



(43) International Publication Date
23 April 2015 (23.04.2015)

(51) International Patent Classification:
G06F 15/16 (2006.01)

(21) International Application Number:
PCT/US2013/065564

(22) International Filing Date:
18 October 2013 (18.10.2013)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **EMPIRE TECHNOLOGY DEVELOPMENT LLC** [US/US]; 2711 Centerville Road, Suite 400, Wilmington, DE 19808 (US).

(72) Inventor: **KURABAYASHI, Shuichi**; B-103, 2-3-5, Engyo, Fujisawa-shi, Kanagawa 252-0805 (JP).

(74) Agent: **LEE, David S.**; Brundidge & Stanger, P.C., 2318 Mill Road, Suite 1020, Alexandria, VA 22314 (US).

(81) Designated States (*unless otherwise indicated, for every kind of national protection available*): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY,

BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

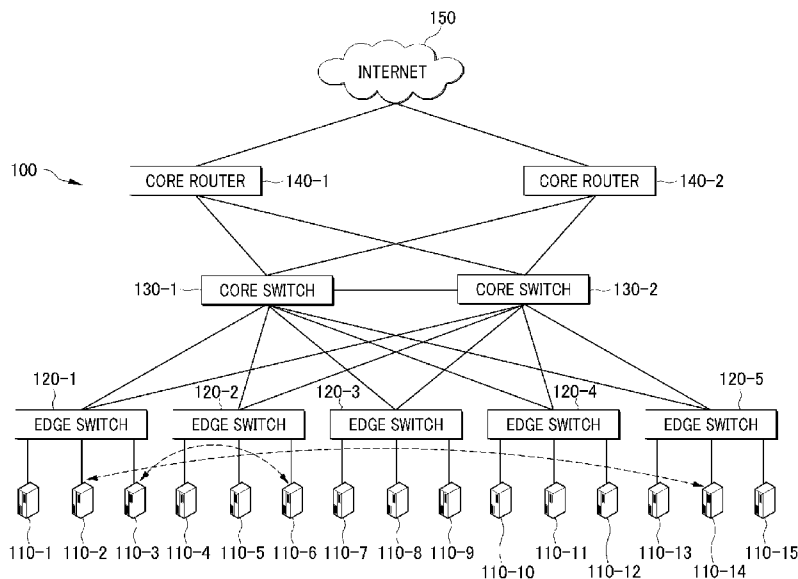
(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU,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

Published:

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

(54) Title: FAILURE RECOVERY SCHEME FOR A CLOUD SYSTEM

FIG. 1



(57) Abstract: Technologies are generally described for a failure recovery scheme for a cloud system. In some examples, the cloud system may include one or more computing nodes, and one or more network switches configured to relay one or more packets among the one or more computing nodes. A respective one of the computing nodes may include a first processor configured to process the one or more packets to communicate with at least one of the network switches, and a second processor configured to process the one or more packets to communicate with at least one of the other computing nodes in the cloud system.

FAILURE RECOVERY SCHEME FOR A CLOUD SYSTEM

BACKGROUND

[0001] Cloud computing services such as, for example, infrastructure-as-a-service (IaaS), platform-as-a-service (PaaS), and software-as-a-service (SaaS), have become popular. A cloud system allows users to enjoy large-scale computing power by distributing tasks into large amount of computing nodes. The computing nodes are connected to the Internet via network switches such as, for example, edge switches, core switches, and/or core routers. If a failure occurs at one of the network switches, such a failure may propagate throughout the entire cloud system, resulting in a large-scale system failure.

SUMMARY

[0002] In an example, a cloud system may include one or more computing nodes, and one or more network switches configured to relay one or more packets among the one or more computing nodes, and a respective one of the computing nodes may include a first processor configured to process the one or more packets to communicate with at least one of the network switches, and a second processor configured to process the one or more packets to communicate with at least one of the other computing nodes in the cloud system.

[0003] In another example, a computing node may include a first processor configured to process one or more packets to communicate with a network switch, and a second processor configured to process the one or more packets to communicate with another computing node.

[0004] In yet another example, a method performed under control of a computing node may include detecting a failure in communicating one or more packets with a network switch based on a first communication protocol, and transferring the one or more packets directly to a target computing node based on a second communication protocol, in response to the detecting of the failure.

[0005] The foregoing summary is illustrative only and is not intended to be in any way limiting. In addition to the illustrative aspects, embodiments, and features described above, further aspects, embodiments, and features will become apparent by reference to the drawings and the following detailed description.

BRIEF DESCRIPTION OF THE FIGURES

[0006] The foregoing and other features of this disclosure will become more apparent from the following description and appended claims, taken in conjunction with the accompanying drawings. Understanding that these drawings depict only several embodiments in accordance with the disclosure and are, therefore, not to be considered limiting of its scope, the disclosure will be described with additional specificity and detail through use of the accompanying drawings, in which:

[0007] **Fig. 1** schematically shows an illustrative example of a cloud system including multiple computing nodes, arranged in accordance with at least some embodiments described herein;

[0008] **Fig. 2** shows a schematic block diagram illustrating an example architecture of a computing node for implementing a failure recovery scheme, arranged in accordance with at least some embodiments described herein;

[0009] **Fig. 3** shows an example flow diagram of a process for implementing a failure recovery scheme, arranged in accordance with at least some embodiments described herein;

[0010] **Fig. 4** schematically shows packet transmission between computing nodes in a cloud system, arranged in accordance with at least some embodiments described herein; and

[0011] **Fig. 5** schematically shows a protocol stack employed by a computing node, arranged in accordance with at least some embodiments described herein.

DETAILED DESCRIPTION

[0012] In the following detailed description, reference is made to the accompanying drawings, which form a part hereof. In the drawings, similar symbols typically identify similar components, unless context dictates otherwise. The

illustrative embodiments described in the detailed description, drawings, and claims are not meant to be limiting. Other embodiments may be utilized, and other changes may be made, without departing from the spirit or scope of the subject matter presented herein. It will be readily understood that the aspects of the present disclosure, as generally described herein, and illustrated in the drawings, can be arranged, substituted, combined, separated, and designed in a wide variety of different configurations, all of which are explicitly contemplated herein.

