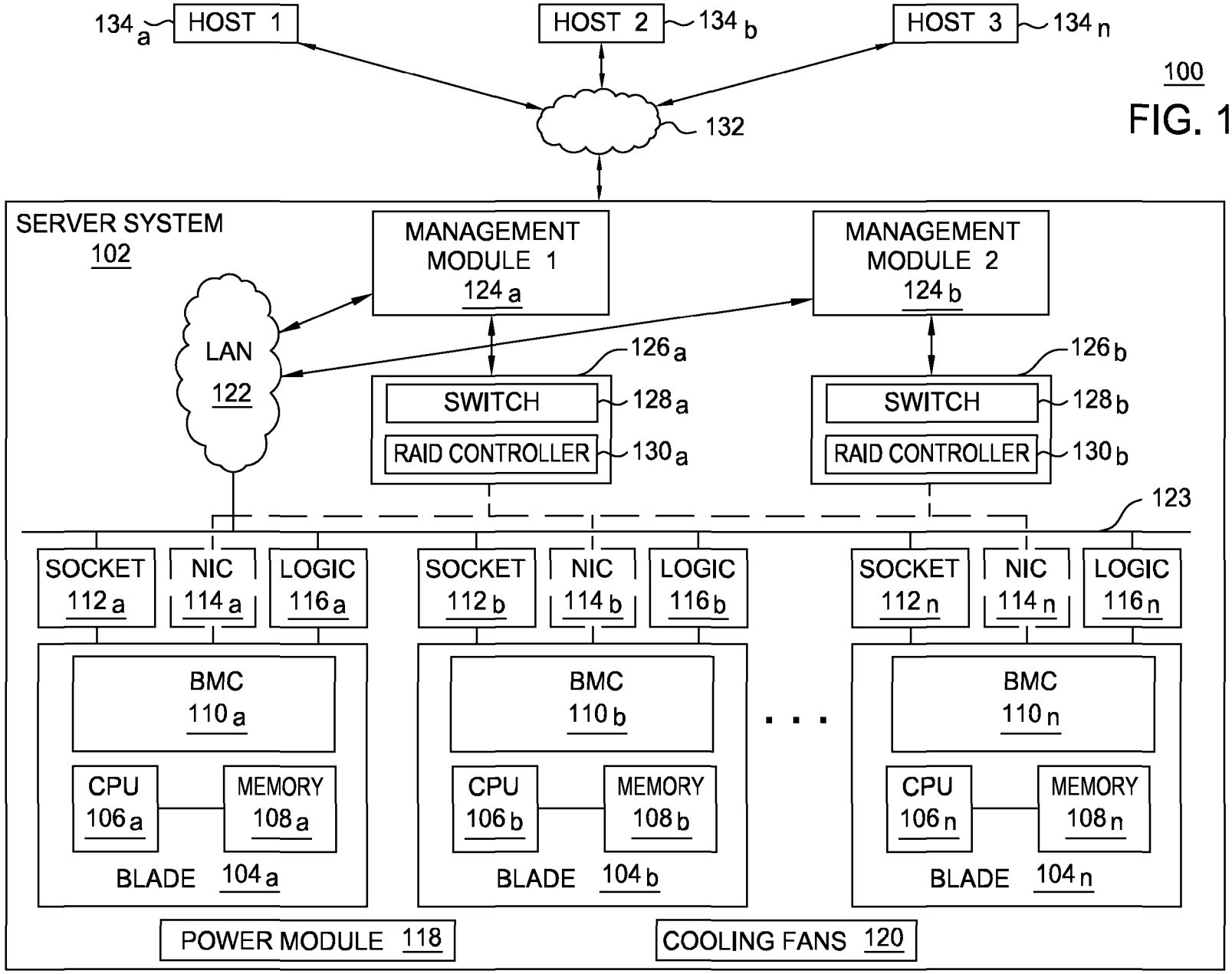

HOST 1
134 a
HOST 2
134 b
HOST 3
134 n
132
100
SERVER SYSTEM
102
MANAGEMENT MODULE 1
124 a
MANAGEMENT MODULE 2
124 b
LAN
122
126 a
126 b
SWITCH
128 a
SWITCH
128 b
RAID CONTROLLER
130 a
RAID CONTROLLER
130 b
123
SOCKET 112 a
NIC 114 a
LOGIC 116 a
SOCKET 112 b
NIC 114 b
LOGIC 116 b
SOCKET 112 n
NIC 114 n
LOGIC 116 n
BMC 110 a
BMC 110 b
BMC 110 n
CPU 106 a
MEMORY 108 a
CPU 106 b
MEMORY 108 b
CPU 106 n
MEMORY 108 n
BLADE 104 a
BLADE 104 b
BLADE 104 n
POWER MODULE 118
COOLING FANS 120

FIG. 1

FIG. 2

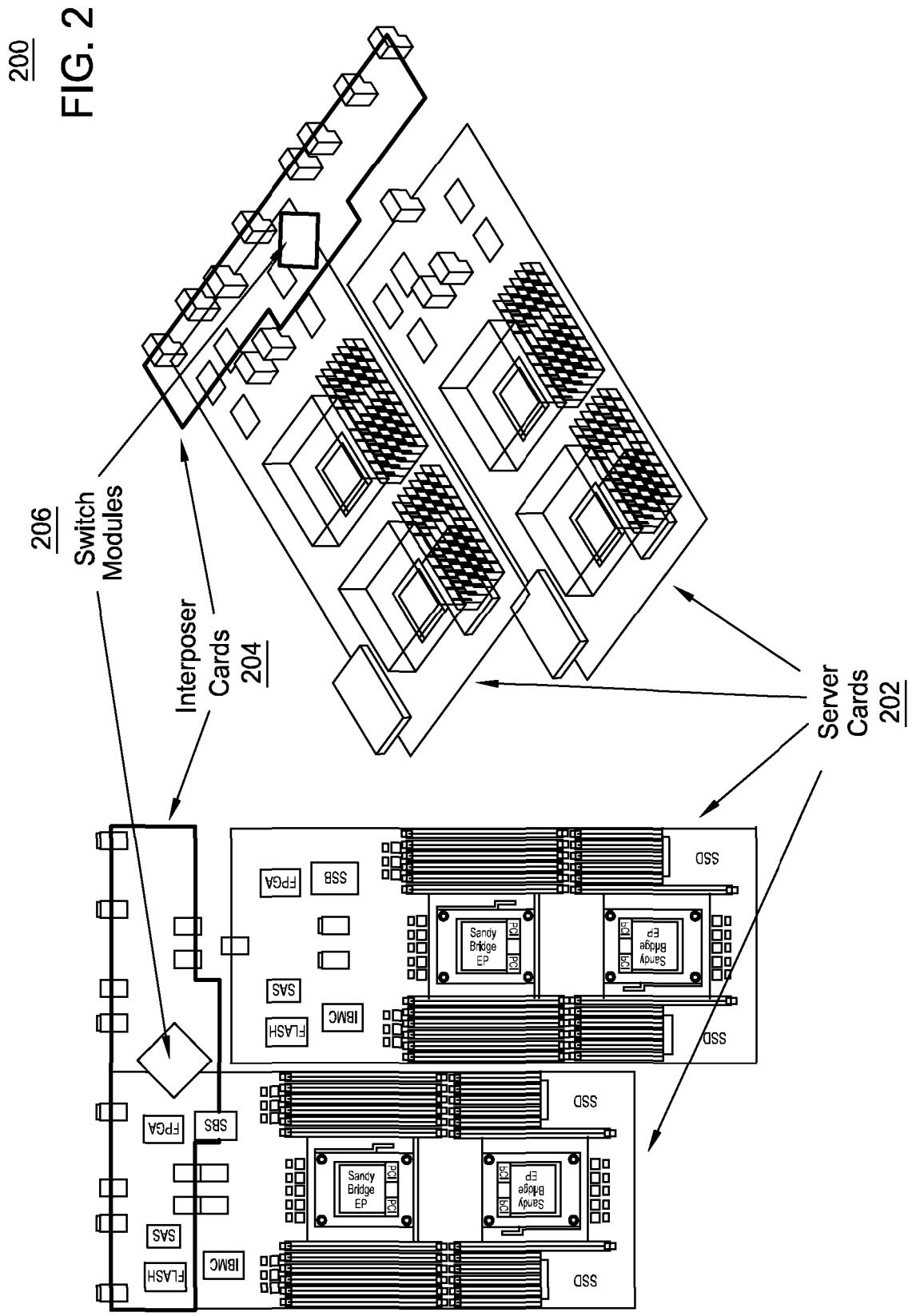
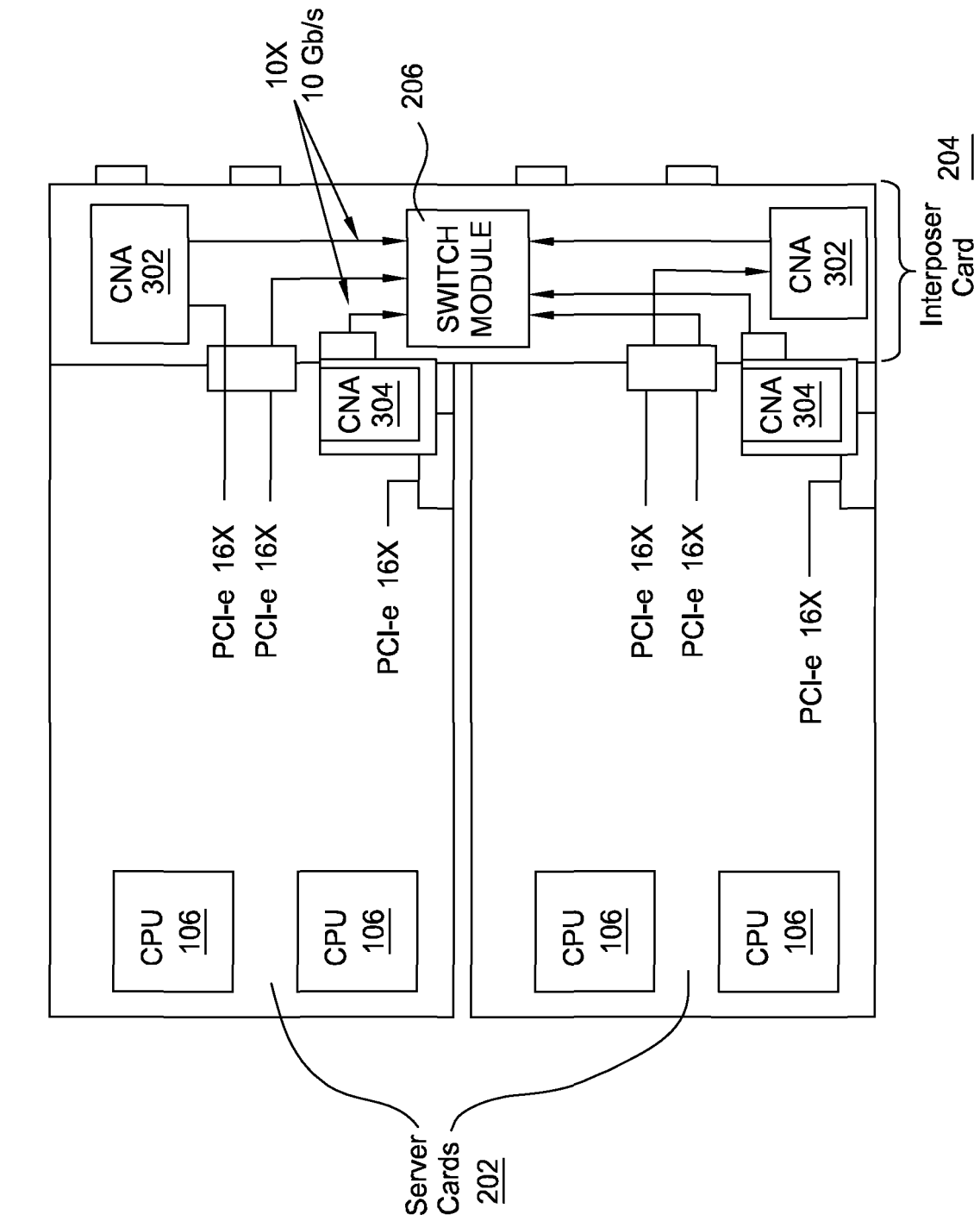

200
206
Switch Modules
Interposer Cards 204
Server Cards 202
SSD
Sandy Bridge EP
PCI
SSB
FPGA
SAS
IBMC
FLASH
SBS

FIG. 3


300
10X
10 Gb/s
206
CNA
302
SWITCH
MODULE
CNA
302
Interposer
Card
204
CNA
304
CNA
304
PCI-e 16X
PCI-e 16X
PCI-e 16X
PCI-e 16X
PCI-e 16X
PCI-e 16X
CPU
106
CPU
106
CPU
106
CPU
106
Server
Cards
202

