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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
4 May 2017 (04.05.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/074595 A1

(51) International Patent Classification:
H04L 12/801 (2013.01)

(21) International Application Number:
PCT/US2016/051865

(22) International Filing Date:
15 September 2016 (15.09.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/246,867 27 October 2015 (27.10.2015) US
15/264,706 14 September 2016 (14.09.2016) US

(71) Applicant: **NEC LABORATORIES AMERICA, INC.** [US/US]; 4 Independence Way, Suite 200, Princeton, NJ 08540 (US).

(72) Inventors: **XU, Qiang**; 392 Windsor Commons, East Windsor, NJ 08512 (US). **LUMEZANU, Cristian**; 10 Landin Lane, Princeton Junction, NJ 08550 (US). **JIN, Cheng**; 1039 29th Avenue SE, Minneapolis, MN 55414 (US). **BAEK, Hyun-Wook**; 814 University Village, Salt Lake City, UT 84108 (US). **JIANG, Guofei**; 5 Danby Court, Princeton, NJ 08540 (US).

(74) Agent: **BITETTO, James, J.**; Tutunjian & Bitetto, P.C., 425 Broadhollow Road, Suite 302, Melville, NY 11747 (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, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

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

(54) Title: VM-TO-VM TRAFFIC ESTIMATION IN MULTI-TENANT DATA CENTERS

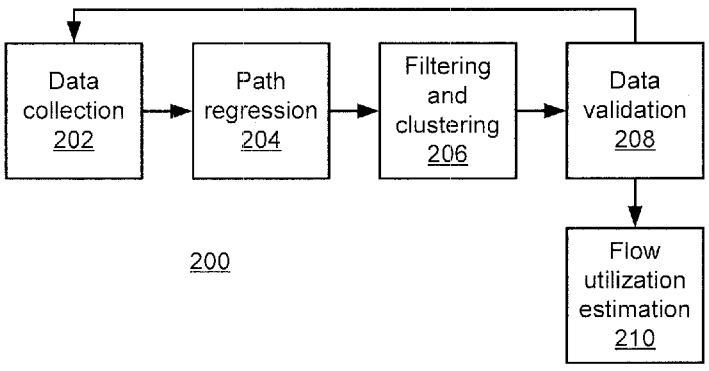

FIG. 2

(57) Abstract: Methods and systems for network management include performing (304) path regression to determine an end-to-end path across physical links for each data flow in a network. A per-flow utilization of each physical link in the network is estimated (314) based on the determined end-to-end paths. A management action is performed (316) in the network based on the estimated per-flow utilization.

WO 2017/074595 A1

VM-TO-VM TRAFFIC ESTIMATION IN MULTI-TENANT DATA CENTERS

RELATED APPLICATION INFORMATION

[0001] This application claims priority to U.S. Provisional Application No. 62/246,867 filed on October 27, 2015, incorporated herein by reference in its entirety.

BACKGROUND

Technical Field

[0002] The present invention relates to data center management and, more particularly, to estimating traffic along virtual flows in a multi-tenant data center.

Description of the Related Art

[0003] Modern data centers may serve multiple clients with a single device using, for example, virtual machines to host the multiple workloads as if they were on separate hardware. In addition, software defined networking provides network administrators with the ability to dynamically reorganize network resources, creating and reorganizing virtual local area networks (VLANs) and virtual extensible local area networks (VXLANs) on the fly.

[0004] However, when multiple different workflows may be transmitted along a single shared physical link, it can be difficult to obtain flow-level information. In a multi-tenant data center, where a single device with a single physical link can host multiple virtual machines and multiple virtual links, it is difficult to break down the utilization of that physical link into the usage of the multiple virtual links.

[0005] Existing attempts to find flow-level information focus on three points: improving the collection of performance counters from distributed network devices (e.g., switches and middleboxes) in data centers, scalable data collection systems, and

instrumenting virtualization platforms. For example, some solutions make use of additional instrumentation at the hypervisor level to perform packet-level inspection. Such inspection offers accuracy in determining information about the flows on a link, but generally involves significant increases in the cost and complexity of network monitoring.

SUMMARY

[0006] A method for network management include performing path regression to determine an end-to-end path across physical links for each data flow in a network. A per-flow utilization of each physical link in the network is estimated based on the determined end-to-end paths. A management action is performed in the network based on the estimated per-flow utilization.

[0007] A method for network management includes performing path regression to determine an end-to-end path across physical links for each data flow in a network by determining a portion of an end-to-end path based on forwarding table information and determining a remainder of the end-to-end path based on one or more inference rules. A number of flows is reduced by filtering the flows according to one or more rules and clustering the filtered flows into flow groups. Per-flow utilization of each physical link in the network is estimated based on the determined end-to-end paths. A management action is performed in the network based on the estimated per-flow utilization.

[0008] A system for network management includes a path regression module having a processor configured to determine an end-to-end path across physical links for each data flow in a network. A flow utilization estimation module is configured to estimate per-flow utilization of each physical link in the network based on the determined end-

to-end paths. A network management module is configured to perform a management action in the network based on the estimated per-flow utilization.

[0009] These and other features and advantages will become apparent from the following detailed description of illustrative embodiments thereof, which is to be read in connection with the accompanying drawings.