[0013] This disclosure is generally drawn, *inter alia*, to methods, apparatuses, systems, devices, and computer program products related to a failure recovery scheme for a cloud system. Further, technologies are herein generally described for a computing node in the cloud system configured to implement the failure recovery scheme.

[0014] In some examples, the cloud system may include one or more computing nodes, and one or more network switches configured to relay one or more packets among the one or more computing nodes. A respective one of the computing nodes may include a first processor configured to process the one or more packets to communicate with at least one of the network switches, and a second processor configured to process the one or more packets to communicate with at least one of the other computing nodes in the cloud system when the first processor is not able to communicate the one or more packets with the at least one of the network switches. The second processor may be a network interface controller (NIC) in which a field programmable gate array (FPGA) is embedded, i.e., an FPGA-embedded NIC. The network switches may include, but are not limited to, edge switches, core switches, and/or core routers.

[0015] In some examples, the respective computing nodes may be configured to detect a failure in communicating the one or more packets with the at least one of the network switches, and transfer the one or more packets directly to at least one of the other computing nodes in the cloud system (i.e., a target computing node) in response to the detecting of the failure. By way of example, but not limitation, the respective computing nodes may calculate a distance between an IP (Internet Protocol) address of a destination computing node (i.e., a computing node to which

the one or more packets to be sent) and an IP address of each of the other computing nodes connected to the computing node, and select the target computing node as one of the other computing nodes that is closest to the destination computing node. In some further examples, when the respective one of the computing nodes detects a recovery from the failure, it may transfer the one or more packets to the at least one of the network switches, in response to the detecting of the recovery. That is, the respective computing nodes may communicate the one or more packets directly with one another when the failure occurs, while communicate the one or more packets via the network switches when there is no failure in the cloud system.

[0016] In some examples, the second processors included in the respective computing nodes may communicate with one another using an overlay network protocol between link layer and network layer, so that the failure recovery scheme described herein may be implemented without modifying existing legacy software and/or hardware.

[0017] Fig. 1 schematically shows an illustrative example of a cloud system 100 including multiple computing nodes, arranged in accordance with at least some embodiments described herein.

[0018] As depicted, cloud system 100 may include computing nodes 110-1, 110-2, 110-3, 110-4, 110-5, 110-6, 110-7, 110-8, 110-9, 110-10, 110-11, 110-12, 110-13, 110-14 and 110-15 (collectively, computing nodes 110). Each of computing nodes 110 may serve as a client or a server (e.g., a web server, an application server, a storage, etc.), depending on the desired implementation. Respective computing nodes 110 may be connected to one of edge switches 120-1, 120-2, 120-3, 120-4 and 120-5 (collectively, edge switches 120). For instance, as depicted, computing nodes 110-1, 110-2 and 110-3 may be connected to edge switches 120-1; computing nodes 110-4, 110-5 and 110-6 may be connected to edge switches 120-2; computing nodes 110-7, 110-8 and 110-9 may be connected to edge switches 120-3; computing nodes 110-10, 110-11 and 110-12 may be connected to edge switches 120-4; and computing nodes 110-13, 110-14 and 110-15 may be connected to edge switches 120-5. Further, respective edge switches 120-1, 120-2, 120-3, 120-4 and 120-5 may be connected to core switches 130-1 and 130-2 (collectively, core switches 130), and respective core

switches 130-1 and 130-2 may be connected to core routers 140-1 and 140-2 (collectively, core routers 140). In this regard, respective computing nodes 110 may communicate with the Internet 150 via at least some of edge switches 120, core switches 130, and core routers 140. That is, each of computing nodes 110 in cloud system 100 may transmit/receive one or more packets via at least some of edge switches 120, core switches 130, and core routers 140 to/from other computing nodes external to cloud system 100 and/or other computing nodes 110 in cloud system 100. Edge switches 120, core switches 130 and core routers 140 may be collectively called as network switches.

[0019] In some embodiments, each of computing nodes 110 in cloud system 100 may also directly communicate with other computing nodes 110 in cloud system 100. For instance, as depicted, computing node 110-2 may directly communicate with computing node 110-14, and computing node 110-3 may directly communicate with computing node 110-6. Such a peer-to-peer (P2P) communication may be activated and/or enabled when at least one of computing nodes 110 detects a failure in communicating one or more packets with network switches 120, 130 and/or 140 (i.e., at least some of edge switches 120, core switches 130, and core routers 140). The peer-to-peer communication among computing nodes 110 may construct a failure recovery network.

[0020] Fig. 2 shows a schematic block diagram illustrating an example architecture of computing node 110 for implementing a failure recovery scheme, arranged in accordance with at least some embodiments described herein.

[0021] Computing node 110 may be one of computing nodes 110-1, 110-2, 110-3, 110-4, 110-5, 110-6, 110-7, 110-8, 110-9, 110-10, 110-11, 110-12, 110-13, 110-14 and 110-15 that may be included in cloud system 100. In this regard, reference may be made to the embodiments depicted and described with reference to Fig. 1.

[0022] As depicted, computing node 110 may include a first processor 210, a second processor 220, a first socket 230 and a second socket 240. Although illustrated as discrete components, various components may be divided into additional components, combined into fewer components, or eliminated while being

contemplated within the scope of the disclosed subject matter. It will be understood by those skilled in the art that each function and/or operation of the components may be implemented, individually and/or collectively, by a wide range of hardware, software, firmware, or any combination thereof.

[0023] First processor 210 may be configured to process one or more packets to communicate with at least one of network switches 120, 130 and/or 140. First processor 210 may be connected to at least one of network switches 120, 130 and/or 140 via first socket 230. By way of example, but not limitation, first socket 230 may be an Ethernet socket.