400

FIG. 4

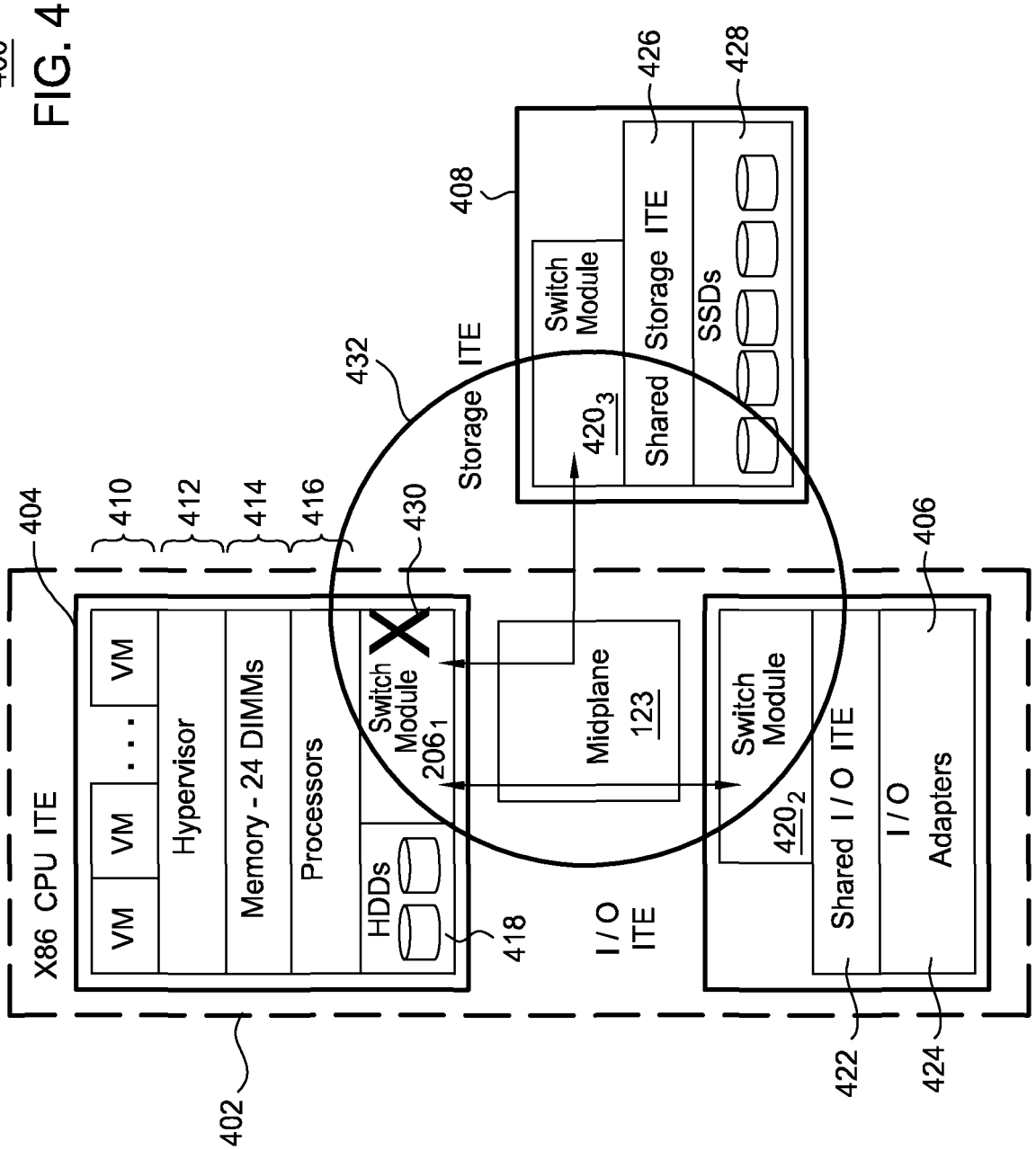
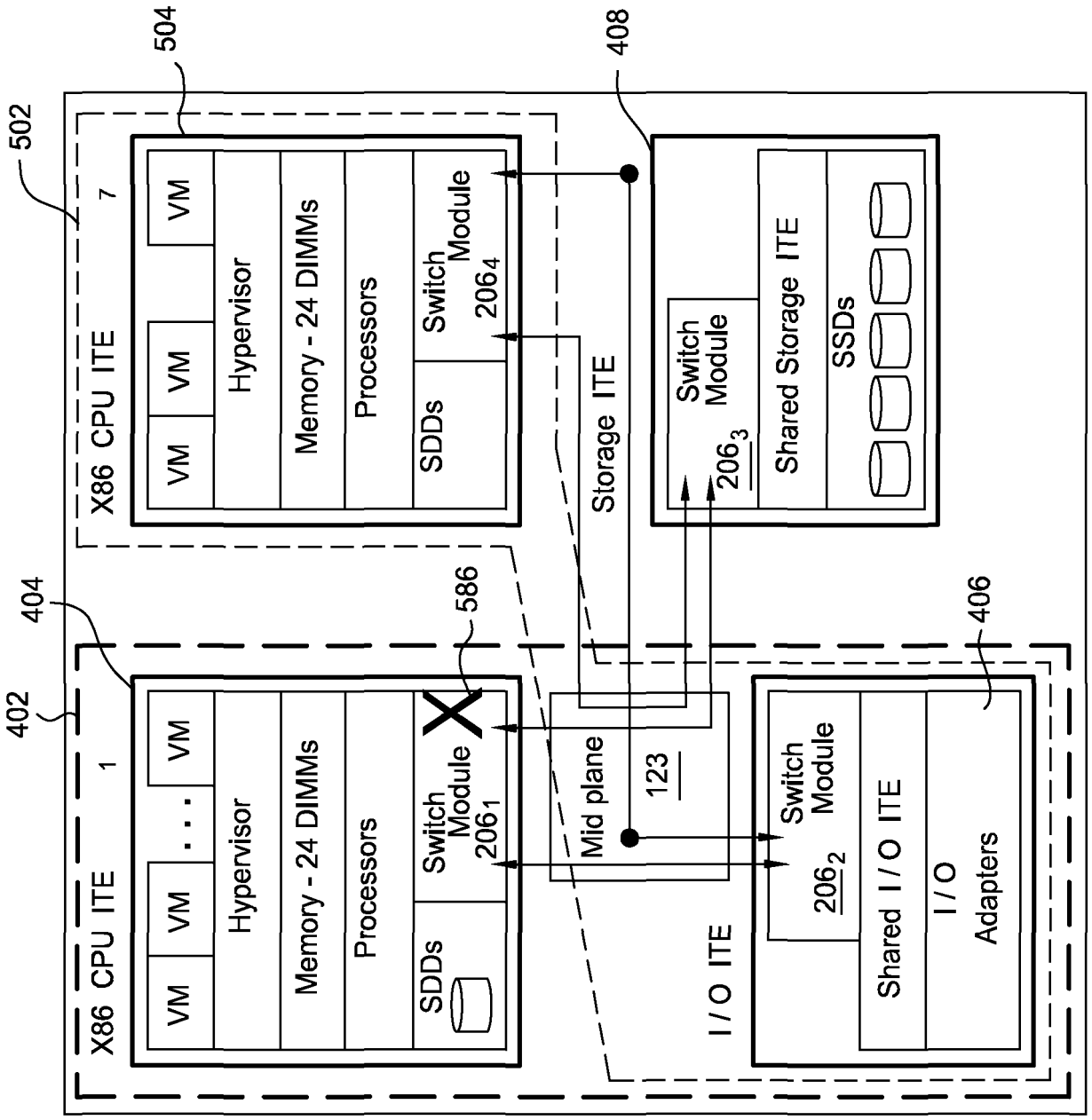
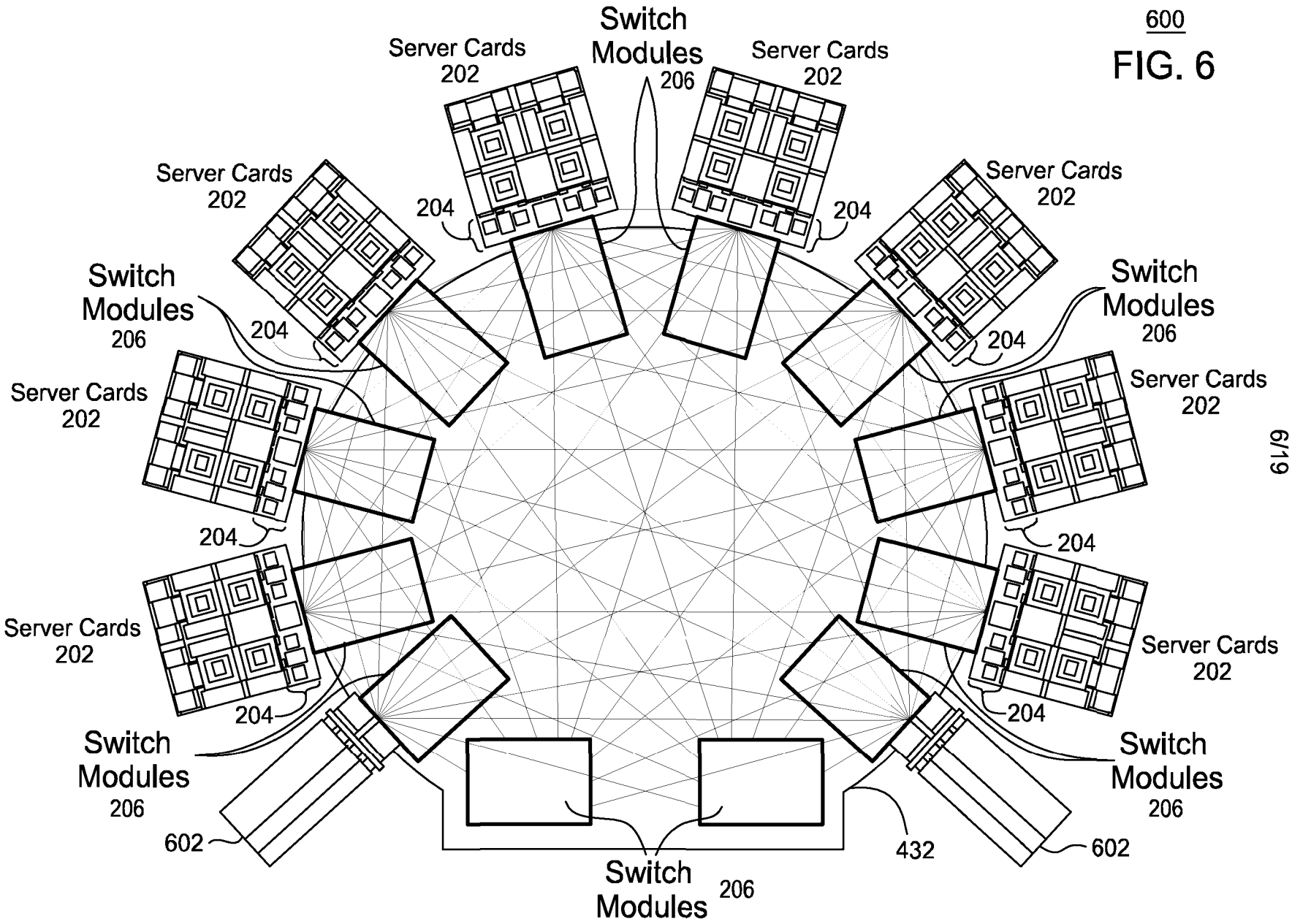

X86 CPU ITE
VM
VM
. . .
VM
Hypervisor
Memory - 24 DIMMs
Processors
HDDs
Switch Module 206 1
Midplane 123
Switch Module 420 2
Shared I / O ITE
I / O Adapters
I / O ITE
Storage ITE
Switch Module 420 3
Shared Storage ITE
SSDs
402
404
406
408
410
412
414
416
418
422
424
426
428
430
432

FIG. 5


500
502
504
408
404
586
406
402
X86 CPU ITE 7
VM
VM
VM
Hypervisor
Memory - 24 DIMMs
Processors
SDDs
Switch Module 206$_4$
Storage ITE
Switch Module 206$_3$
Shared Storage ITE
SSDs
X86 CPU ITE 1
VM
VM
VM
Hypervisor
Memory - 24 DIMMs
Processors
SDDs
Switch Module 206$_1$
Mid plane
123
I / O ITE
Switch Module 206$_2$
Shared I / O ITE
I / O Adapters

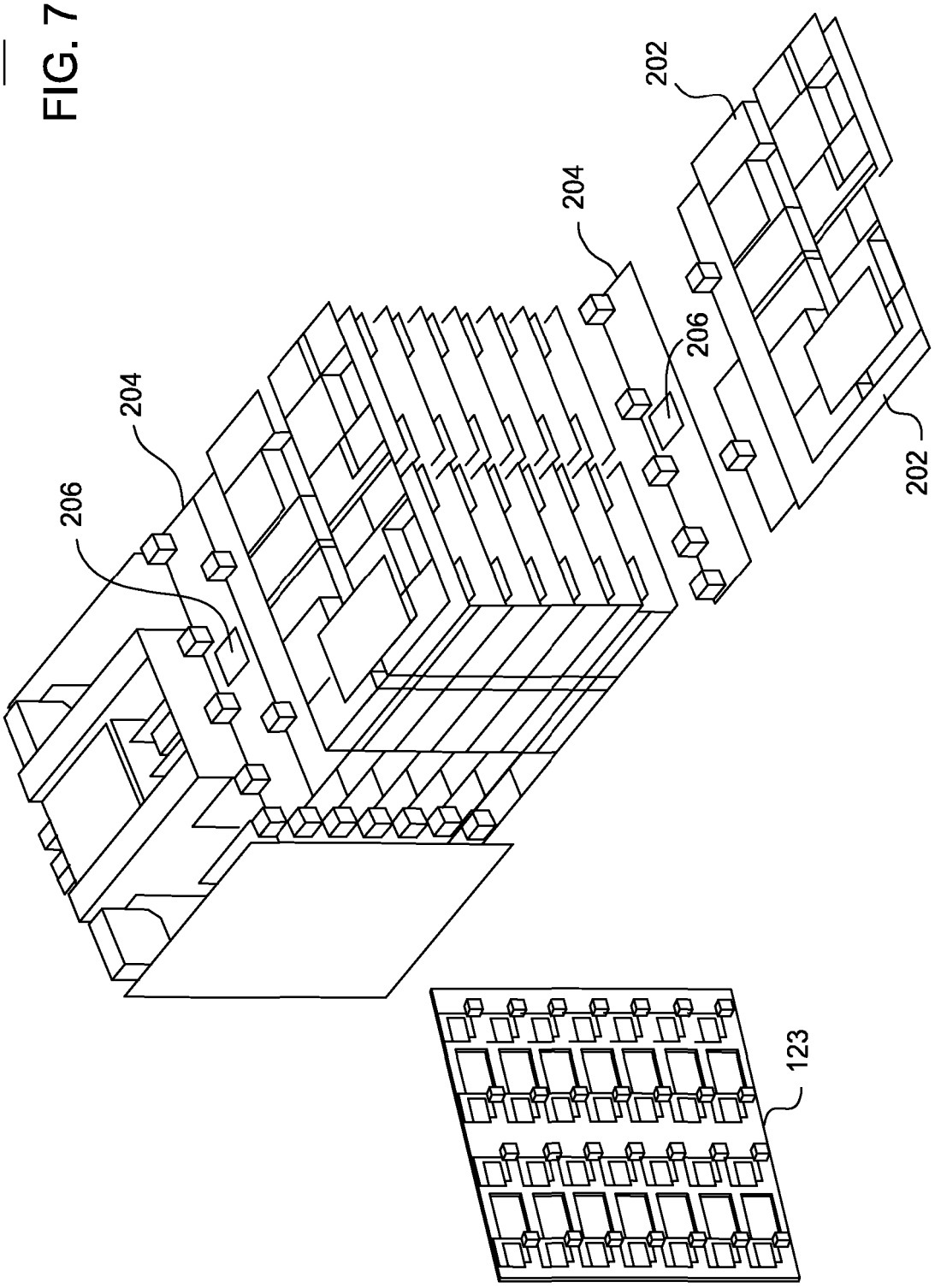
600

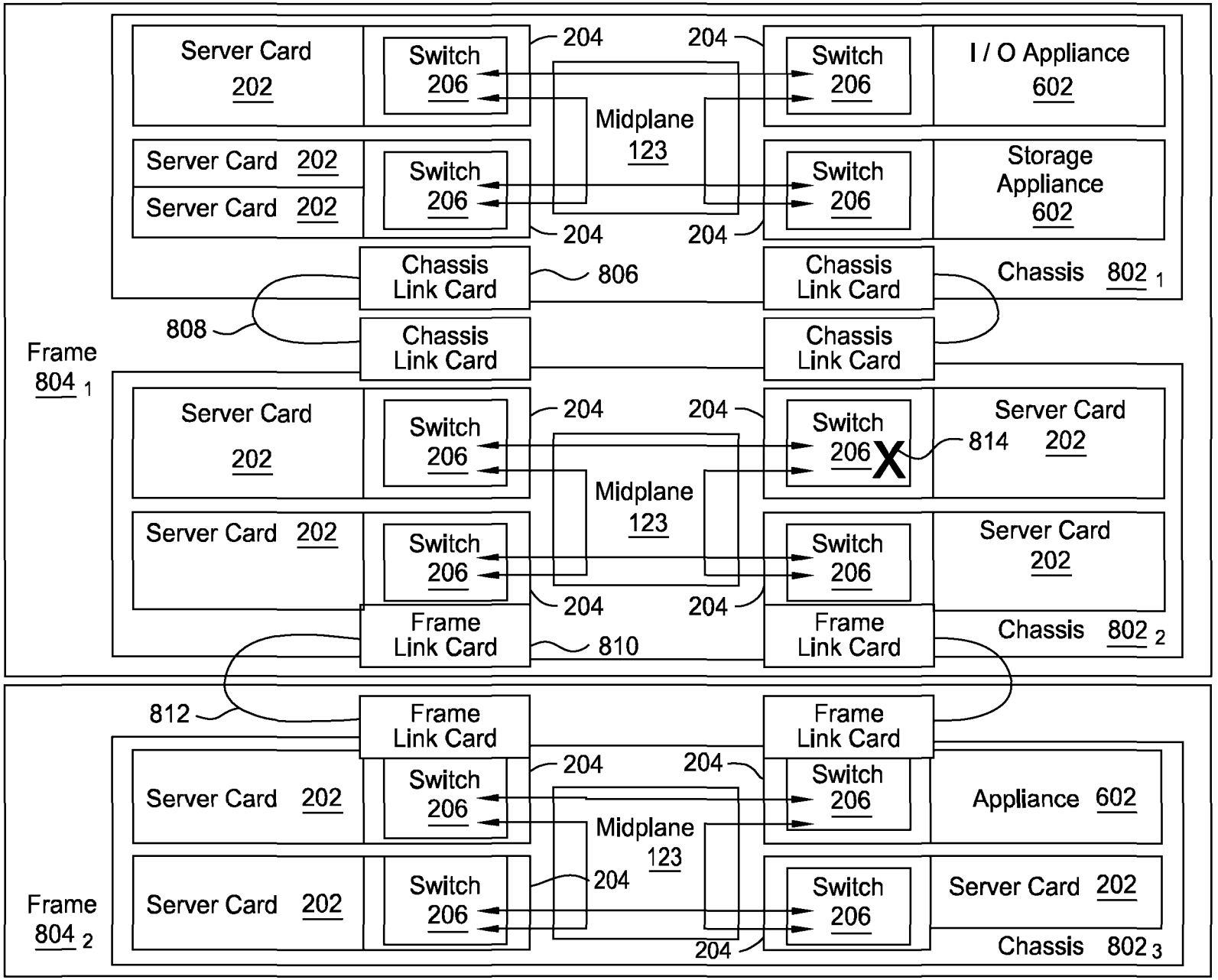
FIG. 6


Switch Modules 206
Server Cards 202
Server Cards 202
204
204
Server Cards 202
Switch Modules 206
204
Switch Modules 206
Server Cards 202
Server Cards 202
204
204
Server Cards 202
Server Cards 202
204
204
Switch Modules 206
Switch Modules 206
602
602
Switch Modules 206
432