BRIEF DESCRIPTION OF DRAWINGS

[0010] The disclosure will provide details in the following description of preferred embodiments with reference to the following figures wherein:

[0011] FIG. 1 is a block diagram of a multi-tenant data center network in accordance with the present principles;

[0012] FIG. 2 is a block/flow diagram of a method/system for estimating flow utilization in accordance with the present principles;

[0013] FIG. 3 is a block/flow diagram of a method for estimating flow utilization in accordance with the present principles;

[0014] FIG. 4 is a block diagram of a system for estimating flow utilization in accordance with the present principles; and

[0015] FIG. 5 is a block diagram of a processing system in accordance with the present principles.

DETAILED DESCRIPTION OF PREFERRED EMBODIMENTS

[0016] Embodiments of the present principles provide flow-level traffic analysis in multi-tenant data centers based on existing performance counters and readily available network forwarding table information. The present embodiments employ path regression to recreate a flow map when forwarding table information has expired.

Filtering and clustering of flows is used to reduce the number of variables in play, making a matrix representation of which flows are active on which links. The reconstructed flow information is then used to generate a set of expected link performance measurements which can be compared to actual link performance measurements to determine a degree of error in the flow traffic output.

[0017] Referring now to FIG. 1, a generalized diagram of a data center network 100 is shown. Multiple virtual machines (VMs) 102 may be implemented on a single physical server 104. The server 104 has a link to a switch 106 which may be connected to other servers 104, other switches 106, and a router 108. This represents the physical network with physical links, but the communications may additionally have a logical layer on top of the physical network layer that forms virtual networks (e.g., virtual local area networks (VLANs) and virtual extensible local area networks (VXLANs)).

[0018] From the perspective of a single VM 102 in the data center network 100, other VMs 102 may appear to be local or may be in separate virtual network segments, with communication mediated by VLAN or VXLAN gateways (which may be implemented as distinct servers 102 within the network and which need not be situated within the most direct physical routing path between two VMs 102). As such, the path a given flow takes through the physical network 100 can be non-trivial, with traffic from a given VM 102 visiting multiple locations before reaching its destination.

[0019] These arrangements are particularly useful for cloud computing environments, where many clients may be served by a single server 104. The virtualization described above allows VMs 102 to act in a simple, easy-to-provision manner without any awareness of the actual network structure. However, when multiple flows exist on a single physical link (and indeed, when those flows sometimes double

back on a given link), it is challenging to identify, for example, what flows are the cause of a local network hotspot.

[0020] The devices in the network 100 can maintain certain information regarding the network's traffic. For example, the simple network management protocol (SNMP) may be implemented in switches 106 and routers 108 to track, for example, forwarding tables, performance counters for physical links, VLAN and VXLAN configurations, etc. Additionally, edge devices may keep information such as VM locations (i.e., the physical server 104 on which a given VM 102 resides), VLAN and VXLAN allocations to clients, etc. It should be noted that this information is collected as a matter of standard practice and that no additional instrumentation need be introduced to determine these points.

[0021] Some of the information, however, may be unreliable. For example, in the case of forwarding tables, the tables may suffer from the expiration of entries. As a result, this information provides only partial visibility for a given source/destination pair.

[0022] Referring now to FIG. 2, a system 200 for flow-level traffic analysis is shown. Block 202 collects data within a network 100. As noted above, a significant amount of data may be tracked as a matter of standard operation within the data center network 100, and this information may be made available in block 202. It should be noted that this data collection may be performed with zero instrumentation or any other kind of alteration to the existing data collection mechanisms within a data structure. While the log format may vary from one virtualization platform to the next, the availability of the data is generally present.

[0023] As noted above, some of the data from the forwarding tables may be incomplete. Block 204 uses the available data to perform path regression to determine

the end-to-end paths of VM pairwise flows. Domain knowledge may be employed to help construct a coarse-grained path. For example, it may be known that the end-to-end path includes intermediate hops such as host servers 104, VXLAN gateways, and switches 106. The forwarding tables collected by block 202 are used to determine the exact intermediate hops—e.g., which VXLAN in particular serves the communication, followed by filling in the hops between known intermediate hops. After all hops in a flow path are filled in, a complete end-to-end path for the given source and destination pair is computed. Thus path regression block 204 generates a many-to-many mapping from flows to physical links (e.g., switch ports, service ports, VM ports, etc.) that may be represented in the form of a matrix.

[0024] To address the problem of forwarding table expiration, rules may be used to infer forwarding behavior. One such rule notes that MAC addresses in a same VLAN are likely to share a same path. In addition, forwarding possibilities are narrowed based on network graph properties and the search space for forwarding is reduced based on the network configuration. Another rule is that, if a particular MAC address fails to appear in a large number of forwarding table snapshots, it implies that the flows originating from that MAC address are negligible.

[0025] In addition to the particular difficulty of accessing complete forwarding table information, performance counters acquired by data collection block 202 may be introduced with noise, time misalignments, conflicts, etc. These cause inaccuracies in the flow traffic analysis but, as described below, feedback based on an error analysis is generated to compensate for the inaccuracies.