[0024] Second processor 220 may be configured to process the one or more packets to communicate with another computing node 110 in cloud system 100. Second processor 220 may be connected to the other computing node 110 via second socket 240. By way of example, but not limitation, second processor 220 may be a network interface controller (NIC) in which a programmable processor, such as a field programmable gate array (FPGA) configured to implement L3 packet processing, may be embedded. By way of example, but not limitation, second socket 240 may also be an Ethernet socket.

[0025] In some embodiments, computing node 110 may be configured to enable second processor 220 to process the one or more packets when first processor 210 is not able to communicate the one or more packets with the at least one of network switches 120, 130 and/or 140 (e.g., when a failure occurs in communicating the one or more packets with the at least one of network switches 120, 130 and/or 140). In some further embodiments, computing node 110 may be configured to enable first processor 210 to process the one or more packets and disable second processor 220, when first processor 210 is normally able to communicate the one or more packets with the at least one of network switches 120, 130 and/or 140 (e.g., when the failure is recovered).

[0026] In some embodiments, when cloud system 100 detects the failure in communicating the one or more packets with the at least one of network switches 120, 130 and/or 140, cloud system 100 may determine, from among other computing nodes connected to computing node 110, a target computing node to transfer the one

or more packets. By way of example, but not limitation, computing node 110 may calculate a distance between an IP (Internet Protocol) address of a destination computing node and an IP address of each of the other computing nodes connected to computing node 110 (e.g., based on integer subtraction between the IP addresses), and select the target computing node as one of the other computing nodes that is closest to the destination computing node. Then, in some embodiments, computing node 110 may transfer the one or more packets directly to the target computing node. Packet transmission between computing node 110 and the other computing nodes in cloud system 100 will be described below with reference to Fig. 4.

[0027] In some embodiments, first processor 210 may communicate the one or more packets with the at least one of network switches 120, 130 and/or 140 based on a first communication protocol, while second processor 220 may transfer the one or more packets to the target computing node based on a second communication protocol. The second communication protocol may be a peer-to-peer communication protocol between computing node 110 and the other computing nodes in cloud system 100. In some embodiments, the second communication protocol may be an overlay network protocol between link layer and network layer, that may be implemented by the field programmable gate array (FPGA) included in second processor 220. Protocol stack that may be employed by second processor 220 will be described below with reference to Fig. 5.

[0028] Fig. 3 shows an example flow diagram of a process 300 for implementing a failure recovery scheme, arranged in accordance with at least some embodiments described herein.

[0029] Process 300 may be implemented in a computing node such as computing node 110 including, for example, first processor 210, second processor 220, first socket 230 and second socket 240. Thus, reference may be made to the embodiments depicted and described with reference to Figs. 1 – 2. Process 300 may include one or more operations, actions, or functions as illustrated by one or more blocks 310, 320 and/or 330. Although illustrated as discrete blocks, various blocks may be divided into additional blocks, combined into fewer blocks, or eliminated, depending on the desired implementation. Processing may begin at block 310.

[0030] At block 310 (Detect Failure in Communicating Packets with Network Switch Based on First Communication Protocol), computing node 110 may detect a failure in communicating one or more packets with a network switch (e.g., at least one of edge switches 120, core switches 130, and core routers 140) based on a first communication protocol. Processing may continue from block 310 to block 320.

[0031] At block 320 (Determine Target Computing Node to Transfer Packets), computing node 110 may determine, from among other computing nodes connected to computing node 110, a target computing node to transfer the one or more packets. In some embodiments, computing node 110 may calculate a distance between an IP (Internet Protocol) address of a destination computing node and an IP address of each of the other computing nodes connected to computing node 110 (e.g., based on integer subtraction between the IP addresses). Then, computing node 110 may select the target computing node as one of the other computing nodes that is closest to the destination computing node, as will be described below with reference to Fig. 4. Processing may continue from block 320 to block 330.

[0032] At block 330 (Transfer Packets Directly to Target Computing Node Based on Second Communication Protocol), computing node 110 may transfer the one or more packets directly to the target computing node based on a second communication protocol, which may be implemented by, for example, a field programmable gate array (FPGA) included in computing node 110. In some embodiments, the second communication protocol may be different from the first communication protocol, in that the second communication protocol may be a peer-to-peer communication protocol between computing node 110 and the other computing nodes. Further, the second communication protocol may be an overlay network protocol between link layer and network layer, as will be described below with reference to Fig. 5.

[0033] One skilled in the art will appreciate that, for this and other processes and methods disclosed herein, the functions performed in the processes and methods may be implemented in differing order. Furthermore, the outlined steps and operations are only provided as examples, and some of the steps and operations may be optional, combined into fewer steps and operations, or expanded into additional

steps and operations without detracting from the essence of the disclosed embodiments.

[0034] Fig. 4 schematically shows packet transmission between computing nodes in cloud system 100, arranged in accordance with at least some embodiments described herein. Reference may be made to the embodiments depicted and described with reference to Figs. 1 – 3.

[0035] As depicted, server racks 410-1, 410-2 ..., and 410-n may respectively include computing nodes 420-11, 420-12, 420-13, 420-14 and 420-15, computing nodes 420-21, 420-22, 420-23, 420-24 and 420-25, ..., and computing nodes 420-n1, 420-n2, 420-n3, 420-n4 and 420-n5. Respective computing nodes 420-11, 420-12, 420-13, 420-14, 420-15, 420-21, 420-22, 420-23, 420-24, 420-25, ..., 420-n1, 420-n2, 420-n3, 420-n4 and 420-n5 (collectively, computing nodes 420) may be components of cloud system 100, such as computing nodes 110.

[0036] In some embodiments, each of computing nodes 420 may include a network interface controller (NIC) in which a field programmable gate array (FPGA) capable of L3 packet processing is embedded (i.e., an FPGA-embedded NIC). Based on such a function of L3 packet processing, computing nodes 420 may directly communicate with one another when a failure occurs in cloud system 100.

