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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
6 March 2003 (06.03.2003)

PCT

(10) International Publication Number
WO 03/019870 A2

(51) International Patent Classification[7]: H04L 12/28, 12/50

(21) International Application Number: PCT/US02/26905

(22) International Filing Date: 23 August 2002 (23.08.2002)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/314,692 24 August 2001 (24.08.2001) US

(71) Applicant: **PERIBIT NETWORKS, INC.** [US/US]; 2855 Bowers Drive, Santa Clara, CA 95051 (US).

(72) Inventors: **BHARALI, Anupam, A.**; 5210 Silver Ridge Court, San Jose, CA 95138 (US). **SINGH, Balraj**; 885 Highlands Circle, Los Altos, CA 94024 (US). **SAMPAT, Manish, H.**; Apt.J-108, 2000 Walnut Avenue, Fremont, CA 94538 (US). **SINGH, Amit, P.**; 1044 Renoir Court, Sunnyvale, CA 94087 (US). **BATRA, Rajiv**; 28020 Audrey Smith Lane, Saratoga, CA 95070 (US).

(74) Agents: **GLORE, James, E.** et al.; Fenwick & West LLP, Two Palo Alto Square, Palo Alto, CA 94306 (US).

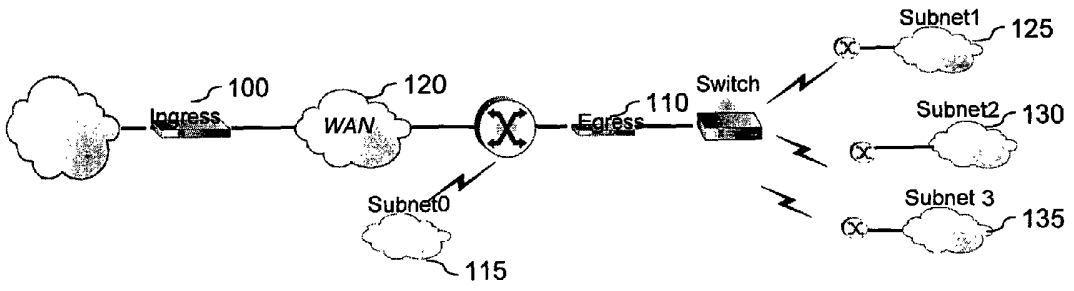
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, OM, PH, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TN, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZM, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, SK, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: EFFICIENT METHOD AND SYSTEM FOR AUTOMATIC DISCOVERY AND VERIFICATION OF OPTIMAL PATHS THROUGH A DYNAMIC MULTI-POINT MESHED OVERLAY NETWORK



(57) Abstract: The present invention provides an efficient system and method for routing information through a dynamic network. The system includes at least one ingress point and one egress point. The ingress and egress point cooperate to form a virtual circuit for routing packets to destination subnets directly reachable by the egress point. The egress point automatically discovers which subnets are directly accessible via its local ports and summarizes this information for the ingress point. The ingress point receives this information, compiles it into a routing table, and verifies that those subnets are best accessed by the egress point. Verification is accomplished by sending probe packets to select addresses on the subnet. Additionally, the egress point may continue to monitor the local topology and incrementally update the information to the ingress to allow the ingress to adjust its compiled routing table.

WO 03/019870 A2

EFFICIENT METHOD AND SYSTEM FOR AUTOMATIC DISCOVERY AND VERIFICATION OF OPTIMAL PATHS THROUGH A DYNAMIC MULTI-POINT MESHED OVERLAY NETWORK

RELATED APPLICATIONS

[0001] This application claims priority under 35 U.S.C. § 119(e) to U.S. Provisional Application serial number No. 60/314,692, entitled "Efficient Method and System for Automatic Discovery and Verification of Optimal Paths through a Dynamic Multi-point Meshed Overlay Network", filed on August 24, 2001, the subject matter of which is incorporated by reference in its entirety herein.

[0002] This application is also related to U.S. Patent Application serial number No. 09/915,939, entitled "Network Architechture and Methods For Transparent On-Line Cross-Sessional Encoding and Transport of Network Communications Data", filed on July 25, 2001 by Amit P. Singh, the subject matter of which is incorporated by reference in its entirety herein.

[0003] This application is also related to U.S. Patent Application serial number No. 09/872,184, entitled "System and Method for Incremental and Continuous Data Compression", filed on May 31, 2001 by Amit P. Singh, the subject matter of which is incorporated by reference in its entirety herein.

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0004] This invention relates generally to the field of Computer Networking and specifically to the field of Routing in Computer Networks.

2. Description of Background Art

[0005] A goal of conventional communication systems which convey packetized information is to be able to quickly and efficiently route individual packets of information from the source computer to the destination computer, without undue delay or loss. However, conventional routed networks rely on individual routers to decide the path along which a packet traverses across a network, e.g., a wide area network (WAN). In conventional routed networks, a router collects routing information from its neighboring routers and/or its local manually configured routes. The conventional router does not examine network statistics and routing information for any part of the network that is not directly connected to the router. The router may also gather information on routes that it has been manually configured to implement. Based on this limited local information, the router decides the next router to forward traffic to.

[0006] Conventional routers are oblivious to the path taken by a packet beyond the next router. Without being able to calculate the entire path taken by the packet, a router may inadvertently send a packet down a path which may dead-end or may significantly degrade transmission times of the packet to the required destination node.

[0007] In addition, conventional routing solutions do not adapt to select an optimal path in the WAN. In addition to the creation of lost or underliverable packets due to a conventional router's poor routing choices, routers may also have different transmission times depending on which router they are sending or receiving data from. Since conventional networks are commonly composed of a variety of different speed lines, it is possible for two separate paths through the WAN to reach the same destination at different times. Based on the local information available to a conventional router, it would be unable to accurately predict which path would reach the intended destination quicker.

[0008] These inefficiencies are exacerbated by the realization that the network is not static. The transmission speed of a router may change between an initial time and a subsequent time. Furthermore, a router may need to be taken off-line, removing a potential path as well as removing any manually configured routes. Conventional routers are unable to recognize when a router closer to the required destination has been taken off line or has slowed down significantly. Since it cannot detect disabled routers further downstream, a conventional router would not be able to intelligently choose a different path to avoid the disabled router. As such, the conventional router cannot efficiently choose its route for any given data packet.

[0009] What is needed is an efficient method and system for routing information in a dynamic multi-point network which (1) can automatically discover and track changes for an entire meshed overlay network, (2) can verify the existence of optimal routing paths through the meshed overlay network, and (3) can select an optimal routing path through the network based on up-to-date network statistics.

SUMMARY OF THE INVENTION

[0010] The present invention is a system and method for automatically identifying and verifying optimal routing paths through a dynamic multi-point meshed overlay network at an ingress point.

[0011] In one embodiment, the system may include at least one ingress router and one potential egress router located on a base network and in communication with each other. Each egress router constantly monitors local network traffic. In one embodiment, this monitoring is done in a passive capacity with the egress router not participating in conventional routing activities. In another embodiment, the egress router actively routes network packets.

[0012] In one embodiment, the egress router compiles the information and statistics regarding destinations, which are directly reached via the egress router. This information is then reported, or advertised, to the ingress router. In one embodiment the ingress router collects the advertised information from its associated egress routers and compiles an initial routing table in order to set up an overlay network.

[0013] In one embodiment, the egress router continues to monitor network traffic once the overlay network is initialized and reports changes to the ingress router. In one

embodiment, these reported changes are advertised through an incremental report, where only changed information, including added and removed destinations and changes in cost to reach a particular destination is included in the report. The ingress router updates its routing table with these incremental changes.

[0014] In one embodiment, once the system is initialized, the ingress router verifies that the destinations advertised by each egress router are reachable through the base network through that router. Additionally, in one embodiment, the ingress router may make a determination as to a total cost for forwarding packets to any given egress router versus an alternate egress router advertising the same destination.

[0015] In one embodiment, the ingress router verifies the optimal overlay paths by sending a plurality of probe packets to each destination subnet. In one embodiment the egress router which receives these packets (and presumably advertised the destination to the ingress router) stops the probe packets and reports back to the ingress router that they were received. In another embodiment, the egress router allows multiple instances of a probe packet to be forwarded on to another egress router downstream to allow the ingress router to verify a cascaded meshed overlay network.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] Figure 1 is an illustration of a virtual network having one ingress and one egress point.