700

FIG. 7



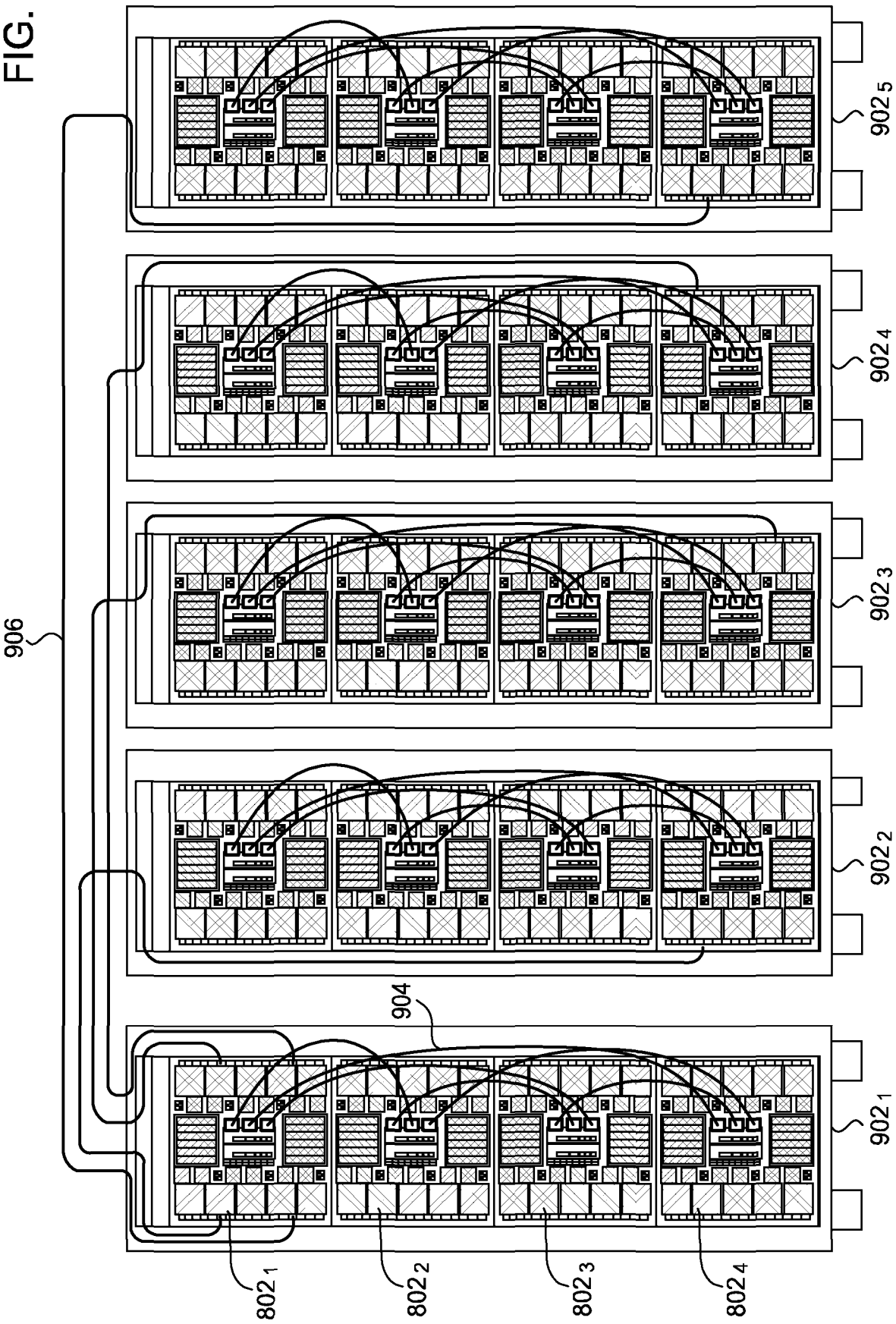

Server Card 202
Switch 206
204
Midplane 123
I / O Appliance 602
Storage Appliance 602
Chassis Link Card
806
Chassis 802 1
808
Frame 804 1
814
Chassis 802 2
Frame Link Card
810
812
Appliance 602
Chassis 802 3
Frame 804 2

800

FIG. 8

900

FIG. 9



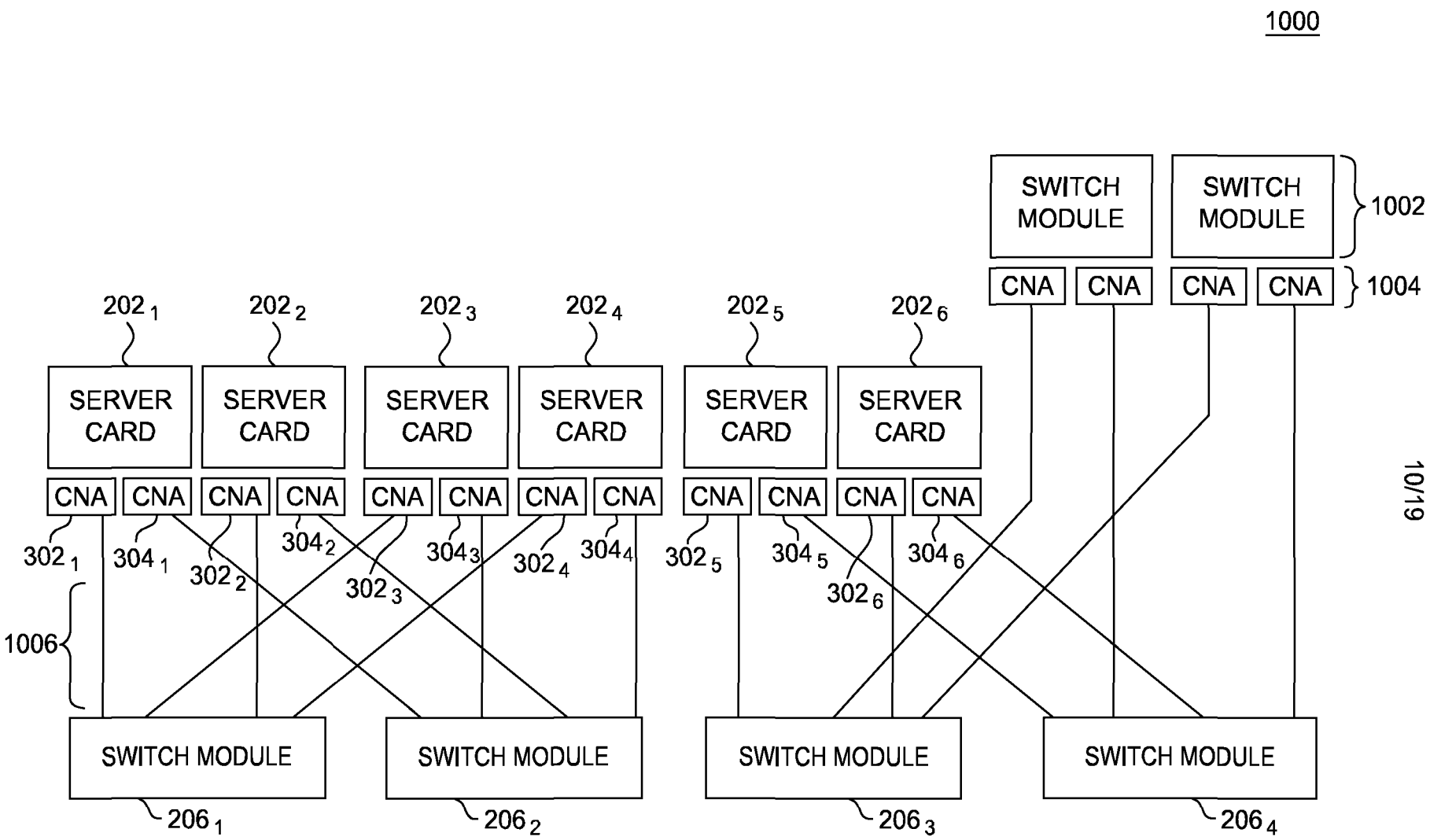

1000
SWITCH MODULE
SWITCH MODULE
1002
CNA
CNA
CNA
CNA
1004
$202_1$
$202_2$
$202_3$
$202_4$
$202_5$
$202_6$
SERVER CARD
SERVER CARD
SERVER CARD
SERVER CARD
SERVER CARD
SERVER CARD
CNA
CNA
CNA
CNA
CNA
CNA
CNA
CNA
CNA
CNA
CNA
CNA
$302_1$
$304_1$
$302_2$
$304_2$
$302_3$
$304_3$
$302_4$
$304_4$
$302_5$
$304_5$
$302_6$
$304_6$
1006
SWITCH MODULE
SWITCH MODULE
SWITCH MODULE
SWITCH MODULE
$206_1$
$206_2$
$206_3$
$206_4$

FIG. 10

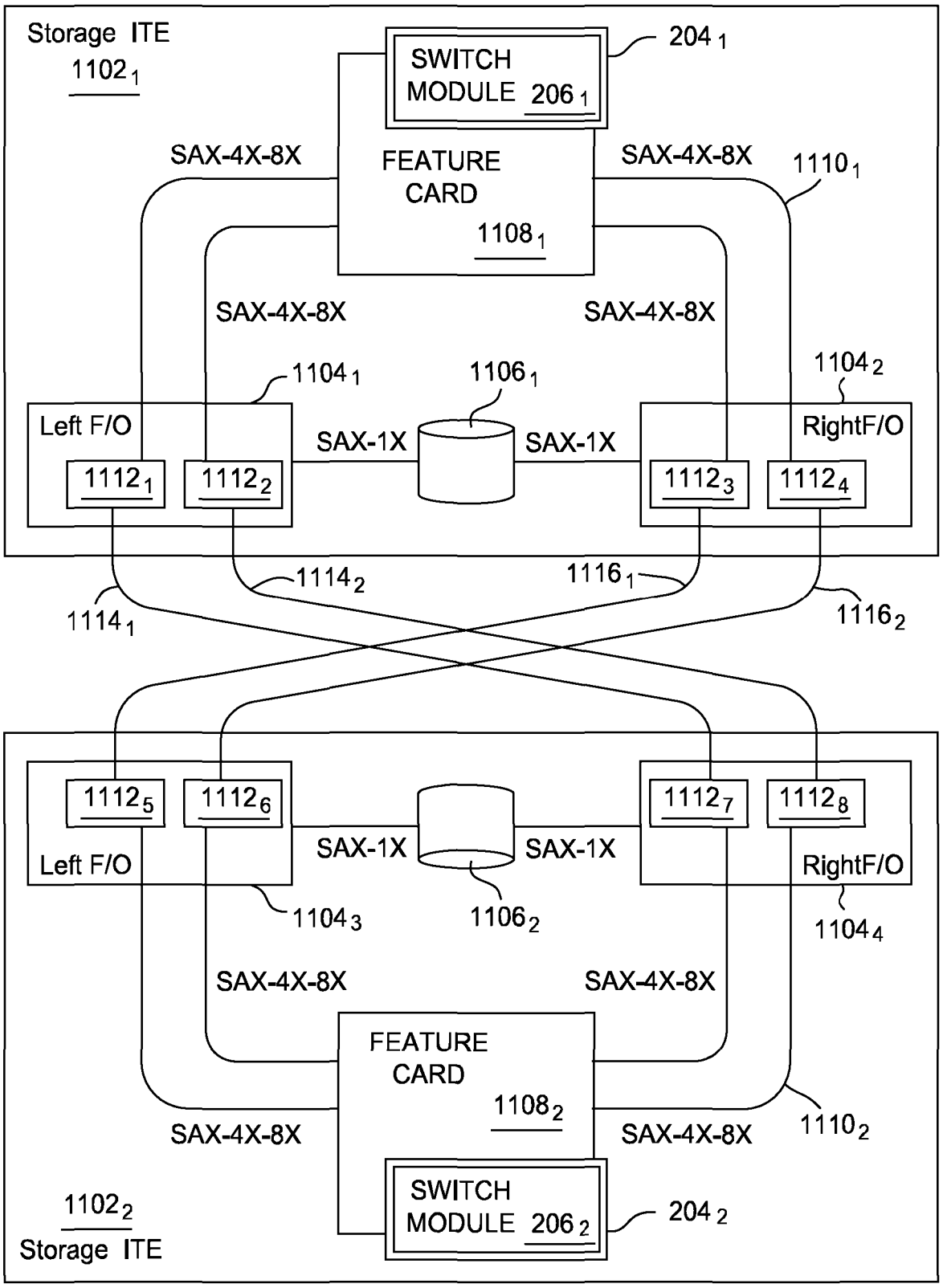
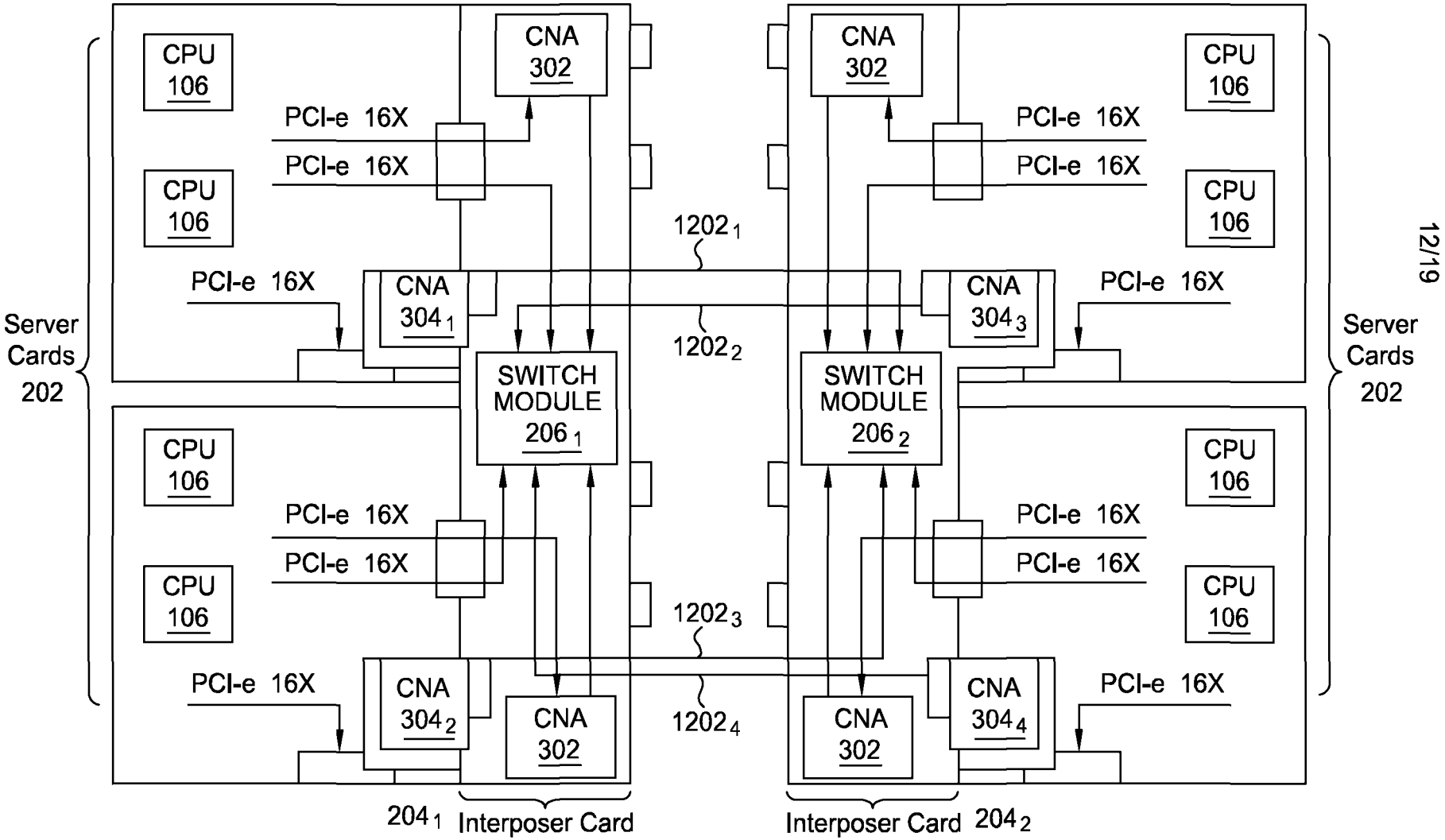

Storage ITE
1102 1
SWITCH MODULE 206 1
204 1
FEATURE CARD
1108 1
SAX-4X-8X
SAX-4X-8X
1110 1
SAX-4X-8X
SAX-4X-8X
1104 1
1106 1
1104 2
Left F/O
SAX-1X
SAX-1X
RightF/O
1112 1
1112 2
1112 3
1112 4
1114 2
1116 1
1114 1
1116 2
1112 5
1112 6
1112 7
1112 8
Left F/O
SAX-1X
SAX-1X
RightF/O
1104 3
1106 2
1104 4
SAX-4X-8X
SAX-4X-8X
FEATURE CARD
1108 2
SAX-4X-8X
SAX-4X-8X
1110 2
SWITCH MODULE 206 2
204 2
1102 2
Storage ITE

FIG. 11

1200

FIG. 12


CPU 106
CNA 302
PCI-e 16X
PCI-e 16X
CPU 106
CPU 106
1202 1
PCI-e 16X
CNA 304 1
CNA 304 3
PCI-e 16X
1202 2
Server Cards 202
SWITCH MODULE 206 1
SWITCH MODULE 206 2
Server Cards 202
CPU 106
CPU 106
PCI-e 16X
PCI-e 16X
PCI-e 16X
PCI-e 16X
CPU 106
CPU 106
1202 3
PCI-e 16X
CNA 304 2
CNA 304 4
PCI-e 16X
CNA 302
1202 4
CNA 302
204 1 Interposer Card
Interposer Card 204 2