[0037] In some embodiments, each of computing nodes 420 may determine, from among other computing nodes connected thereto, a target computing node to transfer one or more packets. By way of example, but not limitation, each of computing nodes 420 may determine the target computing node based on the relationships of IP (Internet Protocol) addresses of respective computing nodes 420, which may be expressed as 32-bit unsigned integers. The IP addresses of respective computing nodes 420 may range uniformly over a 32-bit integer space without overlapping, so that respective computing nodes 420 may calculate a distance with any other computing nodes in cloud system 100.

[0038] In some embodiments, as shown in Fig. 3, the FPGA-embedded NICs of computing nodes 420 are connected densely to nearby computing nodes, while sparsely to far computing nodes. In some further embodiments, the connections with the nearby computing nodes may be via a wireless network such as WLAN (Wireless

Local Area Network), thereby reducing the number of physical wires, while the connections with the far computing nodes may be wired such as via cables.

[0039] In some embodiments, each of computing nodes 420 may determine, from among the other computing nodes connected thereto, the target computing node as a computing node that is closest to a destination computing node. By way of example, but not limitation, upon receiving a packet originated from a source computing node, the FPGA-embedded NIC of each of computing nodes 420 may calculate the distances between the IP address of the destination computing node and the IP addresses of the other computing nodes connected thereto (e.g., by integer subtraction), determine, based on the calculation of the distances, the target computing node that is closest to the destination computing node, and transfer the packet to the determined target computing node. By repeating such packet transfer, the packet may reach the destination computing node. Since a search time converges in logarithmic time in cases where an IP address search range is narrowed to $1/d$ (where d is a natural number not less than 2) on each packet transfer, when the IP address search range is narrowed to half on each packet transfer (i.e., $d = 2$), $\log_2(N)$ may be less than 20 even when the number of computing nodes N is nearly 1 million. Accordingly, when the number of nodes is N , it may be possible to send a packet to the destination computing node with at most $O(\log_2(N))$ packet transfers. Further, fault tolerance may increase since no computing node has to be connected to all of the other computing nodes, and each computing node may operate symmetrically.

[0040] By way of example, but not limitation, it may be assumed that computing node 420-11, the IP address of which is 192.168.0.1, tries to send a packet to computing node 420-n2, the IP address of which is 192.168.0.52. That is, in this example, the source computing node is computing node 420-11, while the destination computing node is computing node 420-n2. In this example, it may also be assumed that source computing node 420-11 has physical connections to the other computing nodes in the same server rack 410-1 (i.e., computing nodes 420-12, 420-13, 420-14 and 420-15), computing nodes 420-21, 420-22, 420-23, 420-24 and 420-25 in neighboring server rack 410-2, and computing node 420-n1 in server rack 410-n, while does not have any direct physical connection to destination computing node

420-n2. Source computing node 420-11 may select a target computing node that is closest to destination computing node 420-n2 from among computing nodes 420-12, 420-13, 420-14, 420-15, 420-21, 420-22, 420-23, 420-24, 420-25 and 420-n1, which are respectively connected to source computing node 420-11. In such cases, source computing node 420-11 may calculate the distance between destination computing node 420-n2 and each of computing nodes 420-12, 420-13, 420-14, 420-15, 420-21, 420-22, 420-23, 420-24, 420-25 and 420-n1, by comparing magnitudes of the IP addresses (e.g., 32-bit unsigned integers) of destination computing node 420-n2 and each of computing nodes 420-12, 420-13, 420-14, 420-15, 420-21, 420-22, 420-23, 420-24, 420-25 and 420-n1. That is, source computing node 420-11 may search the IP address with the least difference from the IP address of destination computing node 420-n2, i.e., 192.168.0.52 (that may be converted into a 32-bit unsigned integer 3232235572). In this example, source computing node 420-11 may determine the target computing node as computing node 420-n1, the IP address of which is 192.168.0.51 (that may be converted into a 32-bit unsigned integer 3232235571), and send the packet to computing node 420-n1. Then, since computing node 420-n1 is connected to destination computing node 420-n2 (that is, computing node 420-n1 and destination computing node 420-n2 are in the same server rack 410-n), computing node 420-n1 may transfer the packet to destination computing node 420-n2.

[0041] Although Fig. 4 illustrates that server racks 410-1, 410-2 ..., and 410-n respectively include five computing nodes, those skilled in the art will readily appreciate that server racks 410-1, 410-2 ..., and 410-n may respectively include any number of computing nodes.

[0042] Fig. 5 schematically shows a protocol stack 500 employed by a computing node, arranged in accordance with at least some embodiments described herein. Reference may be made to the embodiments depicted and described with reference to Figs. 1 – 4.

[0043] As depicted, protocol stack 500 may include a physical layer 510 (e.g., 10GBASE-T), a link layer 520 (e.g., Ethernet), a network layer 530 (e.g., IP), a transport layer 540 (e.g., TCP (Transmission Control Protocol)), and an application layer 550 (e.g., HTTP (Hypertext Transfer Protocol), HTTPS (Hypertext Transfer

Protocol Secure)). An IP overlay network protocol 560 for enabling IP packet transfer between computing nodes may be implemented between link layer 520 and network layer 530.

[0044] In some embodiments, by embedding field programmable gate arrays (FPGAs) that may execute processing for IP overlay network protocol 560 in network interface controllers (NICs) in the computing nodes, thereby executing autonomous packet transfer processing, it may be possible to implement the failure recovery scheme described herein without modifying existing legacy software or the hardware of existing Ethernet networks. Further, by employing the FPGA-embedded NICs as an interface between link layer 520 and network layer 530, it may be possible to implement device drivers in compliance with network device models provided by existing operating systems. By way of example, but not limitation, Network Driver Interface Specification (NDIS), which is a network device driver model for Microsoft Windows, and defines an interface between link layer and network layer, may be applicable for implementing the failure recovery scheme described herein.