[0017] Figure 2 is a flow chart illustrating a route discovery and verification technique according to one embodiment of the present invention.

[0018] Figure 3 is a flow chart illustrating a route discovery and summarization technique according to one embodiment of the present invention.

[0019] Figure 4 is a flow chart illustrating a technique for updating route summarization according to one embodiment of the present invention.

[0020] Figure 5 is an illustration of an example of a virtual network.

[0021] Figure 6 is a flow chart illustrating a route validation technique according to one embodiment of the present invention.

[0022] Figure 7 is a flow chart illustrating virtual network selection according to one embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

[0023] A preferred embodiment of the present invention is now described with reference to the figures where like reference numbers indicate identical or functionally similar elements. Also in the figures, the left most digit of each reference number corresponds to the figure in which the reference number is first used.

[0024] Reference in the specification to "one embodiment" or to "an embodiment" means that a particular feature, structure, or characteristic described in connection with the embodiments is included in at least one embodiment of the invention. The appearances of the phrase "in one embodiment" or "in an embodiment" in various places in the specification are not necessarily all referring to the same embodiment.

[0025] Some portions of the detailed description that follows are presented in terms of algorithms and symbolic representations of operations on data bits within a computer memory. These algorithmic descriptions and representations are the means used by those skilled in the data processing arts to most effectively convey the substance of their work to others skilled in the art. An algorithm is here, and generally, conceived to be a self-consistent sequence of steps (instructions) leading to a desired result. The steps are those requiring physical manipulations of physical quantities. Usually, though not necessarily, these quantities take the form of electrical, magnetic or optical signals capable of being stored, transferred, combined, compared and otherwise manipulated. It is convenient at times, principally for reasons of common usage, to refer to these signals as bits, values, elements, symbols, characters, terms, numbers, or the like. Furthermore, it is also convenient at times, to refer to certain arrangements of steps requiring physical manipulations of physical quantities as modules or code devices, without loss of generality.

[0026] It should be borne in mind, however, that all of these and similar terms are to be associated with the appropriate physical quantities and are merely convenient labels applied to these quantities. Unless specifically stated otherwise as apparent from the following discussion, it is appreciated that throughout the description, discussions utilizing terms such as "processing" or "computing" or "calculating" or "determining" or "displaying" or

"determining" or the like, refer to the action and processes of a computer system, or similar electronic computing device, that manipulates and transforms data represented as physical (electronic) quantities within the computer system memories or registers or other such information storage, transmission or display devices.

[0027] Certain aspects of the present invention include process steps and instructions described herein in the form of an algorithm. It should be noted that the process steps and instructions of the present invention could be embodied in software, firmware or hardware, and when embodied in software, could be downloaded to reside on and be operated from different platforms used by a variety of operating systems.

[0028] The present invention also relates to an apparatus for performing the operations herein. This apparatus may be specially constructed for the required purposes, or it may comprise a general-purpose computer selectively activated or reconfigured by a computer program stored in the computer. Such a computer program may be stored in a computer readable storage medium, such as, but is not limited to, any type of disk including floppy disks, optical disks, CD-ROMs, magnetic-optical disks, read-only memories (ROMs), random access memories (RAMs), EPROMs, EEPROMs, magnetic or optical cards, application specific integrated circuits (ASICs), or any type of media suitable for storing electronic instructions, and each coupled to a computer system bus. Furthermore, the computers referred to in the specification may include a single processor or may be architectures employing multiple processor designs for increased computing capability.

[0029] The algorithms and displays presented herein are not inherently related to any particular computer or other apparatus. Various general-purpose systems may also be used with programs in accordance with the teachings herein, or it may prove convenient to construct more specialized apparatus to perform the required method steps. The required structure for a variety of these systems will appear from the description below. In addition, the present invention is not described with reference to any particular programming language. It will be appreciated that a variety of programming languages may be used to implement the teachings of the present invention as described herein, and any references below to specific languages are provided for disclosure of enablement and best mode of the present invention.

[0030] The language used in the specification has been principally selected for readability and instructional purposes, and may not have been selected to delineate or

circumscribe the inventive subject matter. Accordingly, the disclosure, including the term descriptions set forth below, of the present invention is intended to be illustrative, but not limiting, of the scope of the invention.

[0031] The following terminology is used below in describing various embodiments of the present invention. These characterizations are provided to enhance the readability and understanding of one or more embodiments of the present invention. These descriptions are not intended to limit the invention and are merely illustrative.

[0032] Base Network: A conventional network consisting of interconnected networking elements are referred to as a Base Network.

[0033] Overlay Terminal: The endpoints of a virtual network circuit are referred to as overlay terminals.

[0034] Overlay Network: A collection of virtual network circuits can form an overlay network. An overlay network utilizes the base network components to establish the virtual network circuit and communicate between the overlay terminals. The overlay network can provide encapsulation of data between its terminals. A typical use of an overlay network is to provide secure connections between remote users and corporate private network even if the base network is a public network.

[0035] Overlay Router: In one embodiment, an overlay router is a routing device that can forward incoming traffic in an overlay network. This may be implemented as a software or hardware module in a conventional router. This feature can also be implemented in a standalone networking device or as a module in other networking equipment, for example. For ease of discussion the following description presumes than an overlay router is implemented in a standalone networking device.

[0036] Ingress and Egress points: In one embodiment, for unidirectional network traffic flowing through an overlay network, the overlay terminal at which the traffic enters the overlay network is called the ingress point and the overlay terminal at which the traffic leaves the overlay network is called the egress point. A bi-directional circuit can be considered as a collection of two unidirectional circuits. For bi-directional overlay network, an overlay router serves as the ingress and egress point for traffic flowing in opposite directions.

[0037] Local and remote ports: The ports on the overlay router can be partitioned into local and remote ports. A virtual network circuit can be established between two overlay routers if the overlay terminals for the circuit communicate through their remote ports. Both the ingress and egress point for a circuit can validate this configuration at their respective ends.

[0038] Supported destination subnet: In one embodiment, supported destination subnets are the subnets that are reachable through a local port of the egress point.

[0039] Any network (subnet) is a supported destination subnet if there is a host (d) in this network such that for some host (s), the path from source (s) to destination (d) in the base network includes ingress point and egress point of the circuit in that order. This path is represented as: s -> Ingress -> Egress -> d. In other words, the supported destination subnets can be subnets reachable from the egress point and the normal network path (route) in absence of a virtual network circuit from ingress point to these subnets always traverses through the egress point.

[0040] Supported source subnet: Supported source subnets can be the subnets that are reachable through a local port of the ingress point.

[0041] In one embodiment, the ingress and egress points are co-located on an overlay router device and the supported source subnets of the ingress point are the same as the supported destination subnets of the egress point on this device.

[0042] Multi-point Meshed Overlay Network: In one embodiment, a multi-point meshed overlay network is an overlay network where multiple virtual network circuits can originate from a single overlay terminal and where the virtual network circuit form a meshed (full-mesh or partial-mesh) configuration.

[0043] In one embodiment of the present invention the overlay network emulates the base network routing. Without one embodiment of this invention, one would be forced to manually configure each overlay router with a circuit routing table. Manual configuration is not a scalable solution and is prone to faulty configuration in a dynamically changing network. In one embodiment, the routing is automatically configured resulting in optimal routing in a dynamic network.

[0044] An immediate use of the invention is in the creation of dedicated tunnels at various points in the network and cache information at those elements to use lesser bandwidth. A more detailed description of this is set forth in U.S. patent application number 09/915,939, filed on July 25, 2001 by Amit P. Singh, "Network Architecture for Transparent On-line Encoding and Transport of Network Communication Data" that is incorporated by reference herein in its entirety.