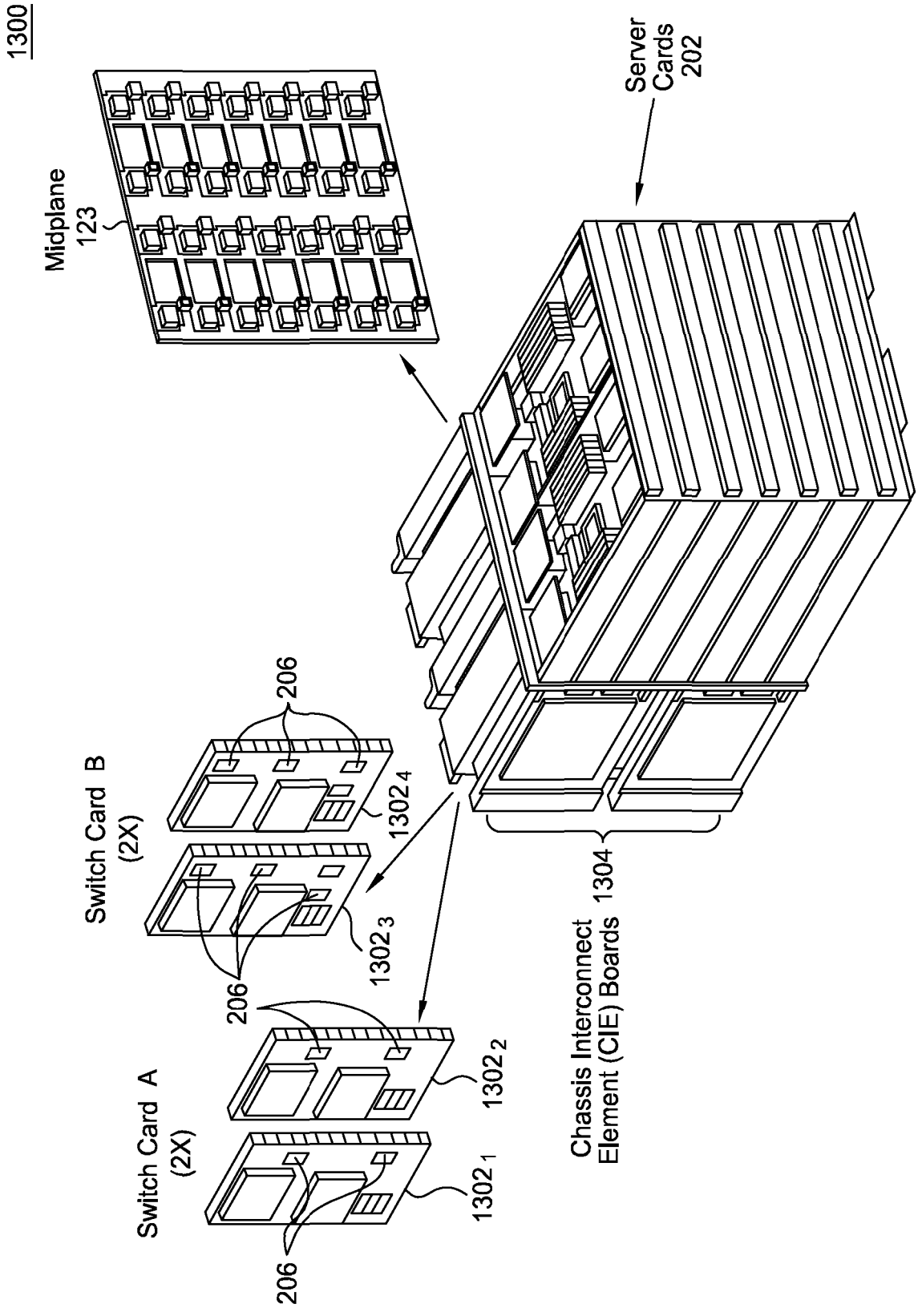

1300
Midplane
123
Server
Cards
202
Switch Card B
(2X)
206
1302 4
1302 3
206
Switch Card A
(2X)
1302 2
1302 1
206
Chassis Interconnect
Element (CIE) Boards 1304

FIG. 13

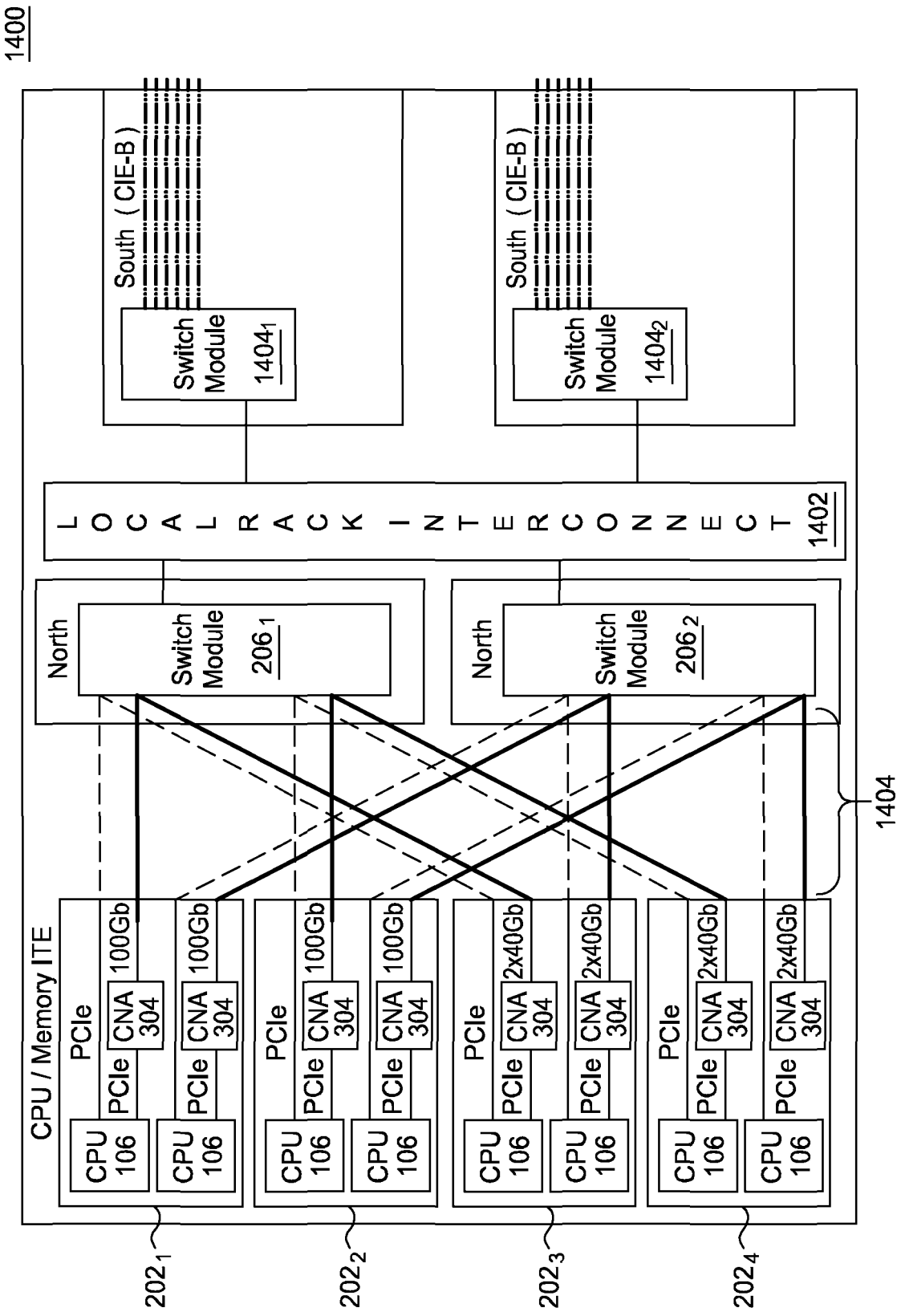

1400
South (CIE-B)
Switch Module 1404 1
South (CIE-B)
Switch Module 1404 2
LOCAL RACK INTERCONNECT 1402
North
Switch Module 206 1
North
Switch Module 206 2
1404
CPU / Memory ITE
PCIe
CNA 304
100Gb
2x40Gb
CPU 106
202 1
202 2
202 3
202 4

FIG. 14

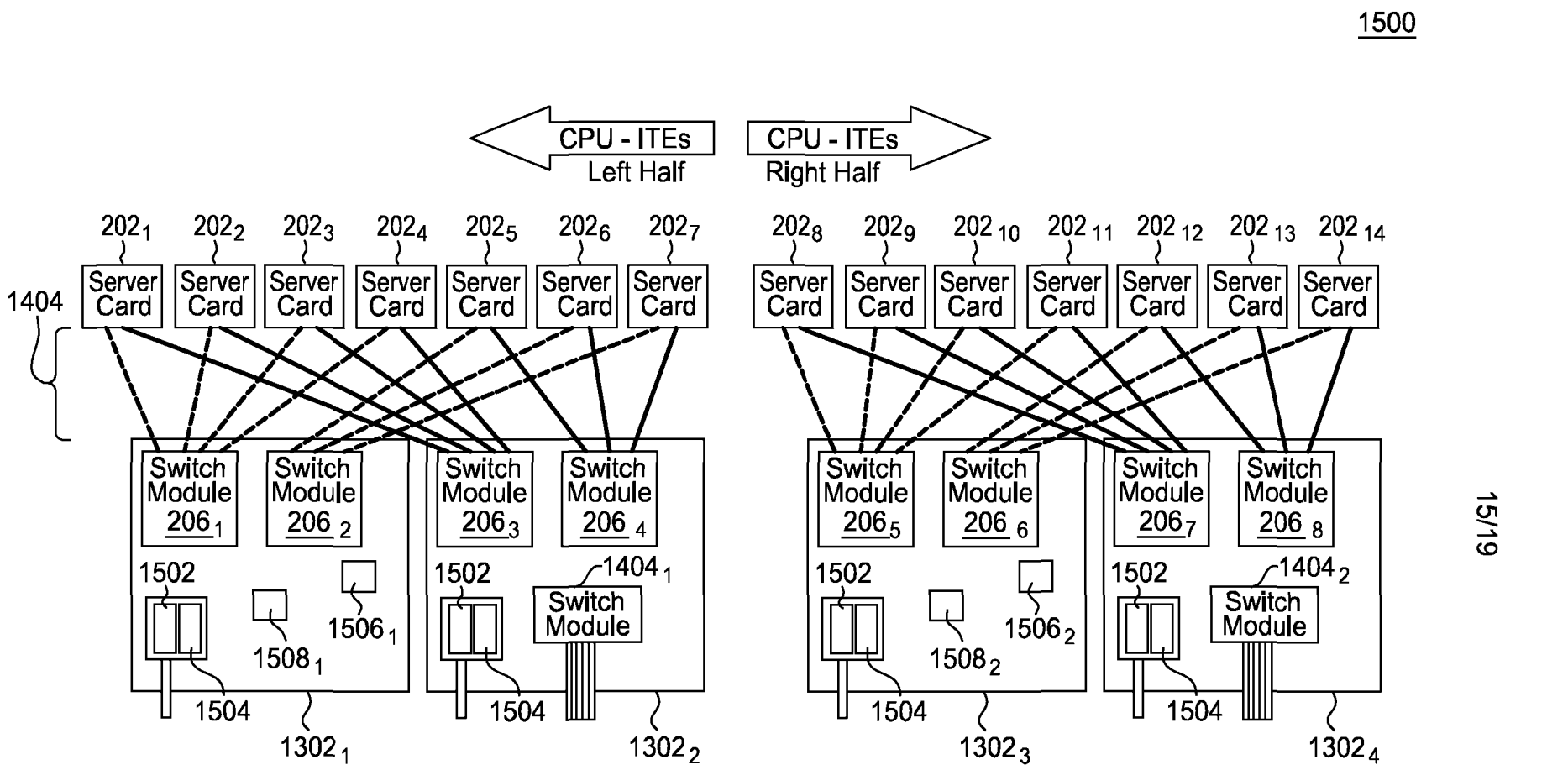



FIG. 15

Key:
Base x8 PCIe Gen3
Optional x8 PCIe Gen3

FIG. 16

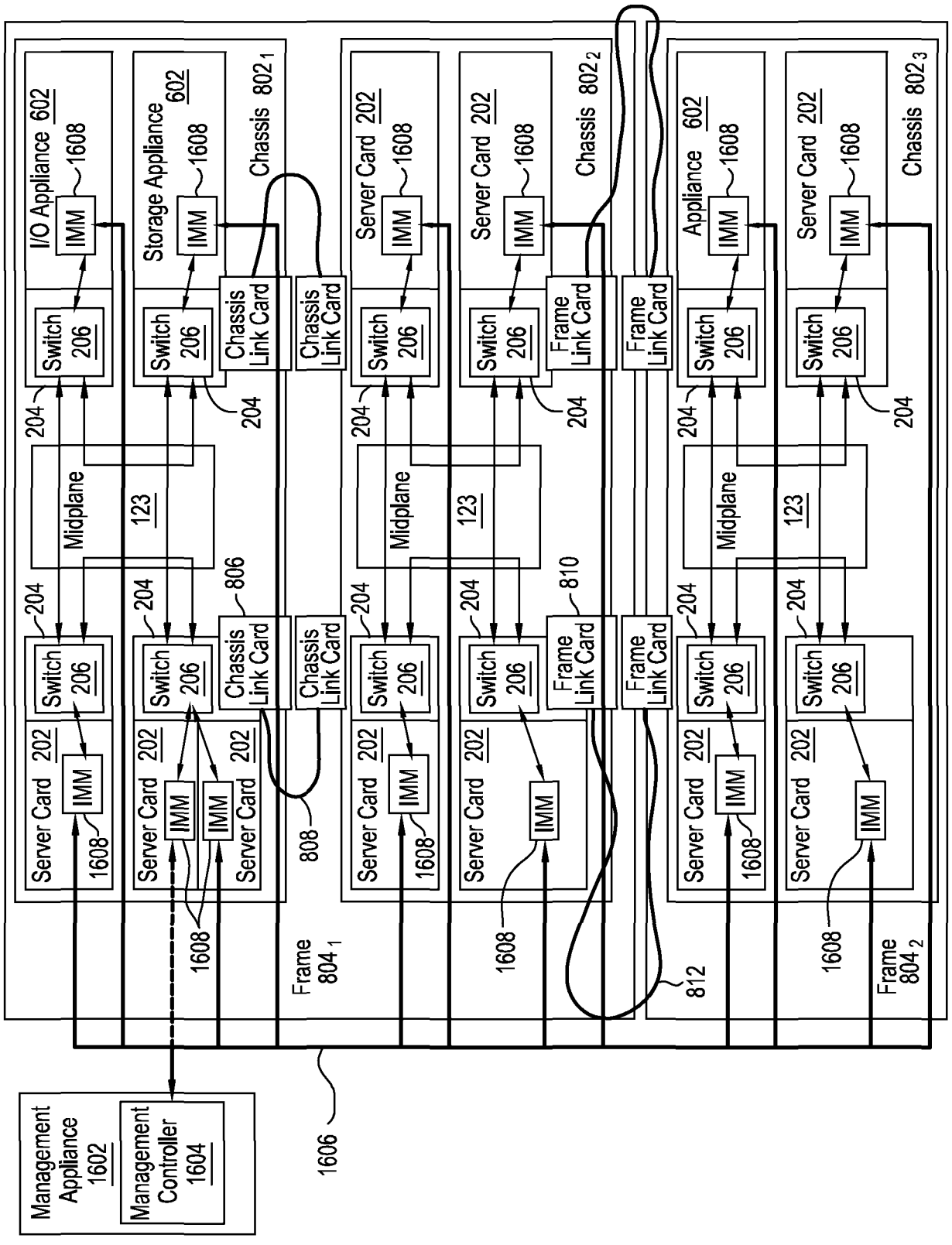
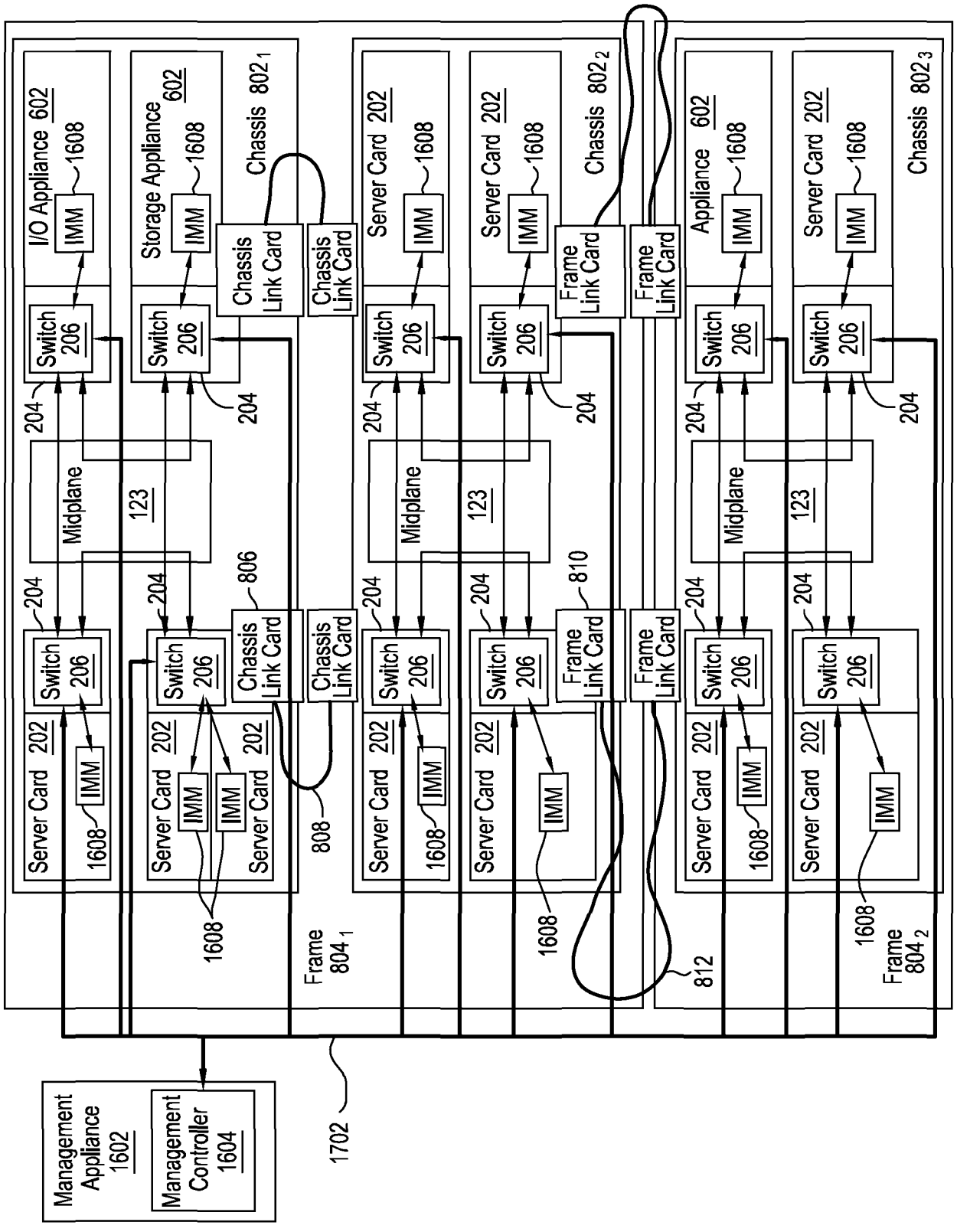