[0045] The present disclosure is not to be limited in terms of the particular embodiments described in this application, which are intended as illustrations of various aspects. Many modifications and variations can be made without departing from its spirit and scope, as will be apparent to those skilled in the art. Functionally equivalent methods and apparatuses within the scope of the disclosure, in addition to those enumerated herein, will be apparent to those skilled in the art from the foregoing descriptions. Such modifications and variations are intended to fall within the scope of the appended claims. The present disclosure is to be limited only by the terms of the appended claims, along with the full scope of equivalents to which such claims are entitled. It is to be understood that this disclosure is not limited to particular methods, reagents, compounds, compositions or biological systems, which can, of course, vary. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting.

[0046] With respect to the use of substantially any plural and/or singular terms herein, those having skill in the art can translate from the plural to the singular

and/or from the singular to the plural as is appropriate to the context and/or application. The various singular/plural permutations may be expressly set forth herein for sake of clarity.

[0047] It will be understood by those within the art that, in general, terms used herein, and especially in the appended claims (*e.g.*, bodies of the appended claims) are generally intended as "open" terms (*e.g.*, the term "including" should be interpreted as "including but not limited to," the term "having" should be interpreted as "having at least," the term "includes" should be interpreted as "includes but is not limited to," *etc.*). It will be further understood by those within the art that if a specific number of an introduced claim recitation is intended, such an intent will be explicitly recited in the claim, and in the absence of such recitation no such intent is present. For example, as an aid to understanding, the following appended claims may contain usage of the introductory phrases "at least one" and "one or more" to introduce claim recitations. However, the use of such phrases should not be construed to imply that the introduction of a claim recitation by the indefinite articles "a" or "an" limits any particular claim containing such introduced claim recitation to embodiments containing only one such recitation, even when the same claim includes the introductory phrases "one or more" or "at least one" and indefinite articles such as "a" or "an" (*e.g.*, "a" and/or "an" should be interpreted to mean "at least one" or "one or more"); the same holds true for the use of definite articles used to introduce claim recitations. In addition, even if a specific number of an introduced claim recitation is explicitly recited, those skilled in the art will recognize that such recitation should be interpreted to mean at least the recited number (*e.g.*, the bare recitation of "two recitations," without other modifiers, means at least two recitations, or two or more recitations). Furthermore, in those instances where a convention analogous to "at least one of A, B, and C, *etc.*" is used, in general such a construction is intended in the sense one having skill in the art would understand the convention (*e.g.*, "a system having at least one of A, B, and C" would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, *etc.*). In those instances where a convention analogous to "at least one of A, B, or C, *etc.*" is used, in general such a construction is

intended in the sense one having skill in the art would understand the convention (*e.g.*, "a system having at least one of A, B, or C" would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, *etc.*). It will be further understood by those within the art that virtually any disjunctive word and/or phrase presenting two or more alternative terms, whether in the description, claims, or drawings, should be understood to contemplate the possibilities of including one of the terms, either of the terms, or both terms. For example, the phrase "A or B" will be understood to include the possibilities of "A" or "B" or "A and B."

[0048] In addition, where features or aspects of the disclosure are described in terms of Markush groups, those skilled in the art will recognize that the disclosure is also thereby described in terms of any individual member or subgroup of members of the Markush group.

[0049] As will be understood by one skilled in the art, for any and all purposes, such as in terms of providing a written description, all ranges disclosed herein also encompass any and all possible subranges and combinations of subranges thereof. Any listed range can be easily recognized as sufficiently describing and enabling the same range being broken down into at least equal halves, thirds, quarters, fifths, tenths, *etc.* As a non-limiting example, each range discussed herein can be readily broken down into a lower third, middle third and upper third, *etc.* As will also be understood by one skilled in the art all language such as "up to," "at least," and the like include the number recited and refer to ranges which can be subsequently broken down into subranges as discussed above. Finally, as will be understood by one skilled in the art, a range includes each individual member. Thus, for example, a group having 1-3 cells refers to groups having 1, 2, or 3 cells. Similarly, a group having 1-5 cells refers to groups having 1, 2, 3, 4, or 5 cells, and so forth.

[0050] From the foregoing, it will be appreciated that various embodiments of the present disclosure have been described herein for purposes of illustration, and that various modifications may be made without departing from the scope and spirit of the present disclosure. Accordingly, the various embodiments disclosed herein are not

intended to be limiting, with the true scope and spirit being indicated by the following claims.

CLAIMS

What is claimed is:

1. A cloud system comprising:
one or more computing nodes; and
one or more network switches configured to relay one or more packets among the one or more computing nodes,
wherein a respective one of the computing nodes includes:
a first processor configured to process the one or more packets to communicate with at least one of the network switches; and
a second processor configured to process the one or more packets to communicate with at least one of the other computing nodes in the cloud system.
2. The cloud system of Claim 1, wherein the respective one of the computing nodes enables the second processor to process the one or more packets when the first processor is not able to communicate the one or more packets with the at least one of the network switches.
3. The cloud system of Claim 1, wherein the second processor included in the respective one of the computing nodes is implemented by a field programmable gate array (FPGA).
4. The cloud system of Claim 1, wherein the second processor included in the respective one of the computing nodes is a network interface controller (NIC) in which a field programmable gate array (FPGA) is embedded.
5. The cloud system of Claim 1, wherein the second processors included in the respective computing nodes construct a failure recovery network.

6. The cloud system of Claim 1, wherein the second processors included in the respective computing nodes communicate with one another using an overlay network protocol between link layer and network layer.

7. The cloud system of Claim 1, wherein the network switches includes at least one of edge switches, core switches, or core routers.

8. A computing node comprising:
a first processor configured to process one or more packets to communicate with a network switch; and
a second processor configured to process the one or more packets to communicate with another computing node.

9. The computing node of Claim 8, wherein the computing node enables the second processor to process the one or more packets when the first processor is not able to communicate the one or more packets with the network switch.

10. The computing node of Claim 8, wherein the second processor is implemented by a field programmable gate array (FPGA).

11. The computing node of Claim 8, wherein the second processor includes a network interface controller (NIC) in which a field programmable gate array (FPGA) is embedded.