[0045] Figure 1 is an illustration of one example of the present invention in a network environment. In Figure 1 an ingress point 100 and an egress point 110 form an overlay network. As noted above, in one embodiment the ingress point 100 and the egress point 110 are each overlay routes within the overlay network which form the required virtual circuit . In one embodiment, for every packet entering the ingress point 100, a decision is made regarding the selection of a virtual network circuit. In a multi-point network, the ingress point 100 serves as an overlay terminal for a number of virtual network circuits whose egress points are in diverse remote locations. Figure 1 illustrates such a virtual network having a single ingress point 100 connected to a single egress point 110 via a WAN 120. In this figure, the egress point 110 has a plurality of supported destination subnets identified as Subnet1 125, Subnet2 130, and Subnet3 135.

[0046] A virtual network circuit may serve multiple destination networks. For example, in Figure 1, the virtual network circuit formed by ingress 100 and egress 110 devices can be used to serve Subnet1 125, Subnet2 130 and Subnet3 135, but not Subnet0 115. Since Subnet0 115 is not directly reached by the egress point 110, nor is it accessible through the local ports of egress point 110, it will not be served by the virtual network circuit formed by ingress 100 and egress 110.

[0047] As described below, one embodiment of the present invention enables an ingress point 100 to automatically determine all supported destination subnets for all circuits originating from this overlay terminal and to automatically select an optimal virtual network circuit for any incoming traffic so that it can forward encapsulated traffic to an egress point 110 for the selected circuit. The notion of optimality is dependent on the application environment, and will be apparent to one skilled in the art.

[0048] Figure 2 is an illustration identifying the route discovery and verification method according to one embodiment of the present invention. One embodiment of the present

invention performs 202 route discovery and summarization then performs 204 route validation and then selects 206 the virtual network circuit. A more detailed description of each of these steps is set forth herein.

[0049] Figure 3 is an illustration of the route discovery and summarization technique 202 according to one embodiment of the present invention. In one embodiment of the present invention, automatic discovery of supported destination subnets begins by automatically discovering 310 subnets that are accessible through local ports at an egress point 110. To discover 310 the subnets that are accessibly thorugh local ports, the egress point 110 compiles a list of routes directly connected to it or reachable through its local ports. The overlay router can gather this information by passively listening to the routing update messages of standard routing protocols (e.g., routing information protocol (RIP), open shortest path first (OSPF), intermediate system to intermediate system (ISIS), border gateway protocol (BGP)) or by actively participating in the routing.

[0050] When in the passive mode, an overlay router is a transparent node in the network. Specifically, in one embodiment the overlay router can implement RIP version 2 protocol in a supplier mode and listen only for RIP updates. In another embodiment, the overlay router can implement the OSPF version 2 protocol in "host mode".

[0051] When in the active mode, the overlay router is a non-transparent device and adds a routing hop to the network path. In addition to compiling information regarding destinations on its local ports, some, or all, of the supported destination subnets may also be configured manually on the overlay router. This configuration may be performed as part of the router's operation in the base network and be implemented using any conventional configuration technique. Once the egress point 110 has collected information regarding its manually configured subnets and discovered subnets directly reachable via its local ports, it summarizes 315 this information. An egress point 110 can collapse, or consolidate, the compiled subnets into destination local subnets through an analysis of network, mask and cost metric. During the summarization 315, the network and mask values of multiple original subnets may be readjusted or consolidated into a single destination local subnet. In another embodiment, the network and mask values are consolidated using any commonly known data compression algorithm. Summarization is an optimization step to reduce the number of destination local subnets and data sent to the ingress point 100. The summarized destination local subnets are then transferred 230 to the ingress point 100. An egress point 110 can transfer 230 the

destination local subnets to multiple ingress points 100 of the overlay network. The automatic discovery and summarization 202 action may be taken at the time of creation of the virtual network circuit.

[0052] In one embodiment, the automatic discovery and summarization 202 step is continually performed by the egress point 110. The overlay router may constantly monitor for any network topology change or an attribute change (e.g., route, gateway, cost, etc.). If a change is detected, these changes are summarized (consolidated) and incrementally transferred to the ingress point 100. This will be discussed in more detail with respect to Figure 4 below.

[0053] In one embodiment of the present invention a route entry for each discovered subnet may include a network destination IP address of the destination, a route mask, a gateway IP address of the next hop router, an interface, a route type, a route protocol, a route age, and a route metric. In one embodiment the route mask is a bit mask that is typically logically ANDed with a destination address before comparing with destination subnet. In this way, a route mask may distinguish the network and host part of the destination address. *E.g.*, a mask value of 255.255.255.255 is used to indicate a host route. The interface entry is used for forwarding packets to the destination network. The interface entry may include user specified names for various network interfaces. In Table 1 below, two entries le0 and le1 are used to indicate that the interface is an Ethernet media type by including the letter "e". Additionally, the 0 and 1 are chosen as interface indices. In one embodiment the interface entry may indicate various forms of interface media including 10 megabit (MB), 100MB, or gigabit Ethernet as well as Token Ring or Fiber Distributed networks. The route type entry may be direct, indirect, invalidated route, or other. The route protocol entry notes the protocol used to discover the route by the egress point 110, *e.g.* RIP, OSPF, BGP, ICMP redirect, or other. The route age entry indicates the last update time of the route in seconds. The route metric entry is typically a value from one to five and reflects the various routing metrics depending on the routing protocol used. The route metric may also be known as a route cost and is typically a dimensionless quantity based on the specific protocol used to implement the route. For instance, for the RIP protocol, the cost is measured as the number of intermediate routers used to reach the destination. For OSPF protocols multiple metrics may be used. Typically the metric is based on the link state, which is the cost assigned to a particular interface. Additionally, the route cost may be used to indicate preferred routes and service providers by providing a lower cost metric for those preferred routes.

[0054] For illustration purposes, consider an overlay network with ingress point IP address 192.168.0.100. The following description refers to Table 1 and Figure 5. Table 1 is an example of routing information collected 310 at an egress point 110 with IP address 192.168.1.100. Figure 5 is an illustration of an example of a network in which the present invention can operate.

Destination	Route Mask	Gateway	Interface	Route/Cost Metric	Protocol
192.168.1.0	255.255.255.0	192.168.1.1	le0	0	RIP
192.168.2.0	255.255.255.0	192.168.1.1	le0	1000	static
192.168.3.0	255.255.255.0	192.168.1.1	le0	2	RIP
192.168.4.100	255.255.255.255	192.168.0.1	le1	5	OSPF
192.168.4.101	255.255.255.255	192.168.0.1	le1	5	OSPF
192.168.4.102	255.255.255.255	192.168.0.1	le1	5	OSPF
192.168.4.103	255.255.255.255	192.168.0.1	le1	5	OSPF
192.168.5.0	255.255.255.0	192.168.1.1	le0	3	RIP

Table 1

[0055] As noted above, one embodiment of the present invention summarizes 315 the subnets discovered at the egress point 110. The route attributes that are relevant to the egress point are the destination network, route mask, metric and the protocol information. The egress point 110 collapses the subnets that were compiled in the automatic discovery process into destination local subnets by analyzing the network, mask and cost metric. The network and mask values of multiple original subnets can be readjusted to a single destination local subnet. This is an optimization step to reduce the number of destination local subnets. The ingress point 100 will subsequently rebuild the values of the multiple subnets based on the condensed address and mask. One skilled in the art may recognize other ways in which to summarize the compiled route information.

[0056] The routing information in the above example is summarized below in Table 2. In this example, the four host routes (for the computers at IP addresses 192.168.4.100, 192.168.4.101, 192.168.4.102, and 192.168.4.103) are summarized into a single subnet route (*i.e.*, 192.168.4.100 with a route mask of 255.255.255.252). As noted above, once this subnet

route and route mask are reported to the ingress point 100, the the ingress point 100 will reconstruct the four host routes based on the summarized information.

Destination	Route Mask	Route/Cost Metric	Protocol
192.168.1.0	255.255.255.0	0	static
192.168.2.0	255.255.255.0	1000	static
192.168.3.0	255.255.255.0	2	RIP
192.168.4.100	255.255.255.252	5	OSPF
192.168.5.0	255.255.255.0	3	RIP

Table 2

[0057] Once the route information is summarized 315, the egress point 110 transfers 320 the destination local subnet information to the ingress point 100 of the overlay network. This action can be taken at the time of creation of the virtual network circuit. Additionally, any time the network topology or attribute change is detected (route, gateway, cost etc.) through the automatic discovery process 310, the changes are summarized 315 and incrementally transferred 320 to the ingress point 100. This process is illustrated in Figure 4.