1600
Management Appliance 1602
Management Controller 1604
1606
Frame 804 1
Frame 804 2
Chassis 802 1
Chassis 802 2
Chassis 802 3
I/O Appliance 602
Storage Appliance 602
Appliance 602
Server Card 202
Switch 206
IMM
1608
Midplane
123
204
Chassis Link Card
Frame Link Card
806
808
810
812

FIG. 17


1700
Management Appliance 1602
Management Controller 1604
1702
Frame 804 1
Frame 804 2
Chassis 802 1
Chassis 802 2
Chassis 802 3
I/O Appliance 602
Storage Appliance 602
Appliance 602
Server Card 202
Switch 206
IMM
1608
Midplane
123
204
Chassis Link Card
806
808
Frame Link Card
810
812

1800

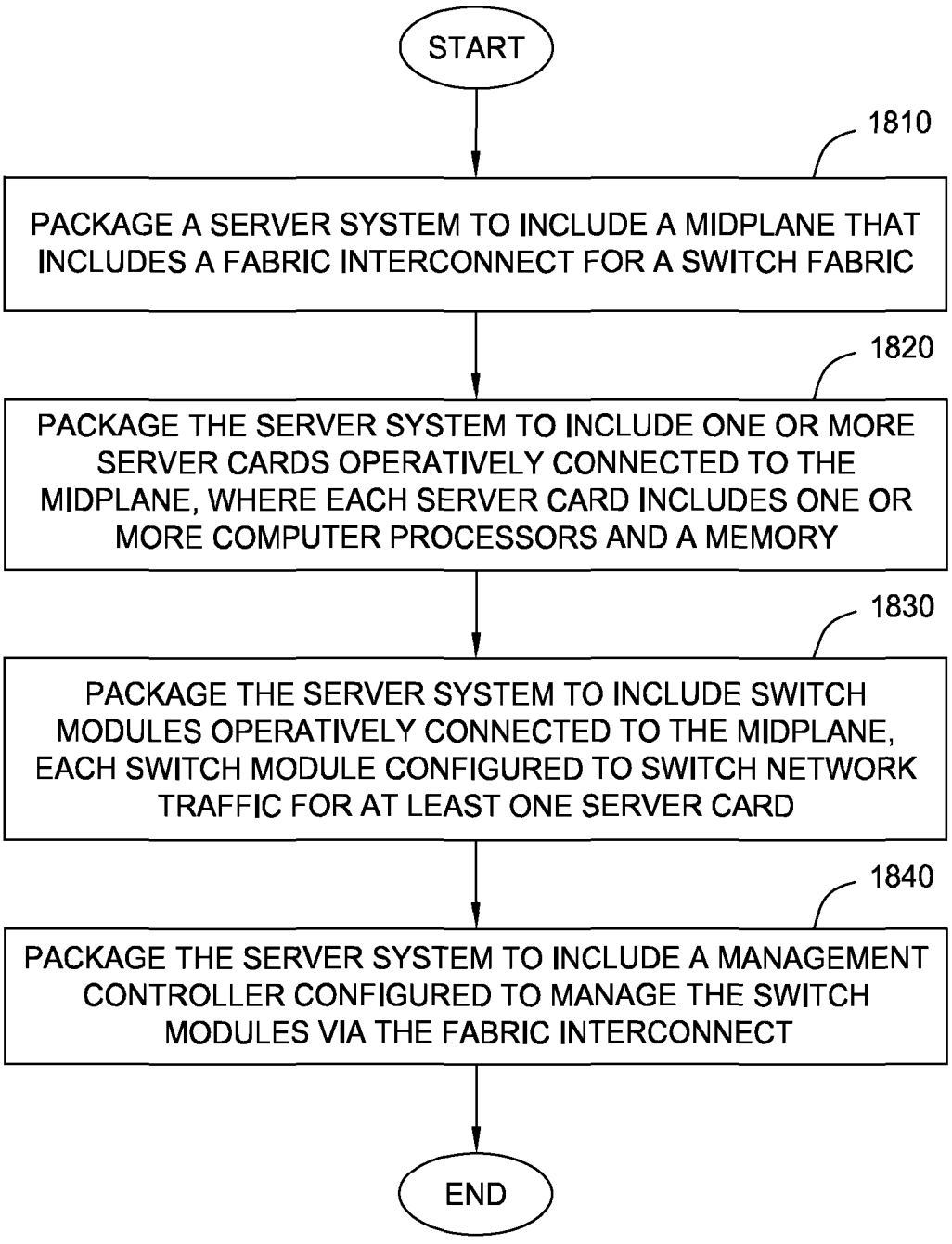

START
1810
PACKAGE A SERVER SYSTEM TO INCLUDE A MIDPLANE THAT INCLUDES A FABRIC INTERCONNECT FOR A SWITCH FABRIC
1820
PACKAGE THE SERVER SYSTEM TO INCLUDE ONE OR MORE SERVER CARDS OPERATIVELY CONNECTED TO THE MIDPLANE, WHERE EACH SERVER CARD INCLUDES ONE OR MORE COMPUTER PROCESSORS AND A MEMORY
1830
PACKAGE THE SERVER SYSTEM TO INCLUDE SWITCH MODULES OPERATIVELY CONNECTED TO THE MIDPLANE, EACH SWITCH MODULE CONFIGURED TO SWITCH NETWORK TRAFFIC FOR AT LEAST ONE SERVER CARD
1840
PACKAGE THE SERVER SYSTEM TO INCLUDE A MANAGEMENT CONTROLLER CONFIGURED TO MANAGE THE SWITCH MODULES VIA THE FABRIC INTERCONNECT
END

FIG. 18

1900

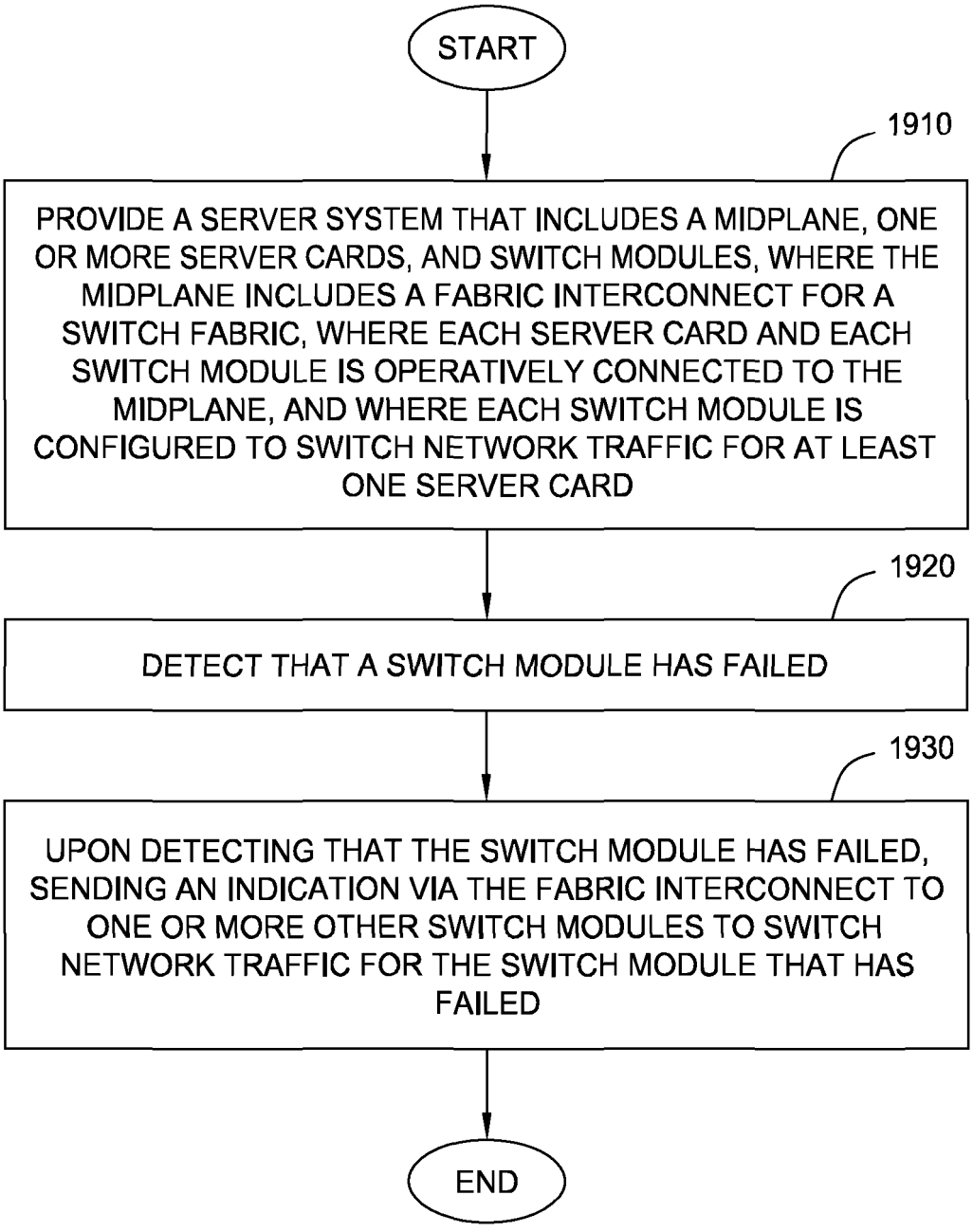

START
1910
PROVIDE A SERVER SYSTEM THAT INCLUDES A MIDPLANE, ONE OR MORE SERVER CARDS, AND SWITCH MODULES, WHERE THE MIDPLANE INCLUDES A FABRIC INTERCONNECT FOR A SWITCH FABRIC, WHERE EACH SERVER CARD AND EACH SWITCH MODULE IS OPERATIVELY CONNECTED TO THE MIDPLANE, AND WHERE EACH SWITCH MODULE IS CONFIGURED TO SWITCH NETWORK TRAFFIC FOR AT LEAST ONE SERVER CARD
1920
DETECT THAT A SWITCH MODULE HAS FAILED
1930
UPON DETECTING THAT THE SWITCH MODULE HAS FAILED, SENDING AN INDICATION VIA THE FABRIC INTERCONNECT TO ONE OR MORE OTHER SWITCH MODULES TO SWITCH NETWORK TRAFFIC FOR THE SWITCH MODULE THAT HAS FAILED
END

FIG. 19

# SWITCH FABRIC MANAGEMENT

## BACKGROUND

Although early computer architectures used standalone, single computers, often referenced as Personal Computers (PCs), more powerful modern computer systems often use multiple computers that are coupled together in a common chassis. An exemplary common chassis is known as a blade chassis, which includes multiple server blades that are coupled by a common backbone within the blade chassis. Each server blade is a pluggable board that includes at least one processor, on-board memory, and an Input/Output (I/O) interface. The multiple server blades are configured to communicate with one another and to share common resources such as storage devices, monitors, input devices, *etc.* Further, one or multiple blade chassis may make up a blade system, which is often dedicated to a single enterprise and/or a particular function, such as processing loans, managing payroll, *etc.*

## SUMMARY

One embodiment of the invention provides a system that includes a midplane, one or more server cards, a plurality of switch modules, and a management controller. The midplane includes a fabric interconnect for a switch fabric. The one or more server cards are operatively connected to the midplane, each server card including one or more computer processors and a memory. The plurality of switch modules are also operatively connected to the midplane, each switch module configured to switch network traffic for at least one of the one or more server cards. The management controller is configured to manage the plurality of switch modules via the fabric interconnect.

Another embodiment of the invention provides a management controller that includes a computer processor and a memory storing management firmware which, when executed on the computer processor, performs an operation. The operation includes managing a plurality of switch modules operatively connected to a midplane of a server system. The midplane comprises a fabric interconnect for a switch fabric. The plurality of switch modules is managed via the fabric interconnect. The midplane is operatively connected to one or more

server cards. Each switch module is configured to switch network traffic for at least one of the server cards.

Yet another embodiment of the invention provides a computer-implemented method that includes providing a server system including a midplane, one or more server cards and a plurality of switch modules. The midplane includes a fabric interconnect for a switch fabric. Each server card and each switch module is operatively connected to the midplane. Each switch module is configured to switch network traffic for at least one of the one or more server cards. The operation also includes managing the plurality of switch modules via the fabric interconnect.

## BRIEF DESCRIPTION OF THE DRAWINGS

So that the manner in which the above recited aspects are attained and can be understood in detail, a more particular description of embodiments of the invention, briefly summarized above, may be had by reference to the appended drawings.

It is to be noted, however, that the appended drawings illustrate only typical embodiments of this invention and are therefore not to be considered limiting of its scope, for the invention may admit to other equally effective embodiments.

Figure 1 is a block diagram of a computing environment having several hosts with access to a server system, according to one embodiment of the invention.

Figure 2 illustrates a configuration in which interposer cards are operatively connected to server cards in a server system, according to one embodiment of the invention.

Figure 3 illustrates a configuration in which an interposer card is operatively connected to two server cards in a server system, according to one embodiment of the invention.

Figure 4 illustrates a server system configured to reduce impact of a repair action on a switch module, according to one embodiment of the invention.

Figure 5 also illustrates a server system configured to reduce impact of a repair action on a switch module, according to one embodiment of the invention.

Figure 6 illustrates a switch fabric for a server system, according to one embodiment of the invention.

Figure 7 illustrates a server system having a midplane that is coupled with multiple interposer cards, according to one embodiment of the invention.

Figure 8 illustrates a server system that includes multiple frames, according to one embodiment of the invention.

Figure 9 illustrates a server system that includes multiple frames, each frame having four chassis, according to one embodiment of the invention.

Figure 10 illustrates a server system that is packaged to include an interposer interconnect, according to one embodiment of the invention.

Figure 11 illustrates a configuration of a server system that eliminates a switch module as an SPOF in a pair of storage ITEs, according to one embodiment of the invention.

Figure 12 illustrates a configuration of a pair of interconnected interposer cards, according to one embodiment of the invention.

Figure 13 illustrates a configuration of a server system that includes multiple switch cards, according to one embodiment of the invention.

Figure 14 illustrates a logical view of a configuration of a server system that includes multiple switch cards, according to one embodiment of the invention.

Figure 15 illustrates a configuration of a server system that includes multiple switch cards, according to one embodiment of the invention.

Figure 16 illustrates a server system that includes a management controller and an independent management interconnect, according to one embodiment of the invention.

Figure 17 illustrates a server system that includes a management controller and that does not include an independent management interconnect, according to one embodiment of the invention.

Figure 18 is a flowchart depicting a method for reducing wiring in a server system that includes a management controller, according to one embodiment of the invention.