12. The computing node of Claim 8, wherein the first processor is further configured to be connected to the network switch via a first Ethernet socket, and the second processor is further configured to be connected to the other computing node via a second Ethernet socket.

13. A method performed under control of a computing node, comprising:

detecting a failure in communicating one or more packets with a network switch based on a first communication protocol; and

transferring the one or more packets directly to a target computing node based on a second communication protocol, in response to the detecting of the failure.

14. The method of Claim 13, further comprising:

determining, from among other computing nodes connected to the computing node, the target computing node to transfer the one or more packets.

15. The method of Claim 14, wherein the determining comprises:

calculating a distance between an IP (Internet Protocol) address of a destination computing node and an IP address of each of the other computing nodes connected to the computing node; and

selecting the target computing node as one of the other computing nodes that is closest to the destination computing node.

16. The method of Claim 15, wherein the calculating is based on integer subtraction between the IP addresses.

17. The method of Claim 13, wherein the second communication protocol is a peer-to-peer communication protocol between the computing node and the other computing nodes.

18. The method of Claim 13, wherein the second communication protocol is an overlay network protocol between link layer and network layer.

19. The method of Claim 13, wherein the second communication protocol is implemented by a field programmable gate array (FPGA) included in the computing node.

20. The method of Claim 13, further comprising:

detecting a recovery from the failure; and

transferring the one or more packets to the network switch based on the first communication protocol, in response to the detecting of the recovery.

