



US009270487B1

(12) **United States Patent**  
**Nguyen et al.**

(10) **Patent No.:** **US 9,270,487 B1**  
(45) **Date of Patent:** **Feb. 23, 2016**

(54) **FULL BISECTION BANDWIDTH NETWORK**

2004/0190454 A1\* 9/2004 Higasiyama ..... 370/238  
2009/0274153 A1\* 11/2009 Kuo et al. .... 370/392  
2011/0302346 A1\* 12/2011 Vahdat et al. .... 710/301

(75) Inventors: **Chinh Kim Nguyen**, San Diego, CA (US); **Curtis Hall Stehley**, Baltimore, MD (US)

**OTHER PUBLICATIONS**

(73) Assignee: **Teradata US, Inc.**, Dayton, OH (US)

Dell Inc., "Dell PowerConnect 6200 Systems CLI Reference Guide", (Oct. 2006).

(\* ) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 1003 days.

Leiserson, Charles E., "Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing", *IEEE Transactions on Computers*, vol. C-34, No. 10, Oct. 1985, (Oct. 1, 1985),892-901.

(21) Appl. No.: **12/577,979**

Davis, David "Preventing network loops with Spanning-Tree Protocol (STP)", [http://www.petri.co.il/csc\\_preventing\\_network\\_loops\\_with\\_stp\\_8021d.htm](http://www.petri.co.il/csc_preventing_network_loops_with_stp_8021d.htm), (Jan. 7, 2009).

(22) Filed: **Oct. 13, 2009**

"IEEE 802.1Q", [http://en.wikipedia.org/wiki/IEEE\\_802.1Q](http://en.wikipedia.org/wiki/IEEE_802.1Q).

(51) **Int. Cl.**  
**H04L 12/46** (2006.01)  
**H04L 12/54** (2013.01)

\* cited by examiner

(52) **U.S. Cl.**  
CPC ..... **H04L 12/4641** (2013.01); **H04L 12/5689** (2013.01)

*Primary Examiner* — Faruk Hamza  
*Assistant Examiner* — Tito Pham

(74) *Attorney, Agent, or Firm* — Howard Speight, PLLC

(58) **Field of Classification Search**  
None  
See application file for complete search history.

(57) **ABSTRACT**

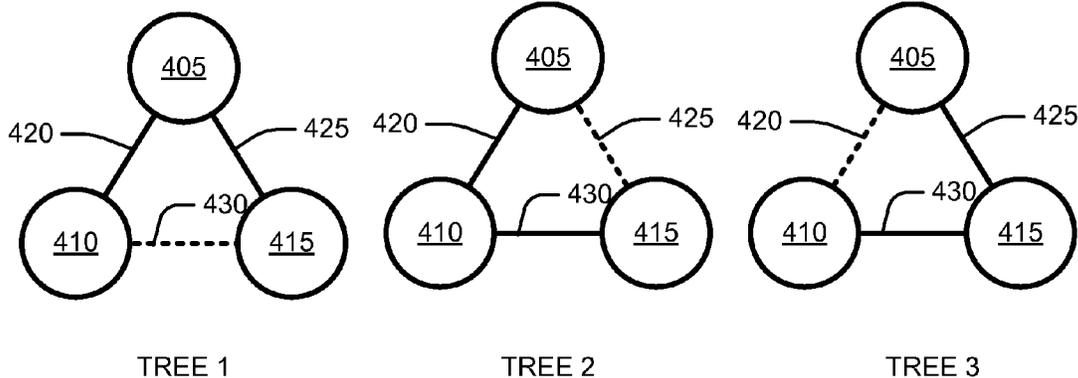
A full bisection bandwidth network, having a plurality of nodes and a plurality of paths among the nodes, is divided into a plurality of Virtual Local Area Networks ("VLANs") by assigning paths to the VLANs such that each VLAN satisfies a spanning tree protocol and all paths are active in at least one VLAN.