[0026] For each physical link, which may include switch ports, server ports, middlebox ports, etc., the total utilization of flows carried by the link can be expressed as an equation to express the performance counter of that interface. It should be

recognized that this representation is likely to include some inaccuracies if some forwarding table information was unavailable. The flows can be represented as a matrix as shown in Table 1 below:

	Flow 1	Flow 2	Flow 3	...	Flow N
Link 1	0	0	0	...	0
Link 2	1	0	1	...	1
Link 3	0	1	0	...	1
...	...	...	...	...	...
Link M	0	1	1	...	1

Table 1

[0027] In this representation, each row may be considered an equation, with the values in each column representing values for the variables of the equation. Looking at the rows another way, each row identifies which flows are traveling on the respective link. The columns, meanwhile, represent the links that a given flow traverses. However, this system will be difficult to solve in this form, because the number of variables dwarfs the number of equations. As such, block 206 performs filtering and clustering to reduce the number of variables by eliminating flows that are unlikely to be active. Exemplary filters that may be used include:

[0028] 1. Link counter based filters—if the traffic counter of a link is below a threshold, all the virtual links traversing through the link can be expected to be inactive;

[0029] 2. Temporal domain-based filters—if the history of a flow follows certain inactive patterns, the flow can be expected to be inactive in the short term; and

[0030] 3. Spatial domain-based filters—some flows' activity may be dependent on other flows and thus, if certain flows are known to be inactive, the other flows depending on them will also be inactive.

[0031] It should be recognized that these filters are included for the purpose of description only and should not be construed as limiting. After filtering, the number of flows being considered should approximately match the number of links in the data center network 100 to make the system of equations solvable. If filtering does not reduce the number of flows sufficiently, the remaining flows may be clustered to further reduce their effective number. Clustering may be performed if, for example, two flows have the same vertical vector in the flow matrix.

[0032] The solution of the system of equations can be expressed as $\mathbf{P} \times \mathbf{F} = \mathbf{L}$, where **P** is the matrix as described above, **F** is a column vector of flow performance, and **L** is a column vector of respective link counters collected from block 202 using, e.g., SNMP data. The flow performance is unknown, but the inferred **P** matrix and the known **L** vector permit the calculation of an inferred flow performance vector **F'**. One exemplary measurement of flow performance may be the size of a flow in terms of a number of bytes.

[0033] Block 208 can then provide data verification based on the inferred flow performance vector **F'**. The **P** matrix and the **F'** vector are used to calculate an expected link counter vector **L'**. The difference between the expected **L'** and the known **L** reflects the quality of the data set. If the performance counters are significantly affected by, e.g., noise or time alignment issues, the error vector **L' - L** will be significant. A classification of errors is possible as follows:

[0034] 1. If errors have locality behavior (i.e., if physical lengths close to one another have high errors), there may be some data collection misconfiguration;

[0035] 2. If the error is close to the averaged error, then the link does not significantly affect data quality; and

[0036] 3. If the error of a link shows some up-and-down behavior over time, it is likely due to a time alignment issue.

[0037] Depending on the error type, data validation block 208 feeds information back to data collection block 202 to improve future data collection.

[0038] After the matrix **P** and inferred flow performance vector **F'** have been calculated, block 210 uses this information to estimate flow utilization.

[0039] Referring now to FIG. 3, a method for flow-level traffic analysis is shown. Block 302 collects data from network devices (e.g., servers 104, switches 106, and routers 108) relating to link and flow performance. This information can be collected using existing SNMP and software defined networking protocols without introducing any new instrumentation in the network device. Block 304 performs path regression based on the collected information to determine an end-to-end path through the network 100 for each flow. As discussed above, the information collected by block 302 may not be sufficient to completely determine complete paths for every flow. As such, block 302 may perform inferences based on rules and other information to fill any gaps in the flow paths.

[0040] Because the number of flows may be much greater than the number of physical links in the network 100, block 306 may filter the flows based on criteria such as those discussed above or any other appropriate rules determined by statistics or domain knowledge. In particular, block 306 filters out flows that can be predicted to have very little activity. If the filtering of block 306 is insufficient to reduce the number of flows being considered to or below the number of physical links, block 308 performs clustering to group flows together to further reduce the number of flow groups being considered.

[0041] Block 310 uses the measured link counters from block 302 and the end-to-end flow paths from block 304 to generate an inferred flow performance vector **F'** as described above. The inferred flow performance vector **F'** is then used with the flow path matrix **P** to generate an expected link counter vector **L'**, which block 310 compares to the measured link counter information to determine an error vector. Based on the error vector, block 312 adjusts the flow path matrix **P** to account for the errors.

[0042] Block 314 estimates the flow utilization using the inferred flow performance vector **F'**. This information reflects which flows use each physical link in the network 100 and their respective contributions to the traffic on those links. Based on the flow utilization estimate, block 316 performs a network management function. For example, in the case of high utilization on a particular link that causes a bottleneck, block 316 may trigger rerouting one or more flows from the overloaded physical link to another path in the network 100. Other network management actions may include prioritizing traffic from particular flows, restricting bandwidth available to particular flows, changing the network topology, changing settings at servers 104, switches 106, or routers 108, or any other network management action within the ability of those having ordinary skill in the art.

[0043] Embodiments described herein may be entirely hardware, entirely software or including both hardware and software elements. In a preferred embodiment, the present invention is implemented in software, which includes but is not limited to firmware, resident software, microcode, etc.

[0044] Embodiments may include a computer program product accessible from a computer-usable or computer-readable medium providing program code for use by or in connection with a computer or any instruction execution system. A computer-usable or computer readable medium may include any apparatus that stores, communicates,

propagates, or transports the program for use by or in connection with the instruction execution system, apparatus, or device. The medium can be magnetic, optical, electronic, electromagnetic, infrared, or semiconductor system (or apparatus or device) or a propagation medium. The medium may include a computer-readable storage medium such as a semiconductor or solid state memory, magnetic tape, a removable computer diskette, a random access memory (RAM), a read-only memory (ROM), a rigid magnetic disk and an optical disk, etc.