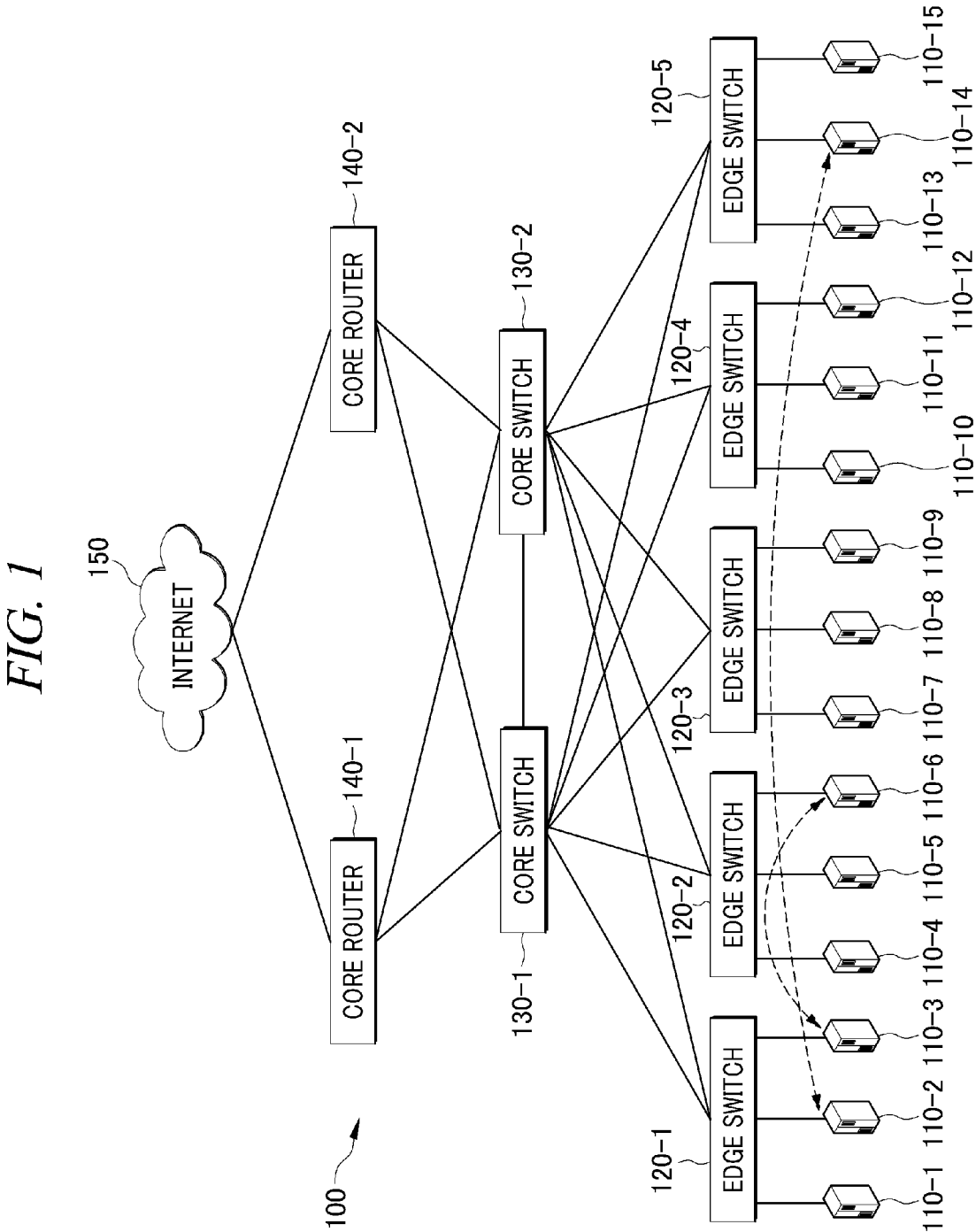


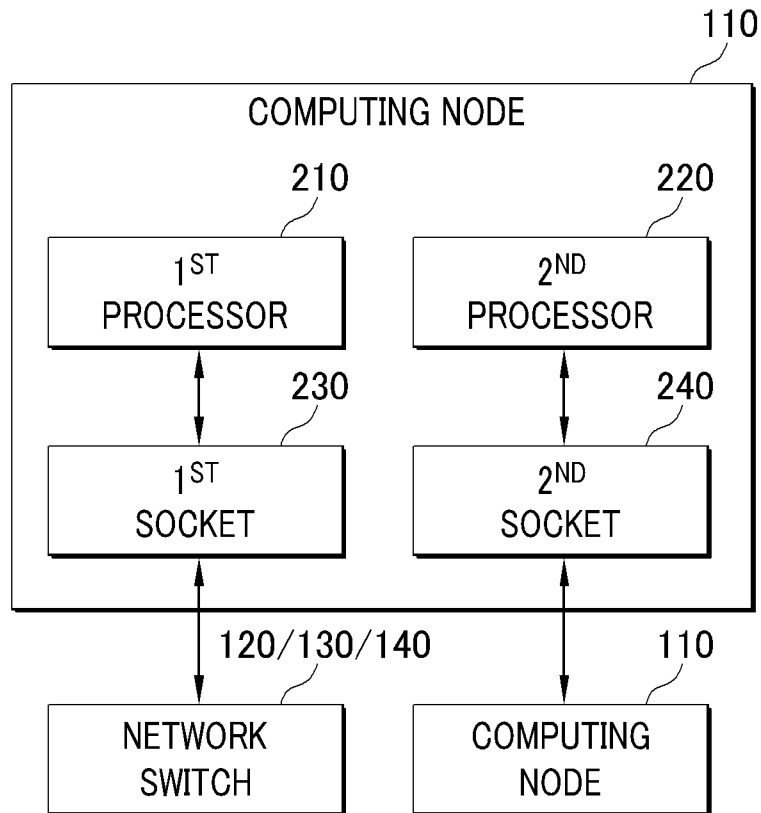
FIG. 2

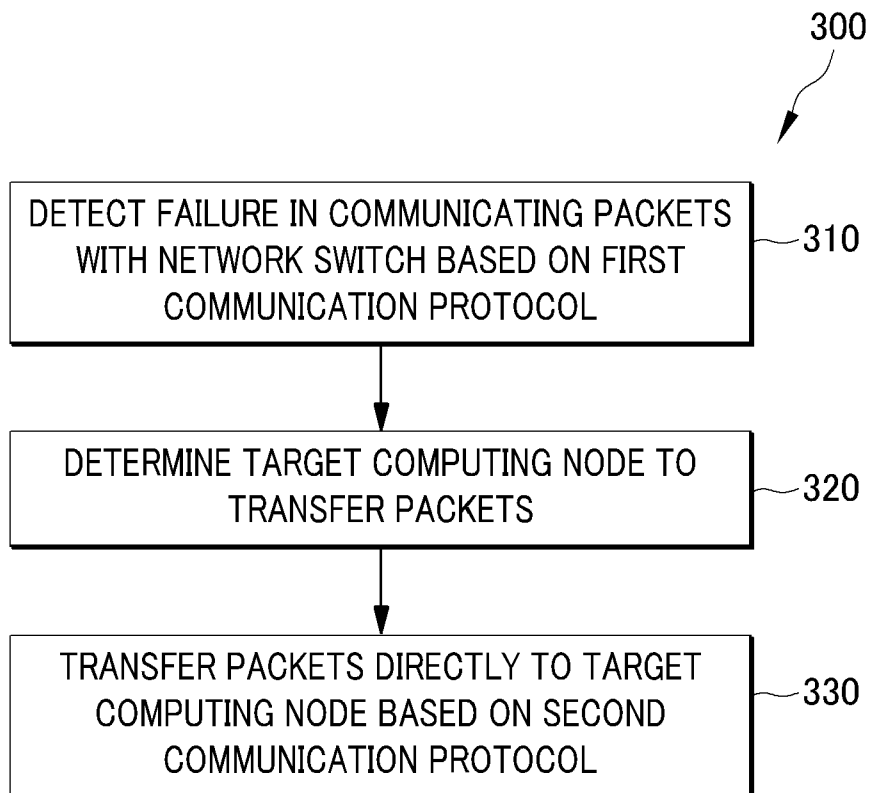
FIG. 3

FIG. 4

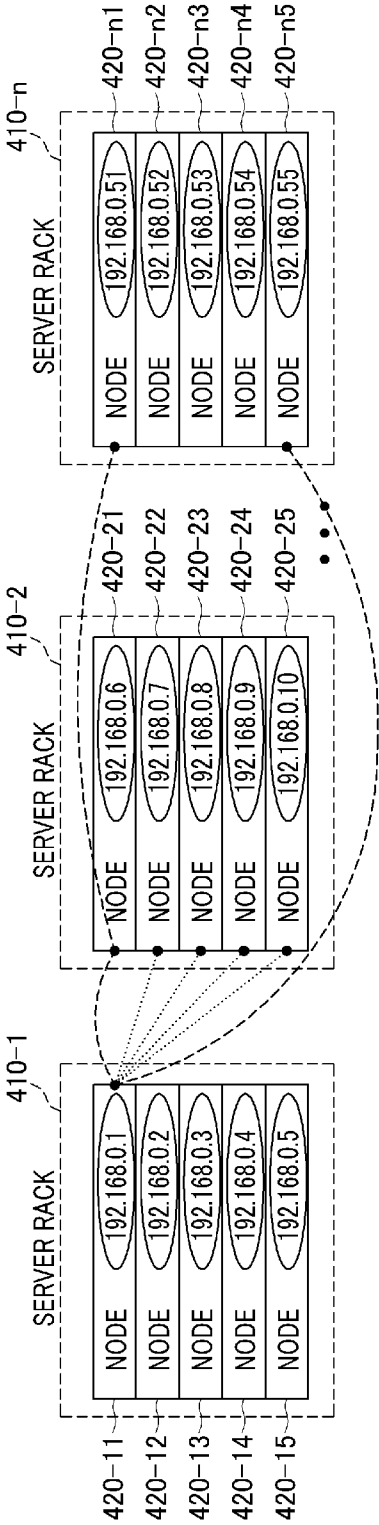
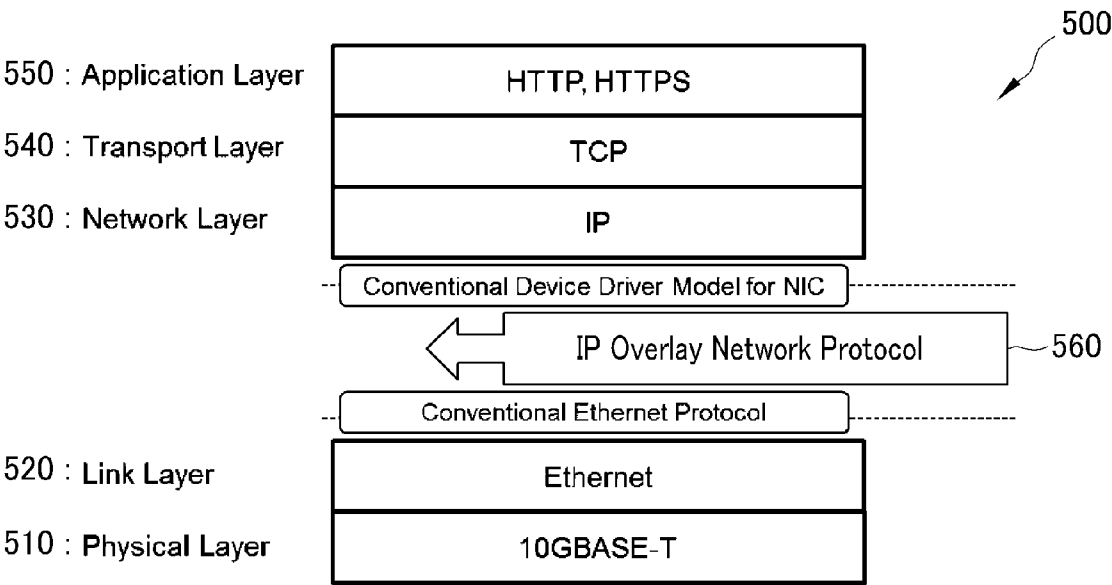