(56) **References Cited**

**U.S. PATENT DOCUMENTS**

7,606,178 B2\* 10/2009 Rahman et al. .... 370/256

**27 Claims, 18 Drawing Sheets**



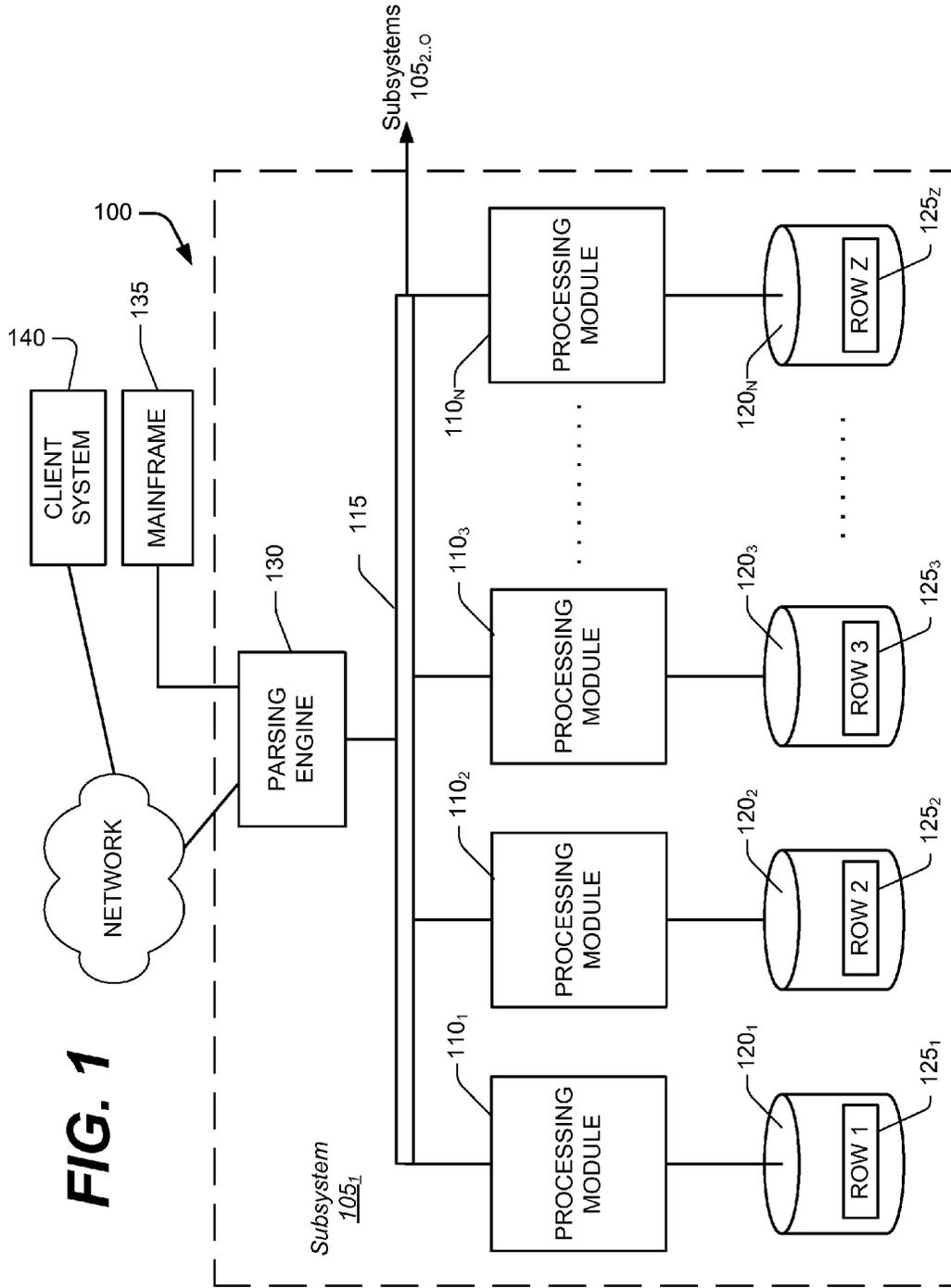
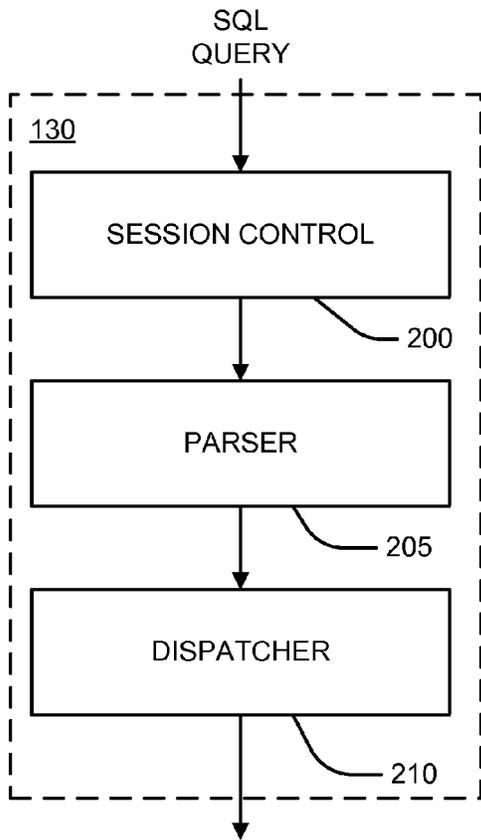
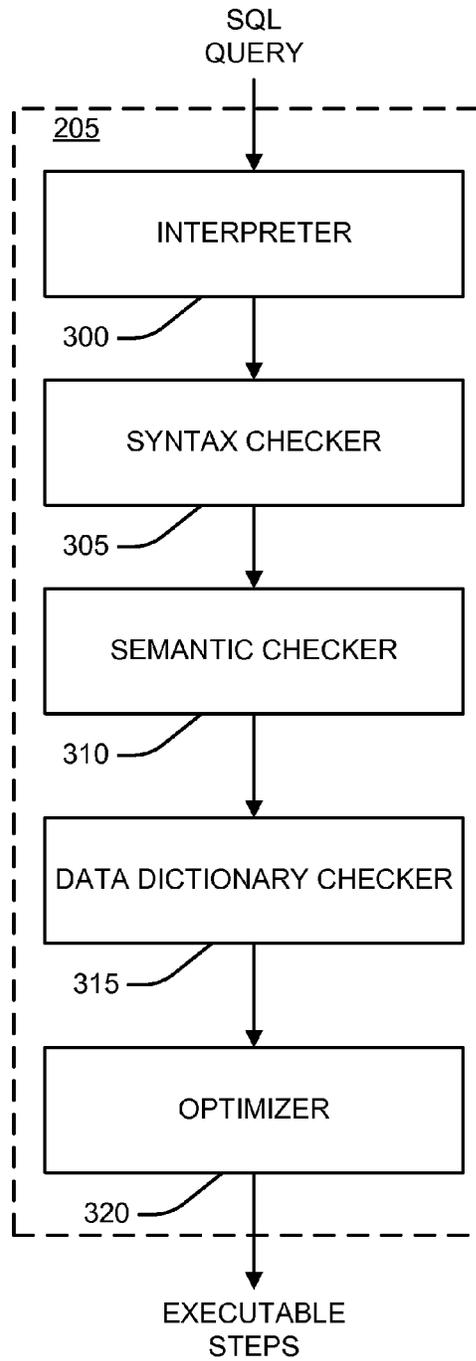