[0045] Each computer program may be tangibly stored in a machine-readable storage media or device (e.g., program memory or magnetic disk) readable by a general or special purpose programmable computer, for configuring and controlling operation of a computer when the storage media or device is read by the computer to perform the procedures described herein. The inventive system may also be considered to be embodied in a computer-readable storage medium, configured with a computer program, where the storage medium so configured causes a computer to operate in a specific and predefined manner to perform the functions described herein.

[0046] A data processing system suitable for storing and/or executing program code may include at least one processor coupled directly or indirectly to memory elements through a system bus. The memory elements can include local memory employed during actual execution of the program code, bulk storage, and cache memories which provide temporary storage of at least some program code to reduce the number of times code is retrieved from bulk storage during execution. Input/output or I/O devices (including but not limited to keyboards, displays, pointing devices, etc.) may be coupled to the system either directly or through intervening I/O controllers.

[0047] Network adapters may also be coupled to the system to enable the data processing system to become coupled to other data processing systems or remote

printers or storage devices through intervening private or public networks. Modems, cable modem and Ethernet cards are just a few of the currently available types of network adapters.

[0048] Referring now to FIG. 4, a network management system 400 is shown. The system 400 includes a hardware processor 402 and a memory 404. The system 400 may further include one or more physical network interfaces 405 that the system 400 uses to communicate with other devices on a network 100. The system also includes one or more functional modules. In one embodiment, the functional modules may be implemented as software that is stored in memory 404 and is executed by processor 402. In an alternative embodiment, some or all of the functional modules may be implemented as one or more discrete hardware components in the form of, e.g., application specific integrated chips or field programmable gate arrays.

[0049] A data collection module 406 accesses the other devices on the network 100 to acquire available information regarding the devices and physical links in the network 100. A path regression module 408 performs path regression using the collected data to fill out end-to-end paths for each flow and data validation module 410 performs data validation to make corrections to the path regression. Flow utilization estimation module 412 determines how the various flows use the physical links in the network 100. If there are too many flows to perform the flow utilization estimation, filtering/clustering module 409 filters out and clusters flows as needed to reduce that number. Network managing module 414 uses the flow utilization estimates to perform network management operations within the network 100 to improve the performance of the network.

[0050] Referring now to FIG. 5, an exemplary processing system 500 is shown which may represent the network management system 400. The processing system

500 includes at least one processor (CPU) 504 operatively coupled to other components via a system bus 502. A cache 506, a Read Only Memory (ROM) 508, a Random Access Memory (RAM) 510, an input/output (I/O) adapter 520, a sound adapter 530, a network adapter 540, a user interface adapter 550, and a display adapter 560, are operatively coupled to the system bus 502.

[0051] A first storage device 522 and a second storage device 524 are operatively coupled to system bus 502 by the I/O adapter 520. The storage devices 522 and 524 can be any of a disk storage device (e.g., a magnetic or optical disk storage device), a solid state magnetic device, and so forth. The storage devices 522 and 524 can be the same type of storage device or different types of storage devices.

[0052] A speaker 532 is operatively coupled to system bus 502 by the sound adapter 530. A transceiver 542 is operatively coupled to system bus 502 by network adapter 540. A display device 562 is operatively coupled to system bus 502 by display adapter 560.

[0053] A first user input device 552, a second user input device 554, and a third user input device 556 are operatively coupled to system bus 502 by user interface adapter 550. The user input devices 552, 554, and 556 can be any of a keyboard, a mouse, a keypad, an image capture device, a motion sensing device, a microphone, a device incorporating the functionality of at least two of the preceding devices, and so forth. Of course, other types of input devices can also be used, while maintaining the spirit of the present principles. The user input devices 552, 554, and 556 can be the same type of user input device or different types of user input devices. The user input devices 552, 554, and 556 are used to input and output information to and from system 500.

[0054] Of course, the processing system 500 may also include other elements (not shown), as readily contemplated by one of skill in the art, as well as omit certain elements. For example, various other input devices and/or output devices can be included in processing system 500, depending upon the particular implementation of the same, as readily understood by one of ordinary skill in the art. For example, various types of wireless and/or wired input and/or output devices can be used. Moreover, additional processors, controllers, memories, and so forth, in various configurations can also be utilized as readily appreciated by one of ordinary skill in the art. These and other variations of the processing system 500 are readily contemplated by one of ordinary skill in the art given the teachings of the present principles provided herein.

[0055] The foregoing is to be understood as being in every respect illustrative and exemplary, but not restrictive, and the scope of the invention disclosed herein is not to be determined from the Detailed Description, but rather from the claims as interpreted according to the full breadth permitted by the patent laws. It is to be understood that the embodiments shown and described herein are only illustrative of the principles of the present invention and that those skilled in the art may implement various modifications without departing from the scope and spirit of the invention. Those skilled in the art could implement various other feature combinations without departing from the scope and spirit of the invention. Having thus described aspects of the invention, with the details and particularity required by the patent laws, what is claimed and desired protected by Letters Patent is set forth in the appended claims.

WHAT IS CLAIMED IS:

1. A method for network management, comprising:

performing path regression (304) to determine an end-to-end path across physical links for each data flow in a network;

estimating per-flow utilization (314) of each physical link in the network based on the determined end-to-end paths; and

performing a management action (316) in the network based on the estimated per-flow utilization.