FIG. 5



INTERNATIONAL SEARCH REPORT

International application No.

PCT/US13/65564

A. CLASSIFICATION OF SUBJECT MATTER

IPC(8) - G06F 15/16 (2014.01)

USPC - 709/238

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)

IPC(8): G06F 15/16, 15/173 (2014.01)

USPC: 709/238, 203, 227

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)

MicroPatent (US-G, US-A, EP-A, EP-B, WO, JP-bib, DE-C,B, DE-A, DE-T, DE-U, GB-A, FR-A); ProQuest; IEEE/IEEEExplore; Google/Google Scholar; node, computer, device, communicate, connect, packet, data, switch, cloud, relay, intercept, multiple, plural, second, processor

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2005/0232261 A1 (WYBENGA, JC et al.) 20 October 2005; paragraphs [0040], [0050].	8
----- Y		1-7, 9-12
Y	US 2002/0112072 A1 (JAIN, S) 15 August 2002; paragraphs [0028], [0030].	1-7, 9
Y	US 2013/0170507 A1 (HSUEH, C et al.) 04 July 2013; paragraphs [0033], [0067].	3, 4, 10, 11
Y	US 2006/0013240 A1 (MA, S et al.) 19 January 2006; paragraph [0058].	12
Y	US 2008/0263255 A1 (BLINICK, SL et al.) 23 October 2008; paragraph [0076].	2, 9
Y	US 6999459 B1 (CALLON, RW et al.) 14 February 2006; column 5, lines 23-26.	5
Y	US 2010/0014533 A1 (HIRANO, J et al.) 21 January 2010; figure 10; paragraphs [0075], [0124].	6

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

* Special categories of cited documents:	"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
"A" document defining the general state of the art which is not considered to be of particular relevance	"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
"E" earlier application or patent but published on or after the international filing date	"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
"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)	"&" document member of the same patent family
"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	

Date of the actual completion of the international search

08 April 2014 (08.04.2014)

Date of mailing of the international search report

06 JUN 2014

Name and mailing address of the ISA/US

Mail Stop PCT, Attn: ISA/US, Commissioner for Patents
P.O. Box 1450, Alexandria, Virginia 22313-1450

Facsimile No. 571-273-3201

Authorized officer:

Shane Thomas

PCT Helpdesk: 571-272-4300
PCT OSP: 571-272-7774

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US13/65564

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

Group I: Claims 1-12; Group II: Claims 13-20

-Please see Supplemental Page-

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:
1-12

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US13/65564

-Continued from Box No. III - Observations where unity of invention is lacking-

This application contains the following inventions or groups of inventions which are not so linked as to form a single general inventive concept under PCT Rule 13.1. In order for all inventions to be examined, the appropriate additional examination fee must be paid.

Group I: Claims 1-12 are directed toward a cloud system and a computing node comprising two processors.

Group II: Claims 13-20 are directed toward a method performed under control of a computing node in the event of a failure.

The inventions listed as Groups I-II do not relate to a single general inventive concept under PCT Rule 13.1 because, under PCT Rule 13.2, they lack the same or corresponding special technical features for the following reasons:

The special technical features of Group I include a first processor configured to process one or more packets to communicate with at least one network switch; and a second processor configured to process the one or more packets to communicate with another computing node, which are not present in Group II.

The special technical features of Group II include detecting a failure based on a first communication protocol; and transferring one or more packets directly to a target computing node based on a second communication protocol in response to detecting the failure, which are not present in Group I.

The common technical feature shared by Groups I-II is a computing node comprising: communicating one or more packets with a network switch; and transferring the one or more packets to a computing node. However, this common feature is previously disclosed by US 2005/0232261 A1 (Wybenga). Wybenga discloses a computing node (route-processing module; paragraph [0040]) for: communicating one or more packets with a network switch (inbound data packets are received and are destined for a switch fabric; paragraphs [0040] and [0050]); and transferring the one or more packets to a computing node (outbound data packets are destined to external network interfaces of external devices; Abstract, paragraphs [0040] and [0050]).

Since the common technical feature is previously disclosed by the Wybenga reference, this common feature is not special and so Groups I-II lack unity.