[0058] Figure 4 is an illustration of a method for updating an existing route. The egress point 110 can remain idle 402. In one embodiment, when a new supported destination subnet is discovered 404, the egress point 110 is no longer idle. In another embodiment, the egress point 110 may also automatically discover 404 when a destination subnet is no longer supported. The egress point 110 updates 406 the list of summarized routes. The egress point 110 then can return to the idle state 402. In one embodiment, the egress point 110 runs an update ingress timer. This timer is used to indicate when the egress point 110 should forward data to ingress point 100. When an update ingress timer expires 410, the egress point exits the idle state and sends an incremental routing update from the list of summarized routes 412. The update ingress timer may then be reset. Alternatively, the update ingress timer may not be started again until new data is discovered 404. In an alternate embodiment, the egress point

110 may forward the incremental report as soon as the new data is discovered 404. Once the data is forwarded to the ingress point 100, the egress point 110 then can return to the idle state 402.

[0059] Once the ingress point 100 has received routing tables from one or more egress points 110, it can then perform route validation 204. Figure 6 is a flowchart of a route validation procedure according to one embodiment of the present invention. A technique for validating routes is as follows: the ingress point generates 606 a number of host addresses covered by a destination local subnet. In one embodiment the number of host addresses is three. These sample addresses may be equally spaced from each other and selected from the set of addresses belonging to a destination local subnet. The ingress point 100 sends 608 this list of host addresses to the egress point 110 indicating that diagnostic packets can be sent to these host addresses.

[0060] After receiving 618 this route validation start information, the egress point 110 acknowledges 620 the receipt of the message, and examines the incoming packets on its remote ports for a probe packet for a time period. In one embodiment the time period is one minute, and is referred to as the probe interval. Once the ingress point 100 receives 610 the acknowledgement, the ingress point 100 sends 612 a specified diagnostic probe packet to each of these host addresses on the base network (not utilizing the overlay network).

[0061] During the probe interval 622, the egress point 110 identifies all the probe packets it observes. In one embodiment, it also terminates the probe packets and does not forward these packets to the actual host destination address. In one embodiment, the egress point 110 forwards the second or higher instance of same message because of the possibility that it is meant for another egress point 110 in cascade on the path to the destination host. At the end 622 of the probe interval, the egress point 110 sends 624 the list of validated addresses to the ingress point 100. After receiving 614 the route validation result, the ingress point 100 updates or creates 616 a circuit routing table.

[0062] For every destination local subnet received 604 from an egress point 110, the ingress point 100 initiates the route validation procedure to ensure that incoming traffic at the ingress point 100 that is destined to any arbitrary destination covered by a destination local subnet indeed traverses through the egress point 110 that advertised the specific destination

local subnet. The following discussion provides an example of the route validation process with reference to Figure 5 and Figure 6.

[0063] The ingress point 100 generates 606 a number of host addresses covered by a destination local subnet. In one embodiment, the number of host addresses is three. These sample probe addresses may be equally spaced from each other and selected from the set of addresses belonging to a destination local subnet. For example, with reference to Figure 5, the probe addresses covering a destination subnet 192.168.2.0 and network mask 255.255.255.0 in our example routing information may be 192.168.2.64, 192.168.2.128 and 192.168.2.192. Similarly, the probe addresses covering the destination subnet 192.168.4.100 and network mask 255.255.255.252 may be 192.168.4.100, 192.168.4.101 and 192.168.4.102. Uniqueness in the generated sample addresses is desired but not required. Thus, if the network and mask combination is too restrictive to generate the required number of unique probe addresses, some or all of the sample addresses may be repeated. It may also be noted that there is no need to validate the destination subnet covered by the egress point 110 (e.g. 192.168.1.0 subnet in the example set forth in Figure 5) since the egress point best represents the subnet which is local to it, and no additional cost will be associated with that particular subnet.

[0064] The ingress point 100 sends 608 the list of generated probe addresses to the egress point 110 indicating that diagnostic packets will be sent to these host addresses. The egress point 110 receives 618 and acknowledges 620 the receipt of the message, and examines the incoming packets on its remote ports for the probe packets for the probe interval. The ingress point 100 receives 610 this acknowledgement and sends 612 the specified diagnostic probe packets to these host addresses on the base network (not utilizing the overlay network). During the probe interval, the egress point 110 identifies all the probe packets it observed. In one embodiment, it also terminates the probe packets. That is, the egress point 110 does not forward the probe packets to the actual host destination address. In one embodiment the egress point 110 will however forward the second or higher instance of same message because of the likelihood that it is meant for another egress point 110 in cascade on the path to the destination host. At the end of the probe interval 622 the egress point 110 sends 624 the list of validated addresses to the ingress point 100 that updates or creates 616 it's routing table.

[0065] Additionally, the overlay router can periodically validate 604 all the destination local subnets. Route validation can also be done when one of the virtual network circuits is terminated (which may be caused by policy change, unavailability of network element or

network breakdown etc.). In this case, the route validation can be done for the full network (i.e., all the routes) or for the routes the virtual network circuit has previously validated.

[0066] Once the routes are validated, the present invention performs 206 virtual network circuit selections at the ingress point 100. In one embodiment, the overlay router serving as the ingress point 100 maintains a circuit routing table. The purpose of this table is to enable the ingress point 100 to select the optimal virtual network circuit for any incoming network traffic. Each entry in the table contains a destination local subnet and the egress point 110 that validated the subnet. This table can be populated and updated by the ingress point 100 after a route validation. The validated destination subnets are listed on this table.

[0067] Figure 7 is a flowchart illustrating the virtual network selection process 206 according to one embodiment of the present invention. For any incoming packet that the overlay router can forward 702, it consults the circuit routing table to find a matching entry for the destination address of the packet to determine 704 if a virtual network circuit is available for the destination. If a match is found, the corresponding egress point is selected and the packet is forwarded 706 on the relevant virtual network circuit to reach the egress point. If no match is found, the incoming packet is forwarded 708 on the base network in its un-encapsulated form. In one embodiment, the incoming packet is encapsulated in another header which is addressed to the egress point 110 selected in step 704. The egress point 110 then strips the additional header and forwards the packet to its destination. In this manner, the ingress point 110 is ensured that that packet took the fastest available route through the base network.

[0068] Table 3 is an example of a circuit routing table at the ingress point 192.168.0.100 and may be used by an ingress point 110 to select a virtual network circuit.

Destination	Route Mask	Metric	Protocol	Egress Point
192.168.1.0	255.255.255.0	0	OSPF	192.168.1.100
192.168.2.0	255.255.255.0	1	OSPF	192.168.2.111
192.168.3.0	255.255.255.0	2	RIP	192.168.1.100
192.168.4.100	255.255.255.252	5	OSPF	192.168.1.100
192.168.4.200	255.255.255.240	4	OSPF	192.168.2.111
192.168.6.0	255.255.255.0	2	RIP	100.200.1.200
192.168.7.0	255.255.255.0	5	OSPF	100.200.2.222

Table 3

[0069] The above table indicates that for the destination subnet 192.168.2.0 and subnet mask 255.255.255.0, the selected egress point is 192.168.2.111. The egress point 110 with IP address 192.168.1.100 also advertised the same destination network as shown in Table 1. The reason for the selection of 192.168.2.111 is because this egress point has validated the probe packets destined for the network and the cost to reach the destination network is smaller (1) from 192.168.2.111 compared to the cost to reach the same destination network from 192.168.1.100 (1000). The destination subnet 192.168.5.0 with network mask 255.255.255.0, which is advertised by egress point 110 with IP address 192.168.1.100 is not included in the circuit routing table for ingress point 100 with IP address 192.168.0.100. The reason is because the egress point could not validate the probe packets destined for this network.

[0070] An efficient way to store information in circuit routing table is similar to a conventional routing table. A Patricia tree, radix tree or other variations of balanced binary tree can be effectively used for fast table lookup. An overlay router examines 704 the incoming packet for the destination address. It performs a table lookup to check if there is a destination local subnet entry that covers the destination address. If multiple circuits are available for a destination local subnet, the circuit list is sorted in increasing order of cost (or any other measures of optimality for an application - function of cost and hop count to the egress point etc.). Thus the front of the list gives the optimal circuit to be used for the incoming packet.