2. The method of claim 1, wherein performing path regression comprises determining a portion of an end-to-end path based on forwarding table information and determining a remainder of the end-to-end path based on one or more inference rules.

3. The method of claim 1, further comprising:

calculating expected link counts based on the determined end-to-end paths;

comparing the expected link counts to measured link counts to generate an error measurement; and

adjusting the determined end-to-end paths based on the error measurement.

4. The method of claim 3, wherein calculating the expected link counts comprises:

representing the end-to-end paths as a path matrix;

representing the measured link counts as a measured link count vector;

determining an inferred flow performance vector based on the path matrix and the measured link count vector; and

determining an expected link count vector based on the path matrix and the inferred flow performance vector.

5. The method of claim 1, further comprising reducing a number of flows by filtering the flows according to one or more rules.

6. The method of claim 1, further comprising clustering the flows into flow groups.

7. The method of claim 1, wherein performing the network management function comprises one or more of rerouting a flow, changing a topology of the network, and changing a setting of a device on the network.

8. A method for network management, comprising:

performing path regression (304) to determine an end-to-end path across physical links for each data flow in a network by determining a portion of an end-to-end path based on forwarding table information and determining a remainder of the end-to-end path based on one or more inference rules;

reducing a number of flows by filtering (306) the flows according to one or more rules and clustering the filtered flows (308) into flow groups;

estimating per-flow utilization (314) of each physical link in the network based on the determined end-to-end paths; and

performing a management action (316) in the network based on the estimated per-flow utilization.

9. The method of claim 8, further comprising:

calculating expected link counts based on the determined end-to-end paths;

comparing the expected link counts to measured link counts to generate an error measurement; and

adjusting the determined end-to-end paths based on the error measurement.

10. The method of claim 9, wherein calculating the expected link counts comprises:

representing the end-to-end paths as a path matrix;

representing the measured link counts as a measured link count vector;

determining an inferred flow performance vector based on the path matrix and the measured link count vector; and

determining an expected link count vector based on the path matrix and the inferred flow performance vector.

11. The method of claim 8, wherein performing the network management function comprises one or more of rerouting a flow, changing a topology of the network, and changing a setting of a device on the network.

12. A system for network management, comprising:

a path regression module (408) comprising a processor configured to determine an end-to-end path across physical links for each data flow in a network;

a flow utilization estimation module (412) configured to estimate per-flow utilization of each physical link in the network based on the determined end-to-end paths; and

a network management module (414) configured to perform a management action in the network based on the estimated per-flow utilization.

13. The system of claim 12, wherein the path regression module is further configured to determine a portion of an end-to-end path based on forwarding table information and to determine a remainder of the end-to-end path based on one or more inference rules.

14. The system of claim 12, further comprising a data validation module configured to calculate expected link counts based on the determined end-to-end paths, to compare the expected link counts to measured link counts to generate an error measurement, and to adjust the determined end-to-end paths based on the error measurement.

15. The system of claim 14, wherein the data validation module is further configured to represent the end-to-end paths as a path matrix, to represent the measured link counts as a measured link count vector, to determine an inferred flow performance vector based on the path matrix and the measured link count vector, and to determine an expected link count vector based on the path matrix and the inferred flow performance vector.

16. The system of claim 12, further comprising a filtering module configured to reduce a number of flows by filtering the flows according to one or more rules.

17. The system of claim 12, further comprising a clustering module configured to cluster the flows into flow groups.

18. The system of claim 12, wherein the network management function comprises one or more of rerouting a flow, changing a topology of the network, and changing a setting of a device on the network.