Figure 19 is a flowchart depicting a method for managing a switch fabric, according to one embodiment of the invention.

## DETAILED DESCRIPTION

Embodiments of the invention reduce wiring in a server system that includes a management controller configured to manage a switch fabric for the server system. As used herein, a switch fabric refers to a network topology where network nodes connect with each other via one or more network switches. In one embodiment, the server system includes a midplane that in turn includes a fabric interconnect for the switch fabric. The server system may also include one or more server cards operatively connected to the midplane. Each server card includes one or more computer processors and a memory. In one embodiment, each server card may be a server blade, also referred to as a blade server or blade. The server system may also include switch modules operatively connected to the midplane, where each switch module switches network traffic for at least one server card. The management controller may be configured to manage the switch modules via the fabric interconnect and not via any other interconnect between the switch modules. Wiring in the server system is reduced because no other interconnect is required to manage the switch modules. Thus, costs and/or complexity of packaging the server system may be reduced.

Further, embodiments of the invention may reduce the impact of a switch failure in a switch fabric. For example, a server system may include a first interposer card and switch module disposed between the midplane and the one or more server cards, operatively connecting the midplane to one or more server cards. Although described with reference to the first interposer card, the midplane may be configured to couple with a plurality of interposer cards. The first interposer card may be hot-swappable from the midplane, and the one or more server cards may be hot-swappable from the first interposer card. By packaging the server system using the techniques disclosed herein, single points of failure (SPOFs) and/or single points of repair (SPORs) may be reduced or minimized. SPOFs are said to be eliminated when the server system can continue to operate in the presence of any component failure. SPORs are said to be eliminated when the server system can continue to operate while any (failed) component is being repaired or replaced.

In the following, reference is made to embodiments of the invention. However, it should be understood that the invention is not limited to specific described embodiments. Instead, any combination of the following features and elements, whether related to different embodiments or not, is contemplated to implement and practice the invention. Furthermore, although embodiments of the invention may achieve advantages over other possible solutions and/or over the prior art, whether or not a particular advantage is achieved by a given embodiment is not limiting of the invention. Thus, the following aspects, features, embodiments and advantages are merely illustrative and are not considered elements or limitations of the appended claims except where explicitly recited in a claim(s). Likewise, reference to "the invention" shall not be construed as a generalization of any inventive subject matter disclosed herein and shall not be considered to be an element or limitation of the appended claims except where explicitly recited in a claim(s).

As will be appreciated by one skilled in the art, aspects of the present invention may be embodied as a system, method or computer program product. Accordingly, aspects of the present invention may take the form of an entirely hardware embodiment, an entirely software embodiment (including firmware, resident software, micro-code, *etc.*) or an embodiment combining software and hardware aspects that may all generally be referred to herein as a "circuit," "module" or "system." Furthermore, aspects of the present invention may take the form of a computer program product embodied in one or more computer readable medium(s) having computer readable program code embodied thereon.

Any combination of one or more computer readable medium(s) may be utilized. The computer readable medium may be a computer readable signal medium or a computer readable storage medium. A computer readable storage medium may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing. In the context of this document, a computer readable storage medium may be any tangible medium that can contain, or store a

program for use by or in connection with an instruction execution system, apparatus, or device.

A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of a carrier wave. Such a propagated signal may take any of a variety of forms, including, but not limited to, electro-magnetic, optical, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer readable storage medium and that can communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device.

Program code embodied on a computer readable medium may be transmitted using any appropriate medium, including but not limited to wireless, wireline, optical fiber cable, RF, *etc.*, or any suitable combination of the foregoing.

Computer program code for carrying out operations for aspects of the present invention may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Smalltalk, C++ or the like and conventional procedural programming languages, such as the "C" programming language or similar programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider).

Aspects of the present invention are described below with reference to flowchart illustrations and/or block diagrams of methods, apparatus (systems) and computer program products according to embodiments of the invention. It will be understood that each block of the flowchart illustrations and/or block diagrams, and combinations of blocks in the flowchart illustrations and/or block diagrams, can be implemented by computer program instructions. These computer program instructions may be provided to a processor of a general purpose computer, special purpose computer, or other programmable data processing apparatus to produce a machine, such that the instructions, which execute via the processor of the

computer or other programmable data processing apparatus, create means for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

These computer program instructions may also be stored in a computer readable medium that can direct a computer, other programmable data processing apparatus, or other devices to function in a particular manner, such that the instructions stored in the computer readable medium produce an article of manufacture including instructions which implement the function/act specified in the flowchart and/or block diagram block or blocks.

The computer program instructions may also be loaded onto a computer, other programmable data processing apparatus, or other devices to cause a series of operational steps to be performed on the computer, other programmable apparatus or other devices to produce a computer implemented process such that the instructions which execute on the computer or other programmable apparatus provide processes for implementing the functions/acts specified in the flowchart and/or block diagram block or blocks.

The flowchart and block diagrams in the Figures illustrate the architecture, functionality and operation of possible implementations of systems, methods and computer program products according to various embodiments of the present invention. In this regard, each block in the flowchart or block diagrams may represent a module, segment or portion of code, which comprises one or more executable instructions for implementing the specified logical function(s). It should also be noted that, in some alternative implementations, the functions noted in the block may occur out of the order noted in the figures. For example, two blocks shown in succession may, in fact, be executed substantially concurrently, or the blocks may sometimes be executed in the reverse order, depending upon the functionality involved. It will also be noted that each block of the block diagrams and/or flowchart illustration, and combinations of blocks in the block diagrams and/or flowchart illustration, can be implemented by special purpose hardware-based systems that perform the specified functions or acts, or combinations of special purpose hardware and computer instructions.

Figure 1 is a block diagram of a computing environment 100 having several hosts with access to a server system 102, according to one embodiment of the invention. Although only three hosts 134a,b,n are depicted for clarity, those skilled in the art will appreciate that additional hosts may have access to the server system 102. The hosts 134a,b,n are connected through a network fabric 132 to the server system 102. Depending on the embodiment, each host

134a,b,n may act as a client that accesses functionality provided by the server system 102 and/or may provide respective server functionality external to the server system 102. The network fabric 132 may be a telecommunications network and/or a wide area network (WAN). In a particular embodiment, the network fabric 132 is the Internet. The server system 102 includes a chassis that houses server blades 104a,b,n. The server blades 104a,b,n are coupled to a midplane 123, which provides mechanical and logical connections (*e.g.*, data and control signal interchange) among the server blades 104a,b,n. Although three server blades 104a,b,n are depicted, those skilled in the art will appreciate that additional server blades may be connected to the midplane 123. Further, although embodiments are described herein with reference to blade systems, other form factors or physical configurations (*e.g.*, rack systems) are broadly contemplated.

Further, although embodiments are described herein with reference to the server blades 104a,b,n being coupled to the midplane 123, those skilled in the art will recognize that more generally, the server blades may be coupled to any printed circuit board (PCB) that serves as a backbone for the chassis, such as a backplane, motherboard, *etc.* Further still, although embodiments are described herein with reference to the server system 102 having a single chassis, those skilled in the art will recognize that in other embodiments, the server system 102 may include multiple chassis. For example, in an alternative embodiment, the server system 102 may be a blade system that includes at least two blade chassis, each having a plurality of blades.

In one embodiment, the server system 102 further includes one or more management modules 124. In the depicted embodiment, the server system 102 includes a primary management module 124a and a backup management module 124b. Each management module 124 is capable of managing multiple server blades 104. During normal operation, one of the management modules 124 is operatively connected to the server blades 104 via a local area network (LAN) 122, the midplane 123 and a Baseboard Management Controllers (BMCs) 110 of each server blade 104 to form an in-band management pathway. In one embodiment, the network fabric 132 serves as an extension to the LAN 122. The LAN 122 and BMC 110 are further discussed below.

In one embodiment, the midplane 123 is mounted in the middle of the chassis of the server system 102 and contains circuitry and sockets 112 into which additional electronic devices or cards, including server blades 104, may be inserted. The midplane 123 includes at least one

bus for secure in-band internal communication via the BMCs 110 and between the management modules 124 and the server blades 104 and/or amongst the server blades 104 themselves.

In one embodiment, when a server blade 104 is inserted into a specific socket 112, a physical address is established for the server blade 104. For example, assume that server blade 104a is inserted into the socket 112a. In one embodiment, control logic 116a detects presence of the server blade 104a in the socket 112a. The control logic 116a may comport with the Electronics Industry Association (EIA) RS485 Standard for data communication. In other embodiments, the control logic 116a may be compliant with the Phillips' Inter-IC (Inter-Integrated Circuit, or $I^2C$) standard or with an Ethernet network standard. The control logic 116a, operating in conjunction with the management module 124a, assigns a physical address on a bus in the midplane 123 to the server blade 104a responsive to insertion of the server blade 104a into the socket 112a. As shown, each server blade 104 is associated with a respective control logic 116 that is operatively connected to the midplane 123. In an alternative embodiment, multiple server blades 104 may share a single control logic 116.

In one embodiment, each server blade 104 is assigned a unique Internet Protocol (IP) address on the midplane 123. That is, the midplane 123 may support intercommunication using IP addressing protocol, in which each device that is operatively connected to the midplane 123 has an IP address assigned by logic (not shown) that is either within or outside the chassis of the server system 102. For example, a Dynamic Host Configuration Protocol (DHCP) server may be used to assign an IP address to the server blade 104a. Communication with the server blade 104a thereafter occurs via a network interface controller (NIC) 114a associated with the server blade 104a. The NIC 114a may be any type of network communications device allowing the server blade 104a to communicate with other server blades 104b,n and/or computers via the LAN 122 and/or the network fabric 132.

In one embodiment, an integrated module 126a is operatively connected to the NIC 114a. The integrated module 126a may be used in pairs (*e.g.*, with integrated module 126b) to provide redundancy. As is known, Small Computer System Interface (SCSI) refers to a set of standards for physically connecting and transferring data between computers and peripheral devices. In one embodiment, the integrated modules 126 include switch modules 128, such as a Serial Attached SCSI (SAS) switch module. The switch modules 128 provide, for the server blades 104, connectivity to Ethernet, Fibre Channel over Ethernet (FCoE), SAS, *etc.*

In one embodiment, each switch module 128 is a switch chip. Depending on the embodiment, the integrated modules 126 may further include redundant array of independent disks (RAID) controllers 130. Each RAID controller 130 is interconnected to RAID devices, such as storage devices in a RAID configuration. The RAID devices may be located within one or more of the server blades 104. The RAID controllers 130 and the RAID devices may collectively be viewed as a RAID subsystem of the server system 102.

In one embodiment, each storage device may be a persistent storage device. Further, each storage device may be a combination of fixed and/or removable storage devices, such as fixed disc drives, floppy disc drives, tape drives, removable memory cards, solid-state drives or optical storage. The memory 108 and the storage device may be part of one virtual address space spanning multiple primary and secondary storage devices.

In one embodiment, each server blade 104 may have at least one central processing unit (CPU) 106 and a memory 108. The CPU 106 is included to be representative of a single CPU, multiple CPUs, a single CPU having multiple processing cores, and the like. Similarly, the memory 108 may be a random access memory. While the memory 108 is shown as a single identity, it should be understood that the memory 108 may comprise a plurality of modules, and that the memory 108 may exist at multiple levels, from high speed registers and caches to lower speed but larger DRAM chips. The memory 108 may be a flash read-only memory ("flash ROM" or "flash memory") that can be erased and reprogrammed in units of memory referred to as "blocks." The memory 108 may also include non-volatile Electrically Erasable Programmable Read Only Memory (EEPROM) that is similar to flash memory, except that EEPROM is erased and rewritten at a byte level and is usually smaller in capacity. Each server blade 104 may be oriented as a processor blade or a storage blade. A processor blade includes one or more processing devices, while a storage blade includes a number of integrated storage devices such as disk drives.

In one embodiment, when the server blade 104 is shipped from a manufacturer, the memory 108 may be pre-burned with firmware, including a basic input/output system (BIOS) and software for monitoring the server blade 104. The monitoring may include controlling storage devices, monitoring and controlling voltages throughout the system, determining the power-on status of the server blade 104, requesting access to a shared keyboard, video, mouse, compact disc read-only memory (CD-ROM) and/or floppy disk drives, monitoring the operating system (OS) running on the server blade 104, *etc.* Examples of operating systems

include UNIX, versions of the Microsoft Windows® operating system, and distributions of the Linux® operating system. More generally, any operating system supporting the functions disclosed herein may be used.

In one embodiment, the management modules 124 are capable of detecting the presence, quantity, type and revision level of each server blade 104, power module 118, and midplane 123 in the system. The management modules 124 may also directly control the operation of each server blade 104 and the power module 118. The management modules 124 may also directly control the operation of cooling fans 120 and other components in the chassis of the server system 102. Directly controlling the operation entails controlling the operation without using the BIOS in the server blades 104. In an alternative embodiment, the management modules 124 may use the BIOS to indirectly control the operation of the cooling fans 120 and the other components in the chassis of the server system 102.

In one embodiment, each server blade 104 includes a baseboard management controller (BMC) 110 that provides local supervisory control of the server blade 104 to which the BMC 110 is associated. Each BMC 110 is configured to communicate with a management module 124 by either using communication path of the LAN 122 (*i.e.*, via an in-band network) or alternatively by using switch modules 128 and NICs 114 (*i.e.*, via an out-of-band network). The management modules 124 may utilize a variety of communications paths in the LAN 122, such as RS485 path, a LAN path, and an $I^2C$ path, to communicate with each server blade 104.

In one embodiment, the LAN 240 is an in-band network also comporting with the Electronics Industry Association (EIA) RS485 Standard for data communication. The management modules 124 – *e.g.*, either the primary management module 124a or the backup management module 124b if the primary management module 124a is down – communicate via the LAN 122 with the BMC 110, which includes logic for coordinating communication with the server blades 104 via the sockets 112.

In one embodiment, the LAN 122 may be configured to allow communications between the server blades 104 and the management modules 124 relating to the remote BIOS settings and BIOS management. The server blades 104 may use BMCs 110 as proxies to communicate with the management modules 124 through the RS485 protocol. Similarly, the management modules may use BMCs 110 as proxies to communicate with the server blades 104 through

the RS485 protocol. In an alternative embodiment, an RS485 connection may be separately made between each server blade 104 and the management modules 124. Additionally, other communications protocols and paths may be utilized over the switch modules 128, such as $I^2C$, TCP/IP, Ethernet, FCoE, *etc.*

Depending on the embodiment, the server system 102 may also be operatively connected to an input device and/or an output device. The input device may be any device for providing input to the server system 102. For example, a keyboard, keypad, light pen, touch-screen, track-ball, or speech recognition unit, audio/video player, and the like may be used. The output device may be any device for providing output to a user of the server system 102. For example, the output device may be any conventional display screen or set of speakers, along with their respective interface cards, *i.e.*, video cards and sound cards. Further, the input device and output device may be combined. For example, a display screen with an integrated touch-screen, a display with an integrated keyboard, or a speech recognition unit combined with a text speech converter may be used.

Figure 2 illustrates a configuration 200 in which interposer cards are operatively connected to server cards in a server system, according to one embodiment of the invention. Depending on the embodiment, the interposer card may also be coupled to the server cards and/or the midplane. In one embodiment, to reduce the impact of a repair action on a failed switch module, the server system is configured to include a midplane 123 and server cards 202, where the midplane and server cards 202 are operatively connected via one or more interposer cards 204. Each server card 202 may correspond to a server blade 104. Each interposer card 204 is hot-swappable and includes one or more switch modules 206. The switch modules 206 switch network traffic for one or more server cards 202 operatively connected with the respective interposer card.

In one embodiment, the server system is configured to detect failure of the switch module 206. Upon detecting failure of the switch module 206, the server system may output for display an indication to perform a repair action on the switch module 206. For example, the indication may be output for display in a graphical user interface (GUI) window or as an alert to be emailed to a user. Depending on the embodiment, the detecting and/or the outputting may be performed by any component of the server system, such as the server cards 202, the switch module 206 and/or firmware contained in the server system. For example, in one

embodiment, the server system includes management firmware that monitors health of the server system and detects failure of the switch module 206.

Accordingly, when a switch module 206 fails, the interposer card 204 that includes the switch module 206 may be replaced with an interposer card having a functional switch module. Further, the interposer card 204 may be replaced without requiring the server system and/or switch fabric to be powered off or rebooted, because of the hot-swappable properties of the interposer cards, server cards and/or midplane. The interposer card having a functional switch module may then be reintegrated into the switch fabric via the configuration tool. Depending on the embodiment, the configuration tool may execute on the server system or on another computer connected to the server system via the network fabric 132.

Accordingly, during the duration of replacing the interposer card with the failed switch module, only the failed switch module and associated server cards are unreachable from the network fabric. During the duration of replacing the interposer card with the failed switch module, other switch modules and/or server cards operatively connected to the midplane remain reachable. Accordingly, the impact of the repair action on the failed switch module is localized to the server cards associated with the failed switch module. In other words, the only network nodes that are unreachable from the network fabric during the repair action are the network nodes associated with the server cards operatively connected to the failed switch module.