[0071] In one embodiment, a probabilistic method is used to select a circuit. In this embodiment the ingress point 100 sends a larger number of diagnostic probe packets covering the destination local subnet. In one embodiment, the number of diagnostic probe packets is 100. Each egress point 110 responds with the number of the probe packets observed by it. These numbers form the basis for the weights to be assigned to the overlay network circuits. The ingress point 100 can probabilistically select any of these egress points 100 in line with

their weights. If not all probe packets are acknowledged by the corresponding egress point(s), the circuit selection mechanism at an ingress point 100 may select the virtual network circuit and the base network at a frequency in line with the number of probe responses received.

[0072] While the invention has been particularly shown and described with reference to a preferred embodiment and several alternate embodiments, it will be understood by persons skilled in the relevant art that various changes in form and details can be made therein without departing from the spirit and scope of the invention.

What is claimed is:

1. An overlay network for transporting a data packet through a base network, comprising:
   an ingress point coupled to the base network, configured to maintain a virtual network routing table; and
   an egress point coupled to the base network, configured to gather information corresponding to a destination local subnet associated with the egress point, but not associated with the ingress point, and to send the gathered information to the ingress point for inclusion in the virtual network routing table;
   the ingress point further configured to select a virtual network circuit for transporting the data packet, the virtual network circuit beginning at the ingress point and ending at the egress point, and the ingress point selecting the circuit responsive to the virtual network routing table and a destination of the data packet.

2. The overlay network of claim 1 wherein:
   the ingress point comprises a first router; and
   the egress point comprises a second router.

3. The overlay network of claim 1 wherein the base network is one of a point-to-point network; a multi-point network, or a dynamic multi-point network.

4. The overlay network of claim 1 wherein the egress point consolidates the gathered information, to reduce the amount of information to be transmitted to the ingress point.

5. The overlay network of claim 1 further comprising a second egress point coupled to the base network configured to gather information corresponding to a second destination local subnet associated with the second egress point, and to send the gathered information to the ingress point for inclusion in the virtual network routing table.

6. The overlay network of claim 1 wherein the egress point is configured to monitor changes to the associated local destination subnet and to transmit updated information corresponding to the associated local destination subnet to the ingress point.

7. The overlay network of claim 1 wherein the gathered information includes at least one of a network IP address of the destination, a route mask, a gateway IP address of the next hop router, an interface type, a route type, a route protocol, a route age, or a route cost.

8. The overlay network of claim 1 wherein the ingress point is further configured to validate a route through the base network to the destination local subnet listed in the virtual network routing table by transmitting a probe packet to an address in the destination local subnet and receiving confirmation of detection from the egress point.

9. The overlay network of claim 8:

further comprising at least one additional egress point coupled to the base network configured to gather information corresponding to at least one additional destination local subnet associated with the at least one additional egress point, and to send the gathered information to the ingress point for inclusion in the virtual network routing table; and

wherein the virtual network routing table comprises at least one additional listing corresponding to the at least one additional destination local subnet, and the ingress point is further configured to validate the at least one additional entry by sending a probe packet to an address located in the at least one additional destination local subnet and receiving a confirmation from the at least one additional egress point that the probe packet was detected.

10. The overlay network of claim 8 wherein the ingress point is configured to transmit a plurality of probe packets to a plurality addresses located in the destination local subnet.

11. The overlay network of claim 8 wherein the egress point terminates the probe packet when detected.

12. The overlay network of claim 8 wherein the ingress point is configured to re-validated the subnet listed in the virtual network routing table.

13. In an overlay network, a method for verifying a virtual network routing table comprising the steps of:

discovering information corresponding to a destination local subnet associated with an egress point;

maintaining the virtual network routing table at an ingress point responsive to the discovered information; and

selecting a virtual network circuit responsive to the virtual network routing table, the virtual circuit beginning at the ingress point and ending at the egress point.

14. The method of claim 13 further comprising the steps of:

consolidating the discovered information to reduce the amount of information; and

transmitting the summarized information to the ingress point.

15. The method of claim 13 further comprising the steps of:

monitoring the destination local subnet for a change at the egress point; and

updating the virtual network routing table responsive to the updated information.

16. The method of claim 13 further comprising the step of discovering information corresponding to a second destination local subnet associated with a second egress point; and wherein the step of maintaining the virtual network routing table is additionally responsive to information corresponding to the second destination local subnet.

17. The method of claim 13 wherein the step of discovering information comprises the egress point passively monitoring routing a protocol message corresponding to the associated destination local subnet on a base network.

18. The method of claim 13 wherein the step of discovering information comprises the egress point actively routing data to the associated destination local subnet on a base network and collecting a routing protocol message corresponding to the associated destination local subnet.

19. The method of claim 13 further comprising the steps of:

sending a probe packet from the ingress point to an address within the destination local subnet corresponding to a routing entry in the virtual network routing table;

confirming detection of the probe packet by the egress point corresponding to the destination local subnet; and

modifying the virtual network routing table responsive to the confirmation to indicate the confirmation of the routing entry.

20. The method of claim 19 further wherein the step of sending a probe packet comprises sending an at least one additional probe packet to an at least one additional address within the destination local subnet.

21. The method of claim 19 wherein the step of sending a probe packet comprises sending a probe packet to the address to reconfirm detection of the probe packet by the egress point.

22. The method of claim 19:

further comprising the step of discovering an additional information corresponding to an at least one additional destination local subnet associated with an at least one additional egress point;

wherein the step of maintaining the virtual network routing table is further responsive to the additional information; the virtual network routing table comprises at least one additional destination local subnet entry corresponding to the transmitted additional information; and the step of sending a probe packet comprises sending an at least one additional probe packet from the ingress point to an address within the at least oneadditional destination local subnet entry.

23. An overlay network for transporting a data packet through a base network comprising:

An egress point means for discovering information corresponding to a destination local subnet associated with the egress point means and for transmitting the discovered information to an ingress point means; and

an ingress point means for maintaining a virtual network routing table responsive to the transmitted information; and for selecting a virtual network circuit responsive to the virtual network routing table, the virtual network circuit beginning at the ingress point means and ending at the egress point means.

24. The overlay network of claim 23 wherein the egress point means further comprises means for consolidating the discovered information to reduce the amount of information to be transmitted to the ingress point means.

25. The overlay network of claim 23 wherein:

the egress point means further comprises means for monitoring the destination local subnet for a change and for transmitting an updated information corresponding to the change to the ingress point means; and

the ingress point means further comprises means for updating the virtual network routing table responsive to the updated information.

26. The overlay network of claim 23 further comprising:

an at least one additional egress point means for discovering an at least one additional information corresponding to an at least one additional destination local subnet

associated with the at least one additional egress point means and for transmitting the at least one additional information to the ingress point means;

wherein the ingress point means additionally maintains the virtual network routing table responsive to the at least one additional destination local subnet.

27. The overlay network of claim 23 wherein the egress point means passively monitors a routing protocol message corresponding to the associated destination local subnet on a base network.

28. The overlay network of claim 23 wherein the egress point means actively routes data to the associated destination local subnet on a base network and collects a routing protocol message corresponding to the associated destination local subnet.

29. The overlay network of claim 23 wherein the ingress point means further comprises:

a means for sending a probe packet from the ingress point means to an address within the destination local subnet corresponding to a routing entry in the virtual network routing table;

a means for confirming receipt of the probe packet by the egress point means; and

a means for modifying the virtual network routing table respoinsive to the confirmation.

30. The overlay network of claim 29 wherein the means for sending a probe packet further comprises means for sending an at least one additional probe packet to an at least one additional address within the destination local subnet.

31. The overlay network of claim 29 wherein the means for sending a probe packet further comprises means for sending a probe packet to the address to reconfirm detection of the probe packet.

32. An overlay network for transmitting a data packet throught a base network, comprising:

an egress point coupled to the base network, configured to forward the data packet to an associated destination local subnet; and

an ingress point coupled to the base network, configured to maintain a virtual network routing table comprising a routing entry corresponding to the destination local subnet; to validate the routing entry by sending a probe packet to an address located in the destination local subnet and by receiveing a confirmation from

the egress point that the probe packet was detected; and to select a virtual network circuit beginning at the ingress point and ending at the egress point responsive to the validated routing entry and the destination of the data packet.