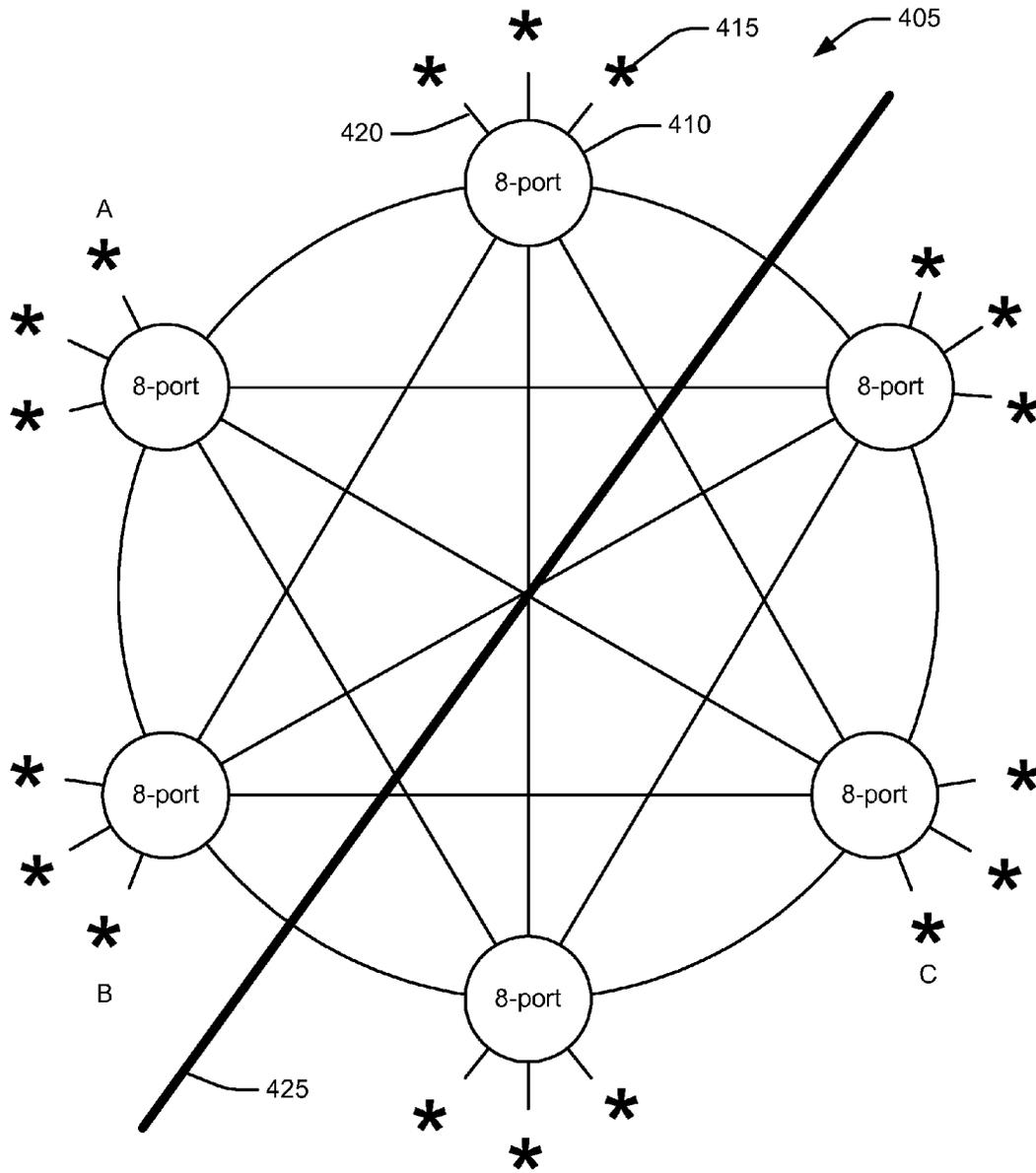
FIG. 1



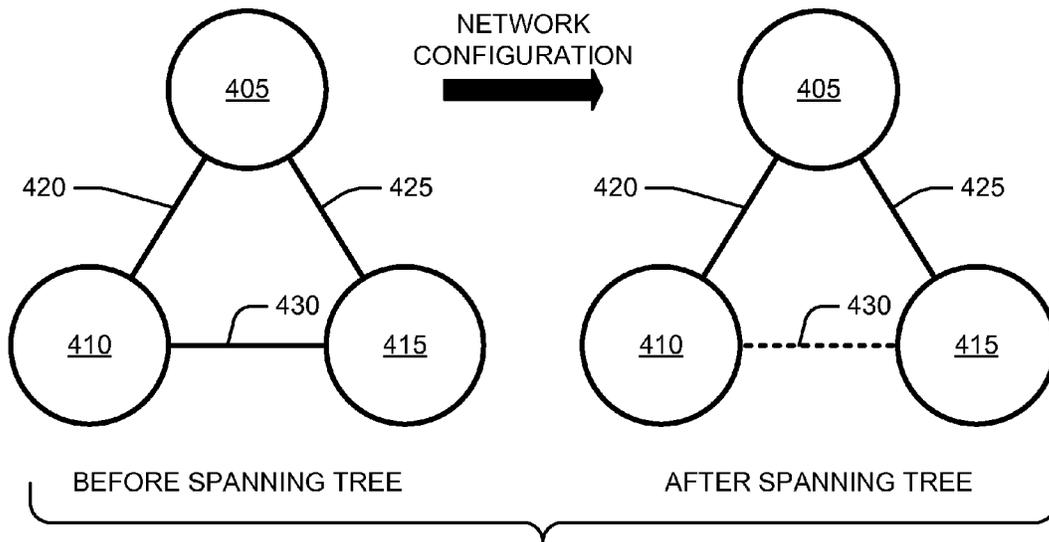
**FIG. 2**



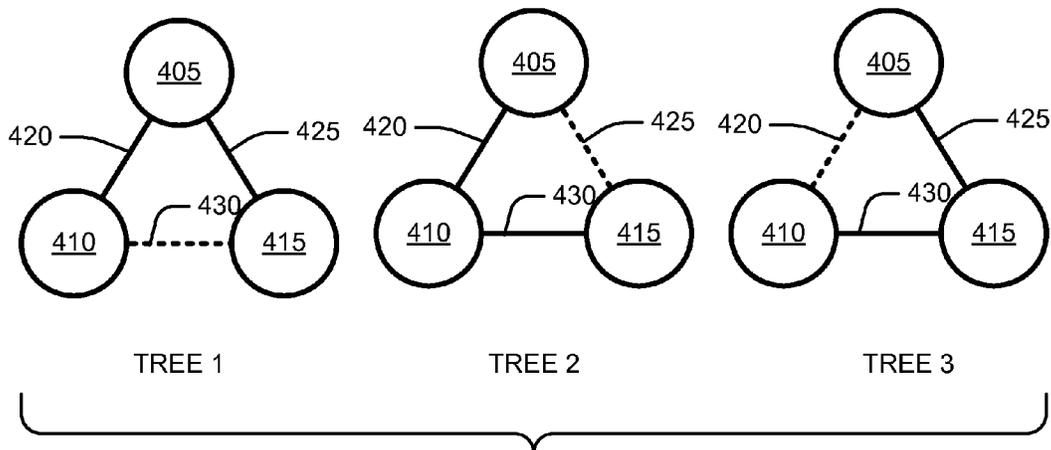