Figure 3 illustrates a configuration 300 in which an interposer card is operatively connected to two server cards in a server system, according to one embodiment of the invention. As shown, the interposer card 204 includes the switch module 206 and two Converged Network Adapters (CNAs) 302. The two server cards 202 may each also include two CPUs 106 and a CNA 304. In one embodiment, the CPUs 106 connect to FCoE with CNAs, which contain both Fibre Channel Host Bus Adapter (HBA) and Ethernet NIC functionality. The CNAs may include one or more physical Ethernet ports and may be configured to offload – from the CPUs 106 – low level frame processing and/or SCSI protocol functionality traditionally performed by Fibre Channel host bus adapters. As described above, the switch module 206 switches network traffic for the server cards 202. If the switch module 206 fails, the configuration 300 allows the switch module 206 to be replaced without requiring reboot of server cards of other interposer cards coupled with the midplane.

Figure 4 illustrates a server system 400 configured to reduce impact of a repair action on a switch module, according to one embodiment of the invention. As shown, the server system 400 includes a logical server 402 configured across a processor information technology element (ITE) 404 and an I/O ITE 406. As used herein, an ITE generally refers to any appliance configured to operatively connect to the midplane 123. In an alternative embodiment, the logical server 402 may also be configured across a storage ITE 408. The I/O ITE 406 and storage ITE 408 are configured to provide additional I/O capacity and storage capacity, respectively, to one or more processer ITEs. Depending on the embodiment, each ITE 404, 406, 408 may be integrated as part of one or more server blades or may be coupled with the midplane 123 as a standalone card. The processor ITE 404 includes one or more virtual machines 410, a hypervisor 412, memory 414, processors 416 and hard disk drives 418. The I/O ITE 406 includes a shared I/O ITE component 422 and I/O adapters 424, while the storage IT 408 includes a shared storage ITE component 426 and solid state drives 428.

In one embodiment, the server blade 404 and the ITEs 406, 408 each further include a switch module 206. Each switch module 206 may be a switch chip and may be included in an interposer card (not shown) that is disposed between the midplane 123 and the processor ITE 404 and/or ITE 406, 408. Collectively, the switch modules 206 provide a switch fabric 432. A failing of the switch module $206_1$ of the processor ITE 404 – denoted by an X symbol 430 – impacts only the processor ITE 404 and not other ITEs operatively connected to the midplane 123. Accordingly, other logical servers configured across the I/O ITE 406 and/or storage ITE 408 remain operational, and connectivity in the switch fabric 432 remains largely operational – *i.e.*, except for connectivity to the processor ITE 404. Depending on the embodiment, the connectivity in the switch fabric 432 that remains operational may also include redundant connectivity in the switch fabric 432. Further, the switch module $206_1$ may be replaced without impacting the other ITEs, logical servers, and/or the switch fabric 432. Accordingly, availability of the switch fabric 432 may be improved.

Figure 5 also illustrates a server system 500 configured to reduce impact of a repair action on a switch module, according to one embodiment of the invention. As shown, the server system 500 includes a first logical server 402 configured across a first processor ITE 404 and the I/O ITE 406. The server system 500 also includes a second logical server 502 configured across a second processor ITE 504 and the I/O ITE 406. In an alternative embodiment, the first logical server 402 and/or the second logical server 502 may also be configured across the storage ITE 408. The ITEs 404, 406, 408, 504 each includes a switch module 206. A failing of the switch

module $206_1$ of the processor ITE 404 – denoted by an X symbol 506, impacts only the first processor ITE 404 and not the second processor ITE 504. Accordingly, the second processor ITE 504, the I/O ITE 406 and the storage ITE 408 remain operational and retain connectivity to the switch fabric during failure and/or replacement of the switch module $206_1$.

Figure 6 illustrates a switch fabric 432 for a server system, according to one embodiment of the invention. As shown, the switch fabric 432 includes a plurality of switch modules 206, each included in a respective interposer card 204. Each interposer card 204 operatively connects two server cards 202 to the switch fabric 432. Depending on the embodiment, the switch modules may be connected to one another in switch fabric via wiring housed in a midplane, cabling external to the midplane, or a combination thereof. Further, one or more of the switch modules 206 may be coupled with appliances 602 other than the server cards 202. Examples of appliances 602 include network appliances, storage appliances, and I/O appliances. Accordingly, failure and/or replacement of a switch module 206 impacts only the server cards 202 coupled with the switch module 206 and does not otherwise impact the rest of the switch fabric 432 and/or the other server cards.

Figure 7 illustrates a server system 700 having a midplane 123 that is coupled with a plurality of interposer cards 204, according to one embodiment of the invention. Each interposer card 204 includes a switch module 206 and operatively connects one or two server cards 202 to the midplane 123. The midplane includes fabric wiring that connects the switch modules 206 to form a switch fabric. The server system 700 is configured such that the server cards 202 are hot-swappable from the interposer cards 204. The server system 700 is further configured such that the interposer cards 204 are hot-swappable from the midplane 123. Accordingly, the packaging and hot-swappable properties of the server system 700 allow a faulty switch module 206 to be replaced while minimizing or reducing impact to the server system 700 and/or the switch fabric.

Figure 8 illustrates a server system 800 that includes multiple frames 804, according to one embodiment of the invention. Each frame includes one or more chassis 802 that are operatively connected via chassis link cards 806 and associated cabling 808. Further, the chassis 802 may be operatively connected across two frames via frame link cards 810 and associated cabling 812. Each chassis 802 houses a midplane 123 that is coupled with one or two server cards 202 via an interposer card 204 according to the techniques disclosed herein. Each interposer card 204 includes a switch module 206 for switching traffic for the server

cards 202. The midplanes 123 include fabric wiring for interconnecting the switch modules 206. Together, the switch modules 206, the fabric wiring, the chassis link cards 806 and associated cabling 808, and the frame link cards 810 and associated cabling 812 form a switch fabric for the server system 800. In other words, the switch fabric for the server system 800 includes cross-chassis and cross-frame interconnects. Accordingly, failure and/or replacement of a switch module 206 – as denoted by an X symbol 814 – merely removes an associated server card 202 from the switch fabric. Operation of the other server cards and/or connectivity of the switch fabric are thereby maintained during the failure and/or replacement of the switch module 206.

Figure 9 illustrates a server system 900 that includes multiple frames, each frame having four chassis 802, according to one embodiment of the invention. The chassis within each frame may be operatively connected via inter-chassis cabling 904. Chassis from different frames may be operatively connected via inter-frame cabling 906. Each chassis houses a midplane having fabric wiring, at least one interposer card having a switch module, and at least one server card according to the techniques disclosed herein. Together, the switch modules, the fabric wiring, the inter-chassis cabling 904, and the inter-frame cabling 906, and any associated link cards form a switch fabric for the server system 900. The server system 900 is thereby configured to increase availability of the switch fabric and/or of the server system 900 during failure and/or replacement of a switch module.

Figure 10 illustrates a server system 1000 packaged to include an interposer interconnect, according to one embodiment of the invention. The packaging of the server system 1000 may include an interposer interconnect 1006 between at least a first interposer card and a second interposer card. And each interposer card includes a switch module 206 coupled with two server cards 202. The interposer interconnect 1006 may include cabling between a network adapter of the first interposer card and a network adapter of the second interposer card. Such cabling may itself be external to the midplane. Each network adapter may be a CNA 302 of the respective interposer card or a CNA 304 of the respective server card. The server system 1000 may also include switch modules 1002 that provide connectivity to external server systems and/or storage controllers. The switch modules 1002 may be operatively connected to the switch modules 206 via one or more CNAs 1004.

Accordingly, when a switch module $206_1$ of the first interposer card fails, a switch module $206_2$ of the second interposer card may switch network traffic for the server cards $202_1$, $202_2$

of the first interposer card – in addition to switching network traffic for the server cards $202_3$, $202_4$ of the second interposer card. Thus, packaging the server system to include the interposer interconnect 1006 eliminates the switch module $206_1$ of the first interposer card as an SPOF. In other words, the server cards $202_1$, $202_2$ of the first interposer card retain connectivity to the switch fabric and/or maintain redundancy thereof, even upon failure of the switch module $206_1$ of the first interposer card.

In one embodiment, the interposer interconnect 1006 may further include cabling between CNAs $302_3$, $302_4$ of the second interposer card to the switch module $206_1$ of the first interposer card. Doing so eliminates the switch module $206_2$ of the second interposer card as an SPOF – in addition to eliminating the switch module $206_1$ as an SPOF. Accordingly, both the switch module $206_1$ of the first interposer card and the switch module $206_2$ of the second interposer card are eliminated as SPOFs.

In one embodiment, the CNA $302_4$ is connected to the switch module $206_1$ via cabling and provides redundancy for the second interposer card. Accordingly, when the switch module $206_2$ of the second interposer card fails, the switch module $206_1$ of the first interposer card may switch network traffic for the server cards $202_3$, $202_4$ of the second interposer card – in addition to switching network traffic for the server cards $202_1$, $202_2$ of the first interposer card.

More generally, packaging the server system 1000 to include the interposer interconnect 1006 between pairs of interposer cards eliminates the switch modules of the each interposer card as an SPOF. Each pair of interposer cards may include two interposer cards that are adjacent to one another according to a predefined axis within a chassis of the server system 1000. In an alternative embodiment, each pair of interposer cards resides in a single rack in the server system 1000. The predefined axis may include an x-axis, a y-axis, a z-axis, or any axis suited for describing relative positions of the interposer cards within the chassis of the server system 1000.

Further, although embodiments are herein described with reference to pairs of interposer cards being interconnected together, other embodiments are broadly contemplated. For example, in an alternative embodiment, three or more interposer cards may be interconnected in a daisy-chained manner. In this example, the interposer interconnect includes: (i) cabling

between a CNA of the first interposer card and a switch module of the second interposer card, (ii) cabling between a CNA of the second interposer card and a switch module of the third interposer card, and (iii) cabling between a CNA of the third interposer card and a switch module of the first interposer card. In an alternative embodiment, the interposer interconnect includes cabling between the CNA of the third interposer card and the switch module of the second interposer card (rather than the first interposer card). Further, to provide additional redundancy, each interposer card may be interconnected with multiple other interposer cards. For example, the interposer interconnect may include: (i) cabling between a first CNA of the third interposer card and a switch module of the first interposer card and (ii) cabling between a second CNA of the third interposer card and a switch module of the second interposer card. Accordingly, if switch module failures occur in the second and third interposer cards, respectively, then the first interposer card is configured to switch network traffic for server cards coupled with the third interposer card. Those skilled in the art will recognize that any predefined number of interposer cards may be interconnected using the techniques disclosed herein.

Figure 11 illustrates a configuration 1100 of a server system that eliminates a switch module 206 as an SPOF in a pair of storage ITEs, according to one embodiment of the invention. The configuration 1100 includes a first storage ITE $1102_1$ and a second storage ITE $1102_2$. Each storage ITE $1102_1$, $1102_2$ may be connected to the midplane via an interposer card 204 that includes a switch module 206. In an alternative embodiment, each storage ITE is coupled with the midplane and includes the switch module 206. As shown, each storage ITE includes a plurality of components, including a feature card 1108, two fanout cards 1104 and a storage device 1106. The feature card 1108 of each storage ITE 1102 may be configured to customize functionality of the respective storage ITE 1102. For example, the feature card 1108 may be configured in order to customize the storage ITE as a RAID ITE, network-attached storage (NAS) ITE, and/or file cache ITE, *etc.* The fanout cards 1104 of each storage ITE 1102 provide network connectivity for the storage ITE 1102 and/or improve availability of the respective storage ITE 1102. Each fanout card 1104 includes one or more fanout components 1112. Each storage ITE 1102 may further include a storage interconnect $1110_1$, $1110_2$ that operatively connects the components of the respective storage ITE 1102. In one embodiment, the storage interconnects $1110_1$, $1110_2$ provide Serial Attached SCSI (SAS) connectivity between the components of the storage ITEs 1102. In alternative embodiments, the storage interconnects provide FCoE or Serial ATA (SATA) connectivity.

In one embodiment, the configuration 1100 further includes a switch interconnect, along with cabling 1114 between the storage interconnects 1110 of the storage ITEs 1102. The cabling 1114 may connect a fanout card $1104_1$ of the first storage ITE $1102_1$ to a fanout card $1104_4$ of the second storage ITE $1102_2$. Depending on the embodiment, to provide increased bandwidth, the switch interconnect may include cabling 1116 between an additional fanout card of the each storage ITE 1102. Accordingly, when the switch module $206_1$ of the first storage ITE $1102_1$ fails, a switch module $206_2$ of the second storage ITE $1102_2$ may switch network traffic for the first storage ITE $1102_1$ – in addition to switching network traffic for the second storage ITE $1102_2$. Packaging the storage ITEs 1102 to include cabling 1114 between the storage interconnects 1110 eliminates each switch module $206_1$, $206_2$ as a SPOF, using the storage interconnects 1110 of the storage ITEs 1102.

Figure 12 illustrates a configuration 1200 of a pair of interposer cards 204, where the configuration 1200 includes an interposer interconnect, according to one embodiment of the invention. Rather than using the storage interconnects to eliminate SPOFs, the configuration 1200 includes cabling between CNAs associated with different interposer cards to eliminate SPOFs. Each interposer card 204 includes a switch module 206 and operatively connects two server cards 202 to a midplane. The switch module 206 of each interposer card 204 is configured to switch network traffic for the server cards 202 coupled with the respective interposer card 204. Each interposer card 204 further includes one or more CNAs 302. Each server card 202 includes one or more CPUs 106. Depending on the embodiment, each server card 202 further includes one or more CNAs 304. The configuration 1200 of the interposer cards 202 may also include cabling 1202 between the interposer cards 202. The cabling 1202 may connect the CNAs $304_1$, $304_2$ of the server cards 202 coupled with the first interposer card $204_1$, to the switch module $206_2$ of the second interposer card $204_2$. The cabling 1202 may also connect the CNAs $304_3$, $304_4$ of the server cards 202 coupled with the second interposer card $204_2$, to the switch module $206_1$ of the first interposer card $204_1$.

Should the switch module $206_1$ of the first interposer card $204_1$ fail, the switch module $206_2$ of the second interposer card $204_2$ manages network traffic for the server cards 202 coupled with the first interposer card $204_1$ – in addition to switching network traffic for the server cards 202 coupled with the second interposer card $204_2$. Thus, the configuration 1200 eliminates each of the switch modules 206 as an SPOF. In other words, the server cards 202

coupled with each interposer card retain connectivity to the switch fabric and/or maintain redundancy thereof, even upon failure of one of the switch modules 206.

As described above, packaging the server system to include one or more switch cards coupled with the midplane may eliminate the switch module as an SPOF. In such a case, the midplane includes a fabric interconnect for a switch fabric. The midplane may couple one or more chassis interconnect element (CIE) boards, which themselves include one or more management processing subsystems. Although embodiments are herein described with reference to the switch cards and the CIE boards as being separate components, depending on the embodiment, some or all of the functionality of the CIE boards may be integrated into the switch cards. The switch cards and/or the CIE boards may be coupled with a first face of the midplane, and one or more server cards may be coupled with a second face of the midplane.

In one embodiment, the switch cards and/or the CIE boards may be aligned along a first axis, and the server cards may be aligned along a second axis. Further, the first axis is at least substantially perpendicular to the second axis. For example, the switch cards may be vertically coupled with the first face of the midplane, and the server cards may be horizontally coupled with the second face of the midplane, or vice versa. Depending on the embodiment, the switch fabric includes wiring that connects each switch card with each server card and/or wiring that connects each switch card to each other switch card. Such wiring provides redundant pathing to reduce and/or eliminate SPORs and/or SPOFs in the switch fabric. Further, coupling the switch cards and the server cards with the midplane along perpendicular axes may simplify and/or reduce the amount of the wiring required (at least in some cases).

Figure 13 illustrates a configuration 1300 of a server system that includes multiple switch cards 1302, according to one embodiment of the invention. As shown, the configuration 1300 includes a midplane 123, server cards 202 and CIE boards 1304. The server cards 202 are horizontally coupled with the midplane 123, and the switch cards 1302 and the CIE boards 1304 are vertically coupled with the midplane 123. The switch cards 1302 and/or the server cards 202 may be hot-swappable from the midplane 123. Each switch card 1302 may include one or more switch modules 206, and each server card 202 may include one or more CPUs.

Depending on the embodiment, the switch fabric may include wiring between each switch card 1302 or each switch module 206 to each server card 202 or each CPU. The switch fabric may also include wiring between each switch card 1302 or switch module 206 to each other switch card 1302 or switch module 206. Accordingly, the configuration 1300 provides

redundant pathing between elements in the switch fabric, thereby eliminating a switch module 206 and/or a switch card 1302 as an SPOF in the switch fabric. Specifically, the server cards 202 retain connectivity to the switch fabric upon failure of a switch module 206 or a switch card 1302. Further, a repair action on a failed switch module 206 or switch card 1302 does not impact the connectivity of the server cards 202 to the switch fabric. The repair action may include replacing the switch card 206 with a second switch card having a functional switch module and without restarting the server system and/or the switch fabric.