33. The overlay network of claim 32 further comprising:

at least one additional egress point coupled to the base network and configured to forward the data packet to at least one additional associated destination local subnet; and

wherein the virtual network routing table comprises at least one additional routing entry corresponding to the at least one additional destination local subnet, and the ingress point is further configured to validate the at least one additional routing entry by sending an at least one additional probe packet to an address located in the at least one additional destination local subnet and receiving a confirmation from the at least one additional egress point that the at least one additional probe packet was detected.

34. The overlay network of claim 32 wherein the ingress point is configured to send multiple probe packets to separate addresses located in the destination local subnet.

35. The overlay network of claim 32 wherein the egress point serves the destination local subnet directly and the ingress point does not need to validate the entry associated with the subnet.

36. The overlay network of claim 32 wherein:

the ingress point is located in a first router; and

the egress point is located in a second router.

37. The overlay network of claim 32 wherein the egress point terminates the probe packet when the probe packet is detected by the egress point..

38. The overlay network of claim 37 wherein the egress point allows a subsequent identical probe packet to continue to the destination local subnet address.

39. The overlay network of claim 32 wherein the ingress point is configured to re-validate the routing entry.

40. In an overlay network, a method for validating a virtual network routing table comprising the steps of:

maintaining a virtual network routing table at an ingress point;

sending a probe packet from the ingress point to an address within a destination local subnet corresponding to an entry in the virtual network routing table;

confirming detection of the probe packet by an egress point corresponding to the destination local subnet;

modifying the virtual network routing table responsive to the confirmation to indicate a validated entry; and

selecting a virtual network circuit responsive to the validated entry in the virtual network routing table, the virtual network circuit selected to begin at the ingress point and to end at the egress point.

41. The method of claim 40 wherein the step of sending a probe packet comprises sending the probe packet on a base network.

42. The method of claim 40 wherein:

the virtual network routing table comprises at least one additional destination local subnet entry; and

the step of sending a probe packet comprises sending an at least one additional probe packet from the ingress point to an address within the at least one additional destination local subnet entry.

43. The method of claim 40 wherein the step of sending a probe packet comprises sending at least one additional probe packet to at least one additional address within the destination local subnet.

44. The method of claim 40 wherein the step of sending a probe packet comprises sending a probe packet to the address to reconfirm detection of the probe packet.

45. A system for validating a virtual network routing table in an overlay network comprising:

an ingress point means for maintaining the virtual network routing table and for sending a probe packet to a network address within a destination local subnet corresponding to a routing entry in the virtual network routing table;

a egress point means associated with the destination local subnet for confirming detection of the probe packet; and

the ingress point means further configured for modifying the virtual network routing table responsive to the confirmation from the egress point means to indicate a validated routing entry; and for selecting a virtual network circuit responsive to the validated routing entry, the virtual network circuit beginning at the ingress point means and ending at the egress point means.

46. The system of claim 45 wherein the ingress point means is further for sending the probe packet on a base network.

47. The system of claim 45 wherein:

the virtual network routing table comprises an at least one additional destination local subnet entry; and

the ingress point means further comprises means for sending an at least one additional probe packet from the ingress point means to an address within the at least one additional destination local subnet entry.

48. The system of claim 45 wherein the ingress point means further comprises means for sending an at least one additional probe packet to an at least one additional address within the destination local subnet.

49. The system of claim 45 wherein the ingress point means further comprises means for sending a probe packet to the address periodically to reconfirm detection of the probe packet.

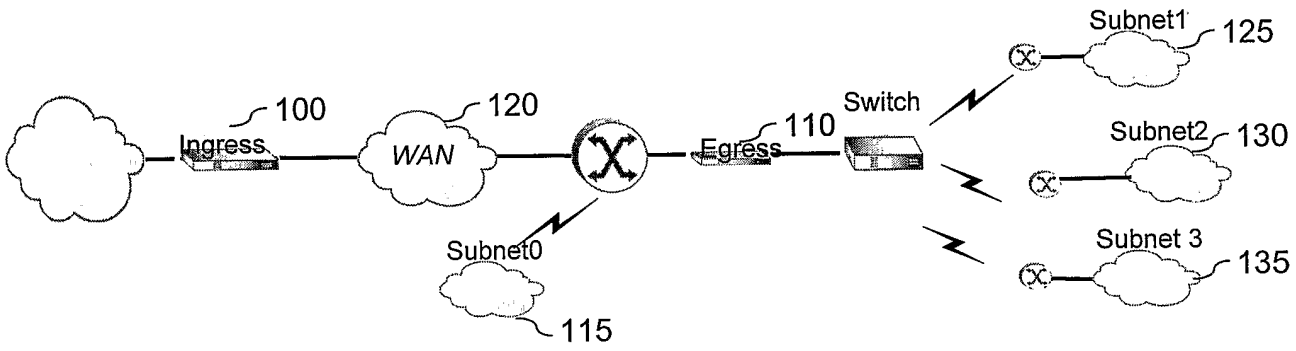

Subnet1
125
100
120
Switch
Ingress
WAN
110
Egress
Subnet2
130
Subnet0
Subnet 3
135
115

Figure 1

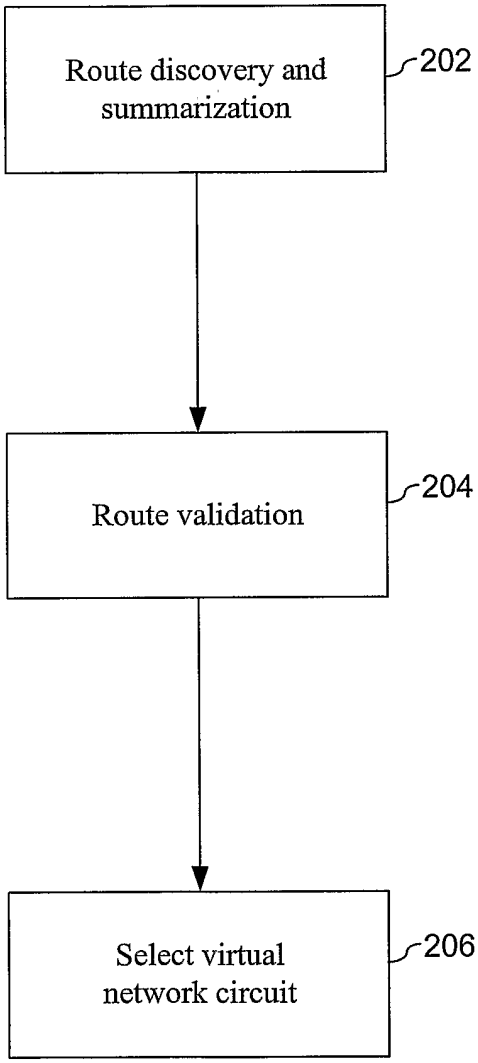

Route discovery and summarization
202
Route validation
204
Select virtual network circuit
206

Figure 2

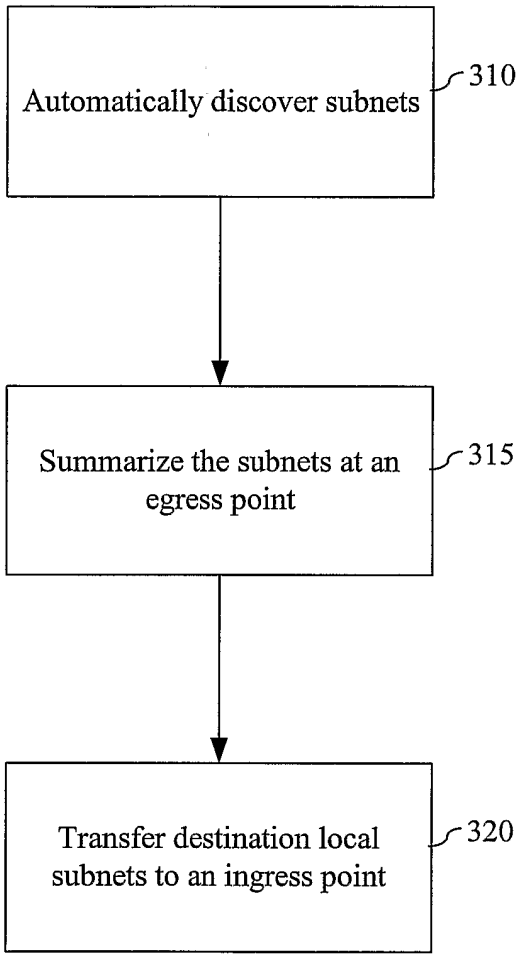

Automatically discover subnets
310
Summarize the subnets at an egress point
315
Transfer destination local subnets to an ingress point
320