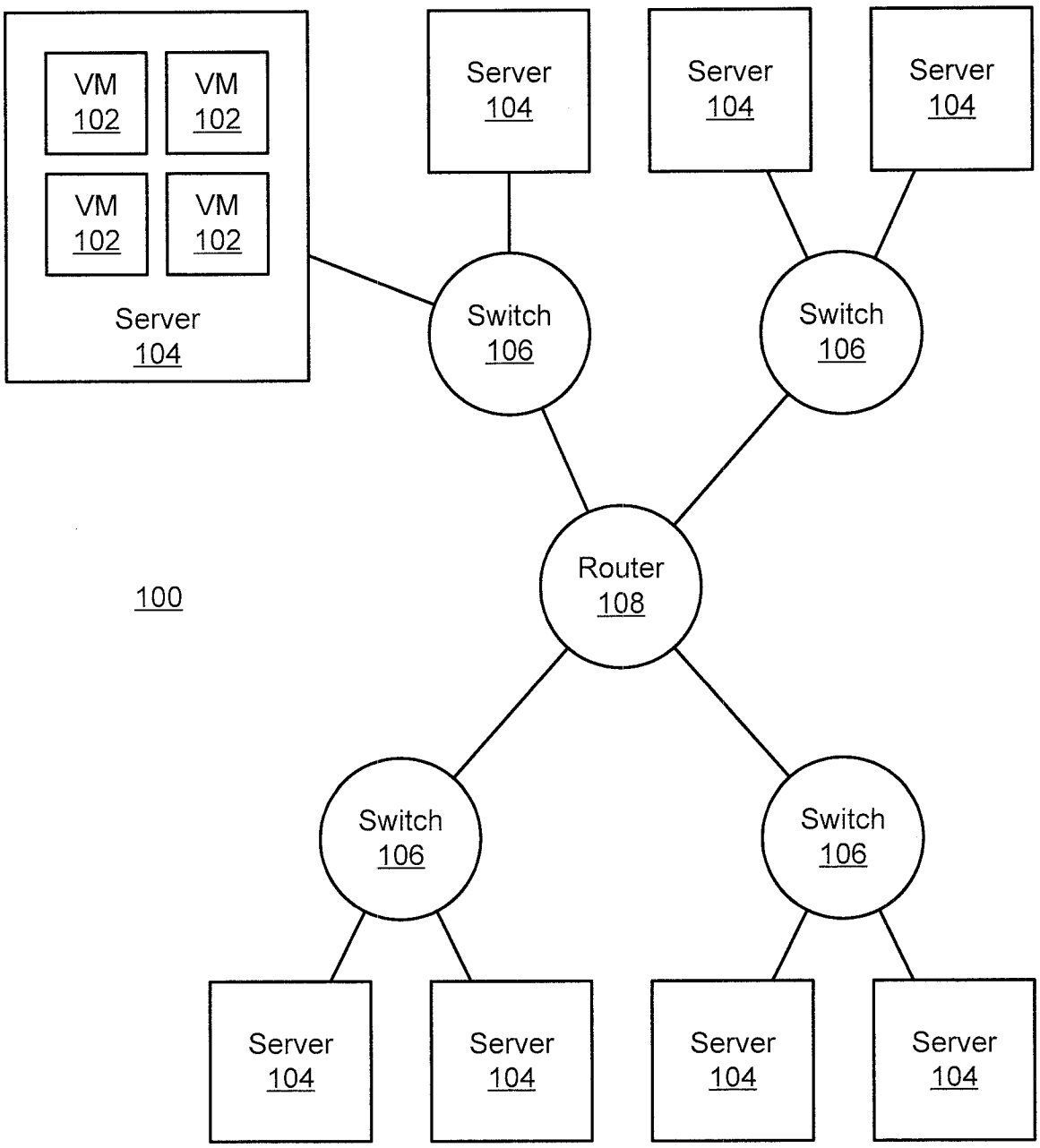
VM
102
VM
102
VM
102
VM
102
Server
104
Server
104
Server
104
Server
104
Server
104
Switch
106
Switch
106
Router
108
100
Switch
106
Switch
106
Server
104
Server
104
Server
104
Server
104

FIG. 1

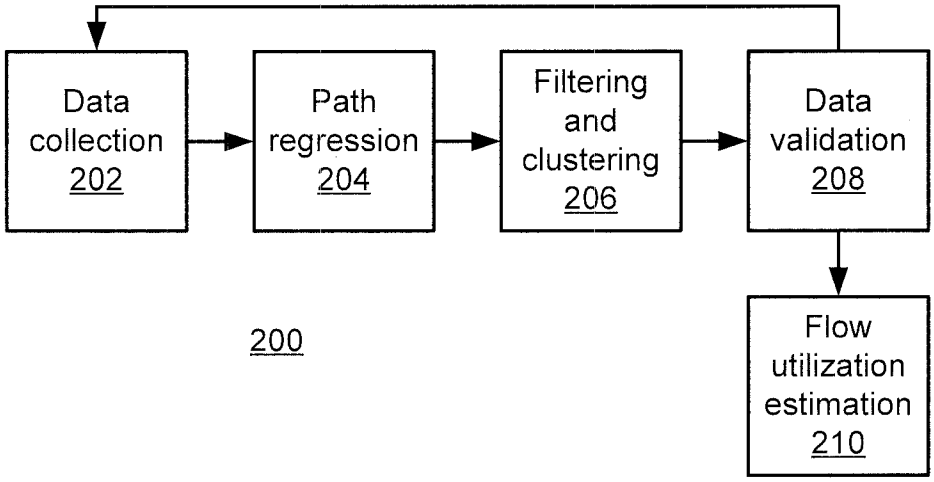
Data collection 202
Path regression 204
Filtering and clustering 206
Data validation 208
Flow utilization estimation 210
200

FIG. 2

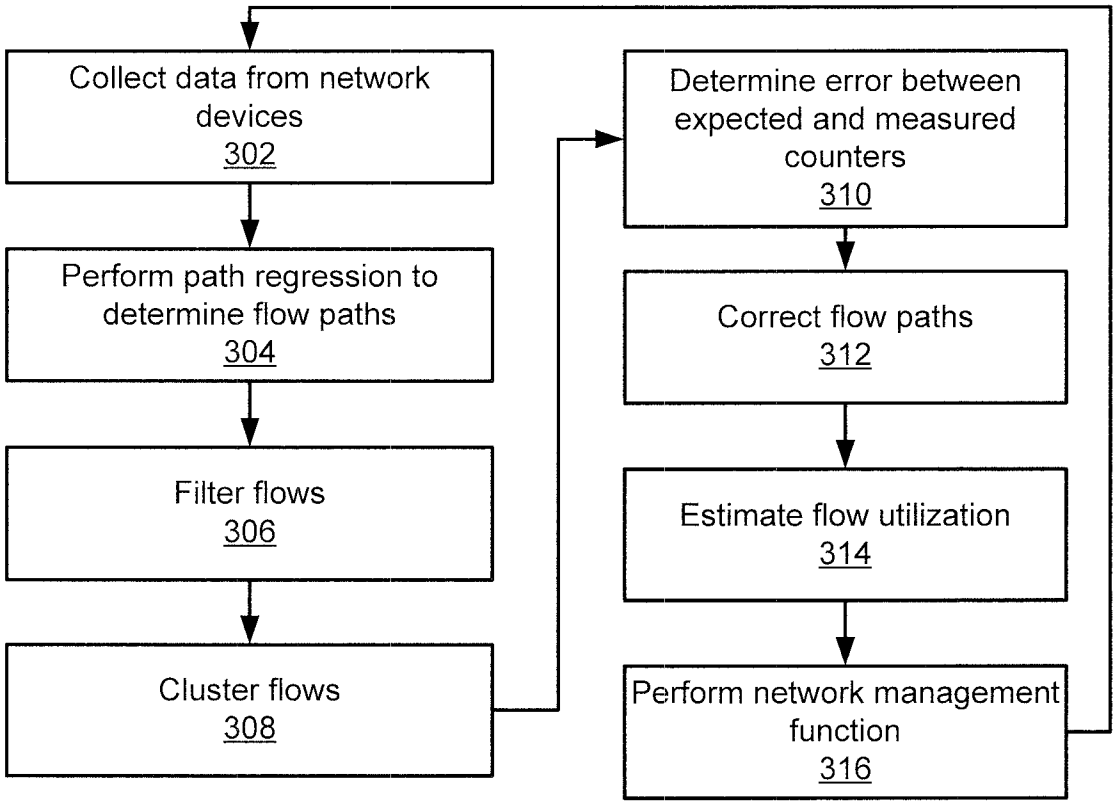
Collect data from network devices 302
Perform path regression to determine flow paths 304
Filter flows 306
Cluster flows 308
Determine error between expected and measured counters 310
Correct flow paths 312
Estimate flow utilization 314
Perform network management function 316