Figure 14 illustrates a logical view 1400 of a configuration of a server system that includes multiple switch cards, according to one embodiment of the invention. As shown, the logical view 1400 includes multiple server cards 202 and multiple switch modules 206, 1304 of a server system. The server system may also include a fabric interconnect for a switch fabric. The switch modules 206 are disposed proximate to the server cards 202 in the switch fabric and provide redundancy in switching network traffic for the server cards 202. The switch modules 206 may also be referred to as north switch modules. The switch modules 1404 are disposed proximate to the north switch modules and provide connectivity between the north switch modules and the rest of the switch fabric. The switch modules 1404 may also be referred to as south switch modules.

In one embodiment, each server card 202 includes two CPUs 106 and two CNAs 304. The switch fabric may include wiring 1404 between each switch module 206 and each server card 202. The switch fabric may further include a local rack interconnect 1402, which in turn includes wiring between the switch modules 206 and the switch modules 1404. Depending on the embodiment, the local rack interconnect 1402 may also include wiring between each of the switch modules 206 and each other one of the switch modules 206, thereby providing all-to-all connectivity among the switch modules 206. Advantageously, the configuration eliminates each switch module 206 as an SPOF in the switch fabric. For example, should switch module $206_1$ fail, then the switch module $206_2$ may route network traffic for each server card 202. Doing so allows each server card 202 to remain connected to the switch fabric, despite the failure in switch module $206_1$.

Figure 15 illustrates a configuration 1500 of a server system that includes multiple switch cards 1302, according to one embodiment of the invention. As shown, the multiple switch cards 1302 each include two switch modules 206, or north switch modules. Each switch module manages network traffic for at least one server card 202. The configuration 1500

includes wiring 1404 between each server switch module 206 and each of one subset of the server cards 202. For example, subsets of the server cards 202 may include a first subset having the server cards $202_{1-7}$ and a second subset having the server cards $202_{8-14}$. The configuration 1500 may also include all-to-all wiring (not shown) among each subset of the switch modules 206. For example, the subsets of the switch modules 206 may include a first subset having the switch modules $206_{1-4}$ and a second subset having the switch modules $206_{5-8}$. In other words, the switch cards 1302 and/or the server cards 202 may be physically partitioned into different subsets in terms of the wiring in the switch fabric, where the wiring provides connectivity and/or redundancy in the switch fabric. The configuration 1500 may also include further wiring across the subsets of switch modules 206.

In one embodiment, the switch cards $1302_{1,3}$ further include one or more management processing subsystems. The management processing subsystems include management firmware configured to monitor health of the server system and/or switch fabric elements, configure the elements, and/or detect and respond to failure of the elements. As shown, the management processing subsystems include an input/output master controller (IoMC) and a chassis service element (CSE). The IoMC manages elements in switch fabric, while the CSE includes manages components in the server chassis. Further, the IoMCs may be operatively connected to one another via an IoMC interconnect on the switch cards 1302, to provide redundancy in monitoring and/or managing the switch fabric, where the IoMC interconnect includes physical wiring between the IoMCs. Upon a failure of a first IoMC $1506_1$, a second IoMC $1506_2$ may be configured to provide switch fabric monitoring and/or management capabilities in lieu of the first IoMC $1506_1$. Advantageously, the configuration 1500 eliminates each switch module 206 as an SPOF and further eliminates each IoMC 1506 as an SPOF in the switch fabric.

In one embodiment, the switch cards $1302_{2,4}$ further include switch modules 1404, or south switch modules. As described above, the south switch modules provide connectivity between the north switch modules and the rest of the switch fabric. Each switch card 1302 further includes local links (L links) 1502 and distance links (D links) 1504. The L link 1502 provides the switch card 1302 with physical connectivity to another chassis within a single physical frame. The D link 1504 provides the switch card 1302 with physical connectivity to another chassis across physical frames. Accordingly, the switch fabric thereby includes additional redundant pathing for elements of the switch fabric.

Figure 16 illustrates a server system 1600 that includes a management controller 1604 and an independent management interconnect 1606, according to one embodiment of the invention. As shown, the server system 1600 includes multiple frames 804, each of which includes one or more chassis 802 that are operatively connected via chassis link cards 806 and associated cabling 808. Further, the chassis 802 may be operatively connected across two frames via frame link cards 810 and associated cabling 812. Each chassis 802 houses a midplane 123 that is coupled with one or two server cards 202 via an interposer card 204 according to the techniques disclosed herein. Each interposer card 204 includes a switch module 206 for switching traffic for the server cards 202. The midplanes 123 include fabric wiring for interconnecting the switch modules 206. Together, the switch modules 206, the fabric wiring, the chassis link cards 806 and associated cabling 808, and the frame link cards 810 and associated cabling 812 form a switch fabric for the server system 1600. The management controller 1604 may be included in a management appliance 1602 and may be configured to manage the switch modules 206 via the management interconnect 1606, independently from the switch fabric.

In one embodiment, the management interconnect 1606 may be configured to provide dedicated pathing between the management controller 1604 and the switch modules 206. Doing so allows the management controller 1604 to manage the switch modules 206. In one embodiment, the management interconnect is an out-of-band network physically isolated from the switch fabric. Such an out-of-band network serves as a dedicated management channel for managing the switch modules 206. For example, the management controller 1604 may be operatively connected to multiple integrated management modules (IMMs) via the dedicated pathing. Each IMM may be a service processor that is operatively connected to a respective switch module 206. Together, the IMMs communicate and/or coordinate with the management controller 1604 via the management interconnect 1606 to manage the switch modules 206 of the switch fabric.

Figure 17 illustrates a server system 1700 that includes a management controller 1604 without also including an independent management interconnect 1606, according to one embodiment of the invention. In one embodiment, the server system 1700 may be configured to perform fabric management via the switch fabric, rather than via the independent management interconnect 1606. In such a case, the switch fabric may include wiring between the management controller 1604 and the switch fabric. In one embodiment, the wiring may physically connect the management controller 1604 with each switch module 206.

Eliminating the independent management interconnect from the packaging of the server system 1700 may reduce wiring in the server system 1700 at least in some cases, reducing costs and/or complexity of packaging the server system.

In one embodiment, the management controller 1604 monitor and respond to predefined events, including, *e.g.*, changes in the switch fabric, such as which a physical addition or removal of a switch module 206, changes in connectivity of a switch module 206, failure of a switch module 206, *etc.* In one embodiment, when managing the switch modules 206, the management controller 1604 may perform one or more operations on the switch modules 206. For example, the management controller may power on (or off) a switch module 206, , communicate with a switch module 206, update firmware in a switch module 206, restart a switch module 206, initialize a switch module 206, monitor a switch module 206, and configure a switch module. In one embodiment, these operations may be performed in response to a user request (*e.g.*, from an administrative user). The operations may also include determining that a switch module 206 has failed and, in response notifying other switch modules in the fabric network traffic that switch module 206 has failed. In response, the other switch modules may be configured to switch network traffic for the failed switch module 206 indication. Further, once the failed switch module is replaced, the management controller 1604 may reintegrate the functional switch module into the switch fabric. The management controller 1604 may then direct the functional switch module to switch network traffic for the failed (and replaced) switch module. The management controller 1604 may also direct the one or more other switch modules to stop switching network traffic for the now-replaced switch module. In one embodiment, to monitor the switch fabric, the management controller 1604 may communicate with multiple switch modules 206 over the switch fabric to capture data characterizing performance of the switch fabric. The management controller 1604 may then use the captured data to provide an end-to-end view of utilization levels of the switch fabric.

One example of a predefined event is an internal error occurring in a component of the switch module 206, causing the switch module 206 to cease its switching function. In one embodiment, the management controller 1604 detects the error over the switch fabric because of adjacent switch modules no longer being able to communicate with the failed switch module 206. Accordingly, the management controller 1604 may indicate for a user to replace the failed switch module 206. After the switch module 206 is replaced, the management

controller 1604 may re-integrate the new switch module into the switch fabric, allowing the new switch module to switch network traffic for the switch fabric.

Although some embodiments are described herein with reference to a server system that does not include an independent management interconnect, other embodiments are broadly contemplated. For example, in an alternative embodiment, the server system may include one or more independent interconnects other than the switch fabric, but the independent interconnects are not used (or configured to be used) by the management controller in managing the switch modules.

Figure 18 is a flowchart depicting a method 1800 for reducing wiring in a server system that includes a management controller, according to one embodiment of the invention. As shown, the method 1800 begins at step 1810, where a provider of a server system packages the server system to include a midplane that in turn includes a fabric interconnect for a switch fabric. At step 1820, the provider of the server system further packages the server system to include one or more server cards operatively connected to the midplane, where each server card includes one or more computer processors and a memory. At step 1830, the provider of the server system further packages the server system to include switch modules operatively connected to the midplane, each switch module configured to switch network traffic for at least one server card. At step 1840, the provider of the server system further packages the server system to include a management controller configured to manage the switch modules via the fabric interconnect. After the step 1840, the method 1800 terminates.

Figure 19 is a flowchart depicting a method 1900 for managing a switch fabric, according to one embodiment of the invention. As shown, the method 1900 begins at step 1910, where a provider of a server system packages the server system to include a midplane, one or more server cards and switch modules. The midplane includes a fabric interconnect for a switch fabric, and each server card and each switch module is operatively connected to the midplane. Further, each switch module is configured to switch network traffic for at least one server card. The server system further includes a management controller. At step 1920, the management controller detects that a switch module has failed. At step 1930, upon detecting that the switch module has failed, the management controller sends, via the fabric interconnect, an indication to one or more other switch modules to switch network traffic for the switch module that has failed. After the step 1930, the method 1900 terminates.

Advantageously, embodiments of the invention reduce wiring requirements in server systems which use a management controller to manage a switch fabric. In one embodiment, the server system may includes a midplane that in turn includes a fabric interconnect for the switch fabric. The server system further includes one or more server cards operatively connected to the midplane, where each server card includes one or more computer processors and a memory. The server system further includes switch modules operatively connected to the midplane, each switch module being configured to switch network traffic for at least one server card. The management controller may be configured to manage the switch modules via the fabric interconnect. Because no other interconnect is required between the switch modules for the switch modules to be managed, wiring in the server system is thereby reduced. Doing so reduces the costs and/or complexity of packaging the server system (at least in some cases).

Further, some embodiments of the invention may also reduce impact of a switch failure in the switch fabric. The server cards may be coupled with the midplane, where each server card is hot-swappable from the midplane. The switch modules may be included in one or more switch cards that are also coupled with the midplane and whereby the one or more switch cards are operatively connected to the one or more server cards. Each switch card may also be also hot-swappable from the midplane. To provide redundant pathing, the switch fabric may include wiring between each switch card and each server card and/or wiring between each switch card and each other switch card. Accordingly, when the switch module of a first switch card fails, the management controller directs a second switch module on the first switch card or on a second switch card to route network traffic for the server cards supported by the switch module of the first switch card. Accordingly, this configuration of the server system eliminates the switch module of the first switch card as an SPOF.

Further still, embodiments of the invention may reduce impact of repair actions on the switch fabric. In one embodiment, a server system may be configured such that upon a failure of the switch module of the first switch card, the first switch card may be replaced with a third switch card having a functional switch module, without powering off or restarting the server system and/or switch fabric. Thus, such a configuration improves availability of the server system and/or the switch fabric while reducing costs associated with repair actions.

While the foregoing is directed to embodiments of the present invention, other and further embodiments of the invention may be devised without departing from the basic scope thereof, and the scope thereof is determined by the claims that follow.

## CLAIMS

1. A system, comprising:

a midplane comprising a fabric interconnect for a switch fabric;

one or more server cards operatively connected to the midplane, each server card comprising one or more computer processors and a memory;

a plurality of switch modules operatively connected to the midplane, each switch module configured to switch network traffic for at least one of the one or more server cards; and

a management controller configured to manage the plurality of switch modules via the fabric interconnect.

2. The system of claim 1, wherein the only interconnect between the switch modules and the management controller is the fabric interconnect.

3. The system of claim 1, wherein the only interconnect used by the management controller in managing the plurality of switch modules is the fabric interconnect.

4. The system of claim 1, wherein managing the plurality of switch modules comprises at least one of powering on, powering off, communicating with, updating firmware in, restarting, initializing, monitoring, and configuring at least one switch module of the plurality of switch modules.

5. The system of claim 1, wherein managing the plurality of switch modules comprises:

performing a predefined action responsive to determining that at least one switch module has failed.

6. The system of claim 5, wherein the predefined action comprises:

sending an indication to one or more other switch modules to switch network traffic for the at least one switch module that has failed.

7. The system of claim 1, wherein at least one of the plurality of switch modules is configured to be replaced without requiring a restart of the system and without requiring a restart of the switch fabric.

8. A management controller, comprising:

a computer processor; and

a memory storing management firmware which, when executed on the computer processor, performs an operation comprising:

managing a plurality of switch modules operatively connected to a midplane of a server system, wherein the midplane comprises a fabric interconnect for a switch fabric, wherein the plurality of switch modules is managed via the fabric interconnect, wherein the midplane is operatively connected to one or more server cards, and wherein each switch module is configured to switch network traffic for at least one of the server cards.

9. The management controller of claim 8, wherein the only interconnect between the switch modules and the management controller is the fabric interconnect.

10. The management controller of claim 8, wherein the only interconnect used by the management controller in managing the plurality of switch modules is the fabric interconnect.

11. The management controller of claim 8, wherein managing the plurality of switch modules comprises at least one of powering on, powering off, communicating with, updating firmware in, restarting, initializing, monitoring, and configuring at least one switch module of the plurality of switch modules.

12. The management controller of claim 8, wherein managing the plurality of switch modules comprises:

performing a predefined action responsive to determining that at least one switch module has failed.

13. The management controller of claim 12, wherein the predefined action comprises:

sending an indication to one or more other switch modules to switch network traffic for the at least one switch module that has failed.

14. The management controller of claim 8, wherein at least one of the plurality of switch modules is configured to be replaced without requiring a restart of the server system and without requiring a restart of the switch fabric.

15. A computer-implemented method, comprising:

providing a server system comprising a midplane, one or more server cards, and a plurality of switch modules, wherein the midplane comprises a fabric interconnect for a switch fabric, wherein each server card and each switch module is operatively connected to the midplane, and wherein each switch module is configured to switch network traffic for at least one of the one or more server cards; and

managing the plurality of switch modules via the fabric interconnect and by operation of one or more computer processors.

16. The computer-implemented method of claim 15, wherein the only interconnect between the switch modules and the management controller is the fabric interconnect.

17. The computer-implemented method of claim 15, wherein the only interconnect used by the management controller in managing the plurality of switch modules is the fabric interconnect.

18. The computer-implemented method of claim 15, wherein managing the plurality of switch modules comprises at least one of powering on, powering off, communicating with, updating firmware in, restarting, initializing, monitoring, and configuring at least one switch module of the plurality of switch modules.

19. The computer-implemented method of claim 15, wherein managing the plurality of switch modules comprises:

performing a predefined action responsive to determining that at least one switch module has failed.

20. The computer-implemented method of claim 19, wherein the predefined action comprises:

sending an indication to one or more other switch modules to switch network traffic for the at least one switch module that has failed.

INTELLECTUAL PROPERTY OFFICE

| | | | |
|---|---|---|---|
| **Application No:** | GB1208064.4 | **Examiner:** | Dr Mark Lewney |
| **Claims searched:** | 1-20 | **Date of search:** | 28 August 2012 |

## Patents Act 1977: Search Report under Section 17

**Documents considered to be relevant:**

| Category | Relevant to claims | Identity of document and passage or figure of particular relevance |
|---|---|---|
| X | 1-20 | WO2007/064466 A2 (MOTOROLA) - See especially fig. 2 and accompanying description. |
| X | 1-20 | US2002/124114 A1 (BOTTOM ET AL.) - See especially fig. 1 A and paragraphs [0022-33]. |
| X | 1-6, 8-13 & 15-20 | US2007/083707 A1 (HOLLAND ET AL.) - See especially fig. 2b and accompanying description. |
| X | 1, 8 & 15 | US2010/165983 A1 (AYBAY ET AL.) - See especially fig. 4 and accompanying description. |

Categories:

| | | | |
|---|---|---|---|
| X | Document indicating lack of novelty or inventive step | A | Document indicating technological background and/or state of the art. |
| Y | Document indicating lack of inventive step if combined with one or more other documents of same category. | P | Document published on or after the declared priority date but before the filing date of this invention. |
| & | Member of the same patent family | E | Patent document published on or after, but with priority date earlier than, the filing date of this application. |

**Field of Search:**

Search of GB, EP, WO & US patent documents classified in the following areas of the UKC$^{X}$ :

Worldwide search of patent documents classified in the following areas of the IPC

G06F; H04L; H05K

The following online and other databases have been used in the preparation of this search report

WPI, EPODOC.

**International Classification:**

| Subclass | Subgroup | Valid From |
|---|---|---|
| G06F | 0013/40 | 01/01/2006 |