Figure 3

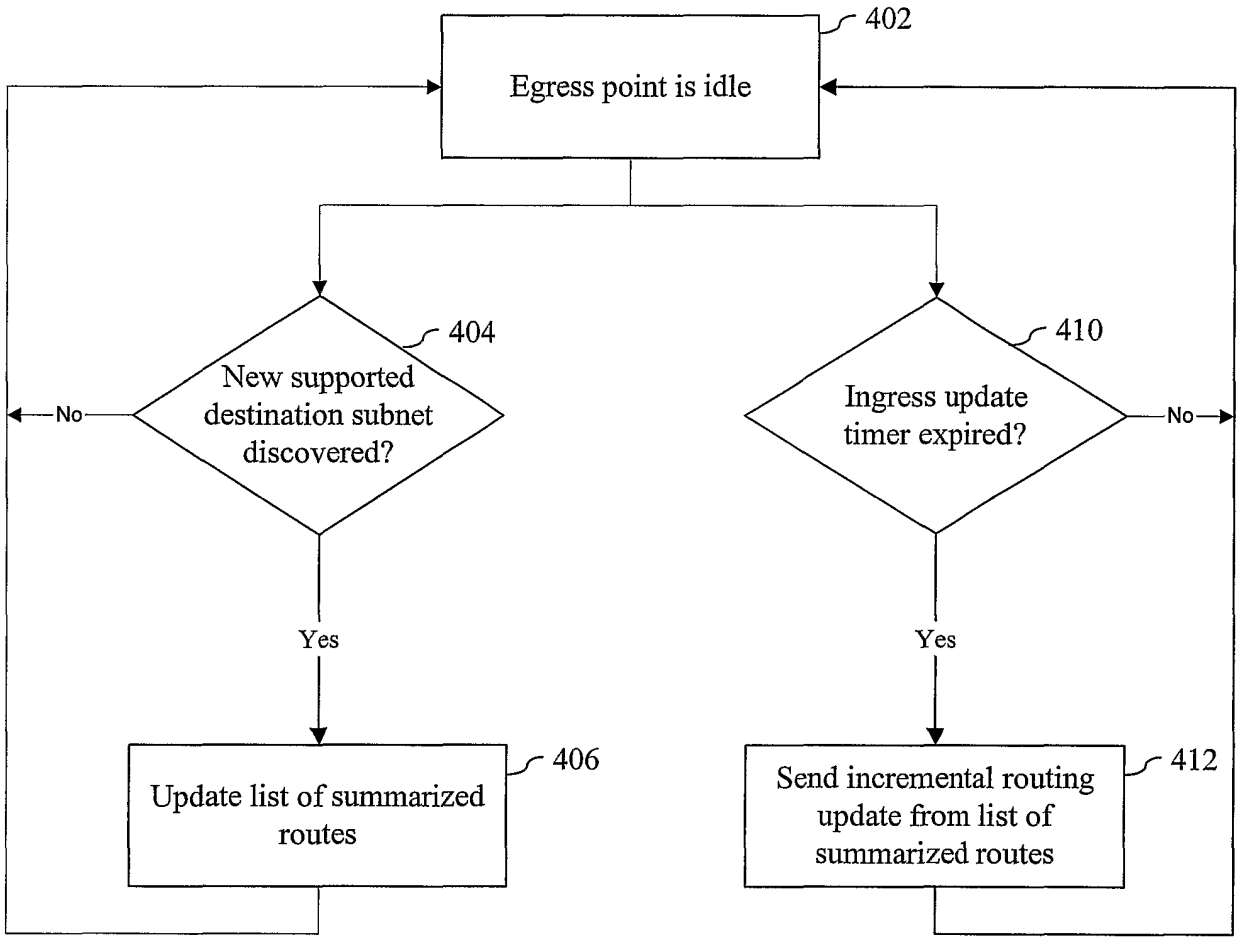

402
Egress point is idle
404
New supported destination subnet discovered?
No
Yes
406
Update list of summarized routes
410
Ingress update timer expired?
No
Yes
412
Send incremental routing update from list of summarized routes

Figure 4

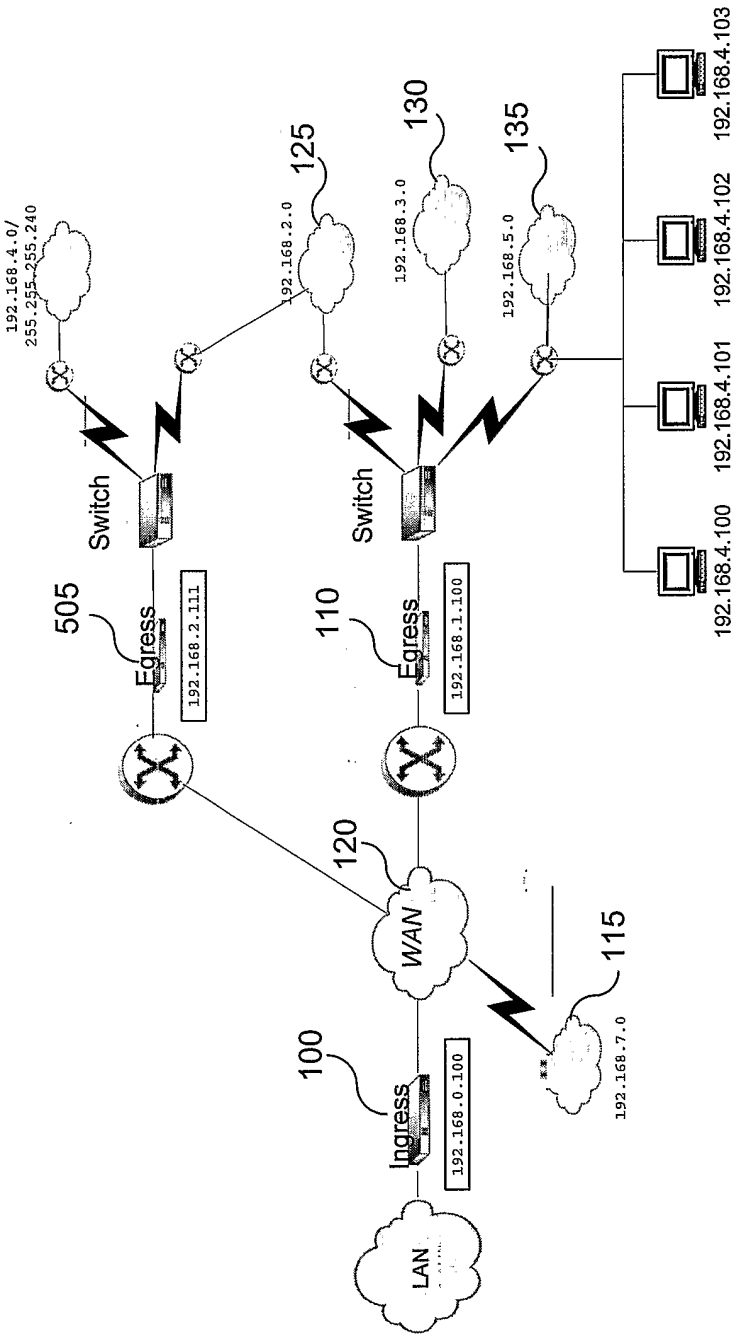

192.168.4.0/
255.255.255.240
125
192.168.2.0
130
192.168.3.0
135
192.168.5.0
192.168.4.103
192.168.4.102
192.168.4.101
192.168.4.100
Switch
Switch
505
Egress
192.168.2.111
110
Egress
192.168.1.100
120
WAN
115
192.168.7.0
100
Ingress
192.168.0.100
LAN