FIG. 3

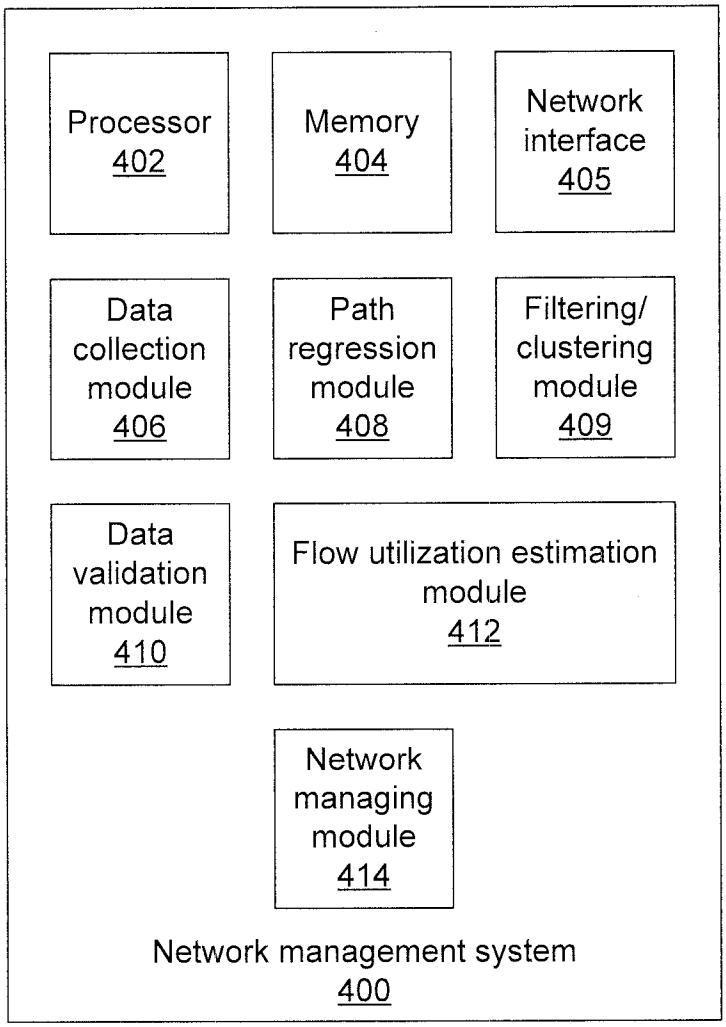
Processor
402
Memory
404
Network interface
405
Data collection module
406
Path regression module
408
Filtering/ clustering module
409
Data validation module
410
Flow utilization estimation module
412
Network managing module
414
Network management system
400