Figure 5

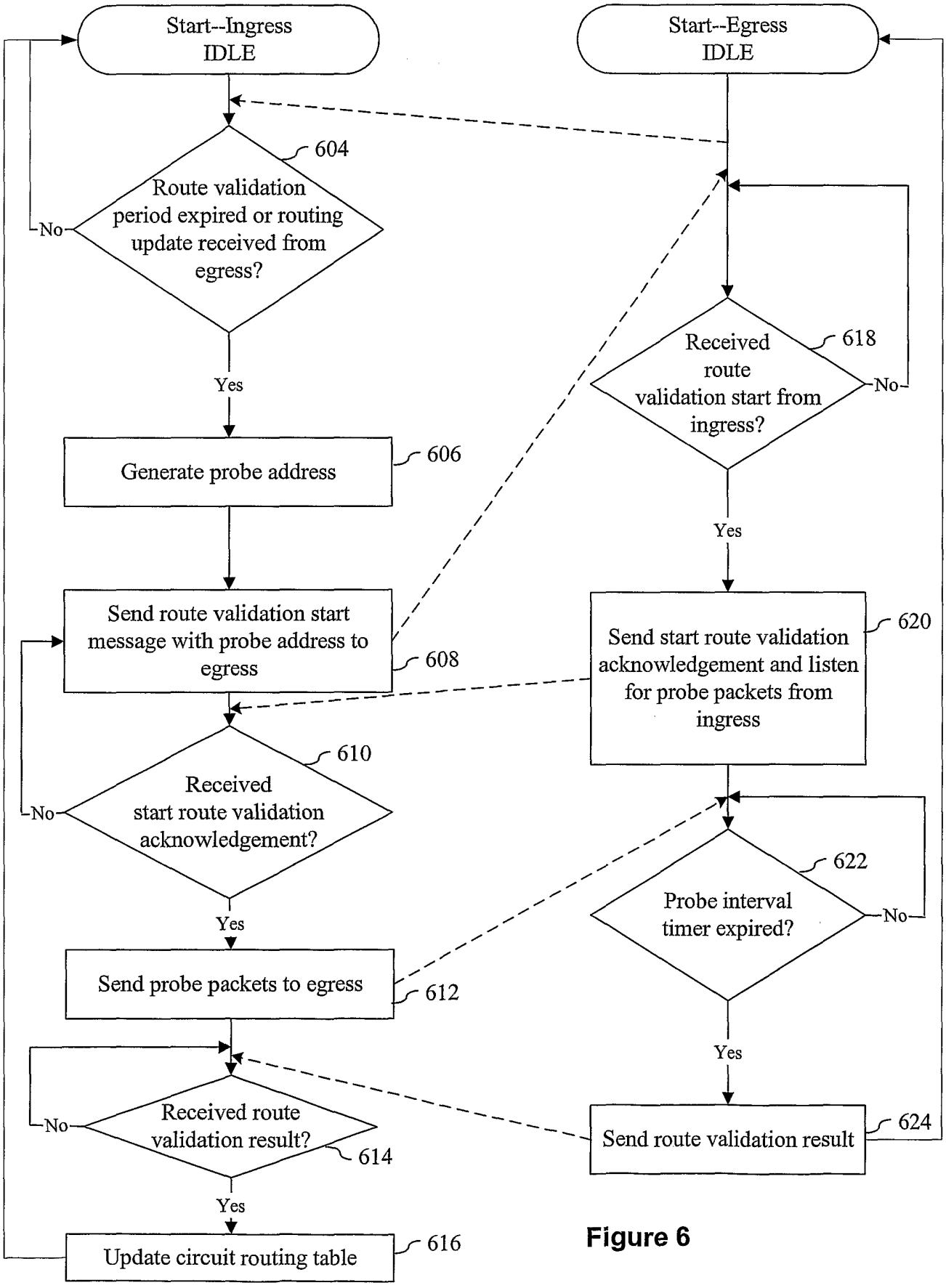

Start--Ingress
IDLE
Start--Egress
IDLE
604
Route validation period expired or routing update received from egress?
No
Yes
Generate probe address
606
Send route validation start message with probe address to egress
608
Received start route validation acknowledgement?
610
No
Yes
Send probe packets to egress
612
Received route validation result?
614
No
Yes
Update circuit routing table
616
Received route validation start from ingress?
618
No
Yes
Send start route validation acknowledgement and listen for probe packets from ingress
620
Probe interval timer expired?
622
No
Yes
Send route validation result
624

Figure 6

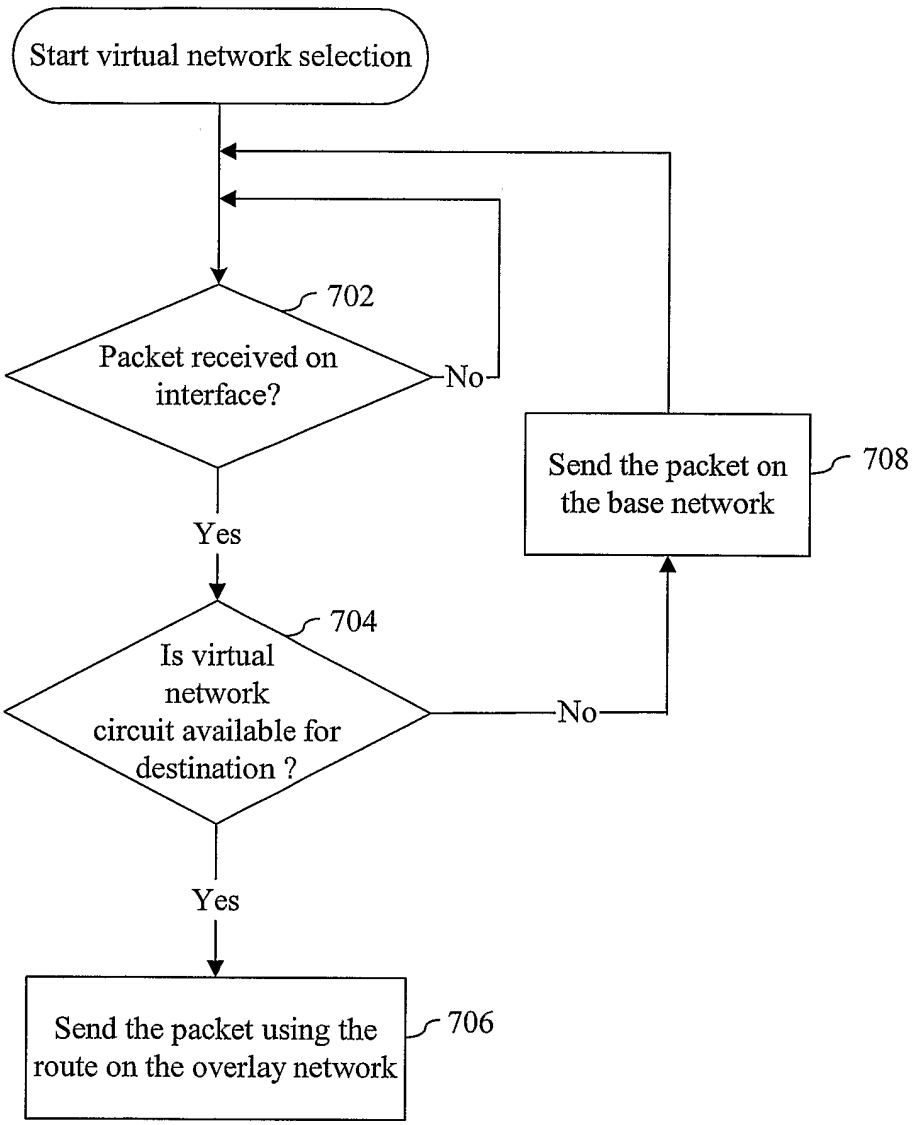

Start virtual network selection
702
Packet received on interface?
No
Yes
704
Is virtual network circuit available for destination ?
No
Yes
708
Send the packet on the base network
706
Send the packet using the route on the overlay network

Figure 7