FIG. 4

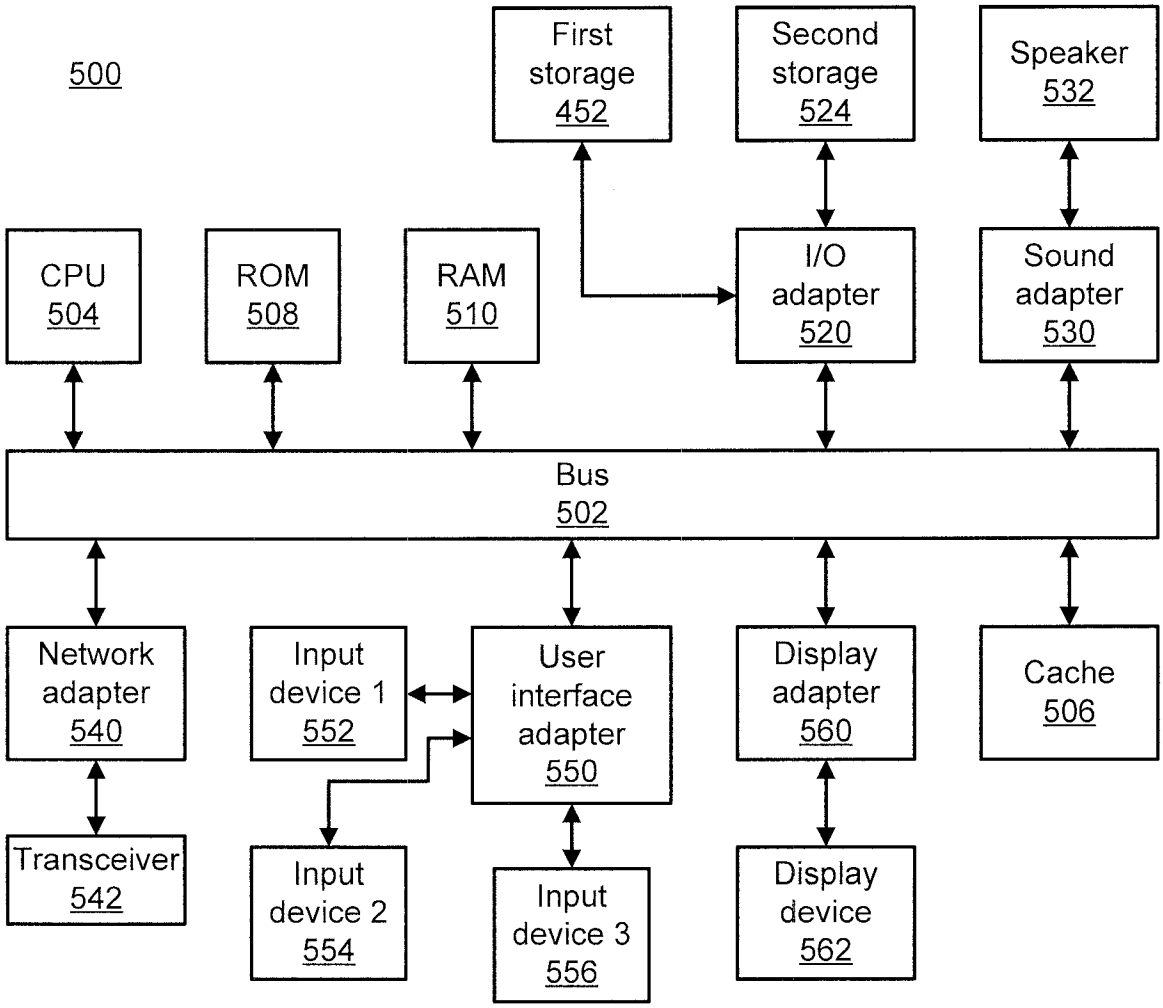
500
First storage 452
Second storage 524
Speaker 532
CPU 504
ROM 508
RAM 510
I/O adapter 520
Sound adapter 530
Bus 502
Network adapter 540
Input device 1 552
User interface adapter 550
Display adapter 560
Cache 506
Transceiver 542
Input device 2 554
Input device 3 556
Display device 562

FIG. 5

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2016/051865

A. CLASSIFICATION OF SUBJECT MATTER

H04L 12/801(2013.01)i

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
H04L 12/801; H04L 12/24; H04L 29/08; H04J 3/16; H04L 29/06; H04L 12/28; H04W 72/00; G01R 31/08

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: per-flow, utilization, estimation, evaluate, path, performance, link, count

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 8351331 B2 (THOMAS KARAGIANNIS et al.) 08 January 2013 See column 1, line 55 - column 2, line 2; column 2, line 62 - column 3, line 17; column 5, lines 28-44; figure 2; and claim 1.	1-2,5-8,11-13 ,16-18
A		3-4,9-10,14-15
Y	US 2014-0096256 A1 (UNIVERSITY OF WASHINGTON THROUGH ITS CENTER FOR CO.) 03 April 2014 See paragraph [0010].	1-2,5-8,11-13 ,16-18
A	US 8125907 B2 (JOHN EARNEST AVERI et al.) 28 February 2012 See column 3, line 46 - column 12, line 31; and figures 1A-2.	1-18
A	US 9112809 B2 (TELEFONAKTIEBOLAGET LM ERICSSON (PUBL)) 18 August 2015 See column 4, line 21 - column 13, line 2; and figures 1-5.	1-18
A	US 7453804 B1 (AZEEM FEROZ et al.) 18 November 2008 See column 17, line 64 - column 22, line 2; and figures 3-4D.	1-18

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

☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents,such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 December 2016 (14.12.2016)	**21 December 2016 (21.12.2016)**

Name and mailing address of the ISA/KR
International Application Division
Korean Intellectual Property Office
189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea
Facsimile No. +82-42-481-8578

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

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/US2016/051865

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 8351331 B2	08/01/2013	US 2011-310735 A1	22/12/2011
US 2014-0096256 A1	03/04/2014	None	
US 8125907 B2	28/02/2012	US 08274891 B2	25/09/2012
		US 08644164 B2	04/02/2014
		US 09100338 B2	04/08/2015
		US 2009-0310485 A1	17/12/2009
		US 2012-117273 A1	10/05/2012
		US 2012-314578 A1	13/12/2012
		US 2014-185445 A1	03/07/2014
		US 2016-006658 A1	07/01/2016
US 9112809 B2	18/08/2015	CN 104798356 A	22/07/2015
		EP 2923479 A1	30/09/2015
		MX 2015006471 A	14/08/2015
		US 2014-0143300 A1	22/05/2014
		WO 2014-081370 A1	30/05/2014
US 7453804 B1	18/11/2008	None	