**FIG. 3**



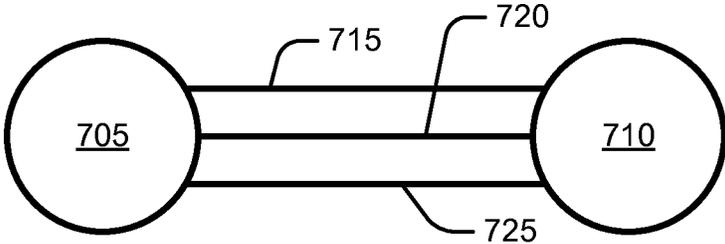
**FIG. 4**



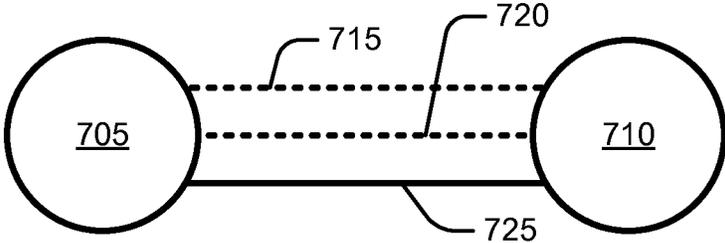
**FIG. 5**



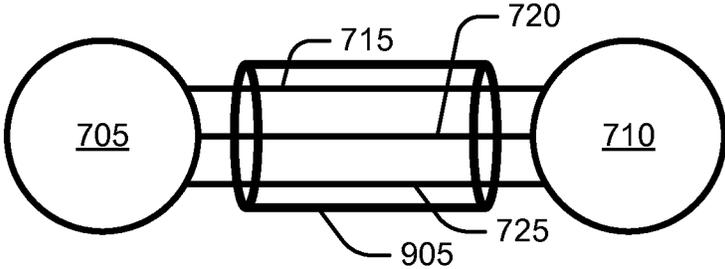
**FIG. 6**



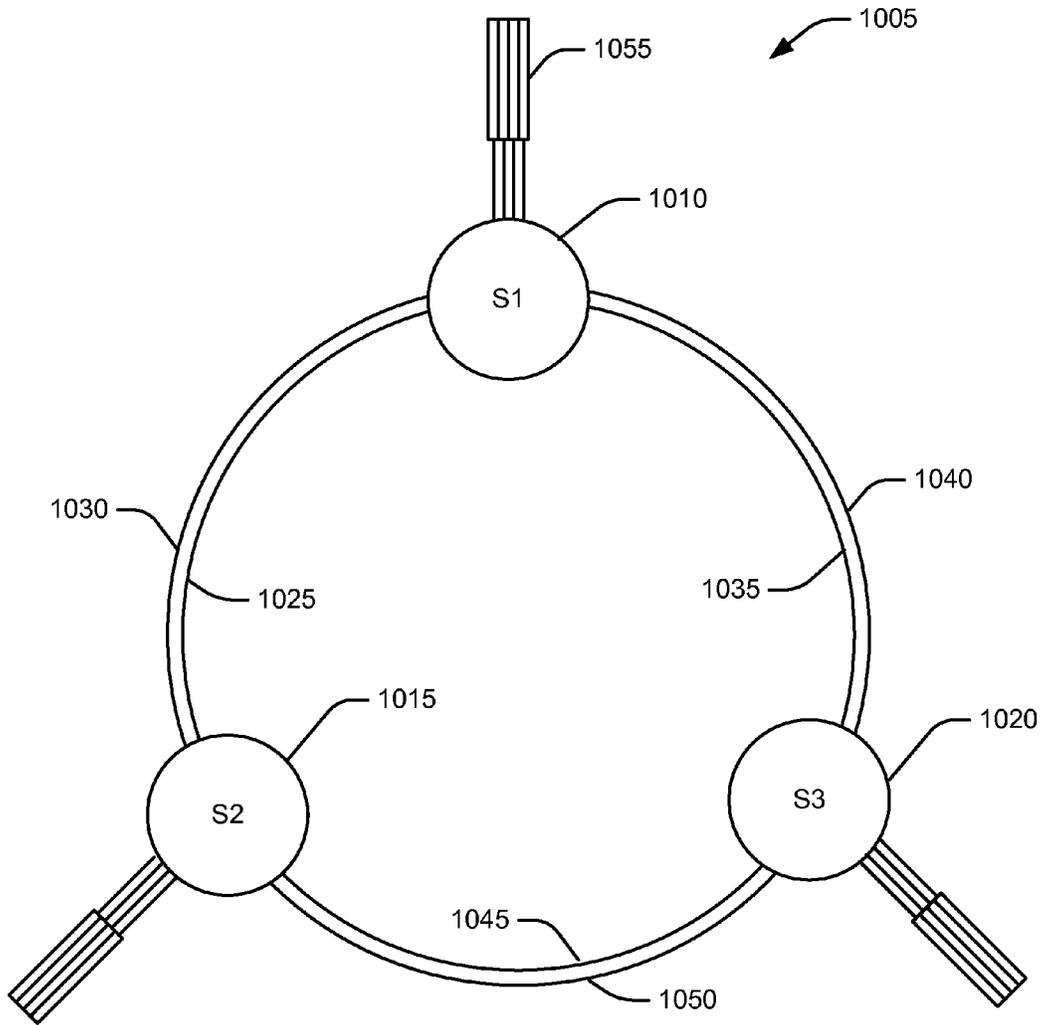
**FIG. 7**



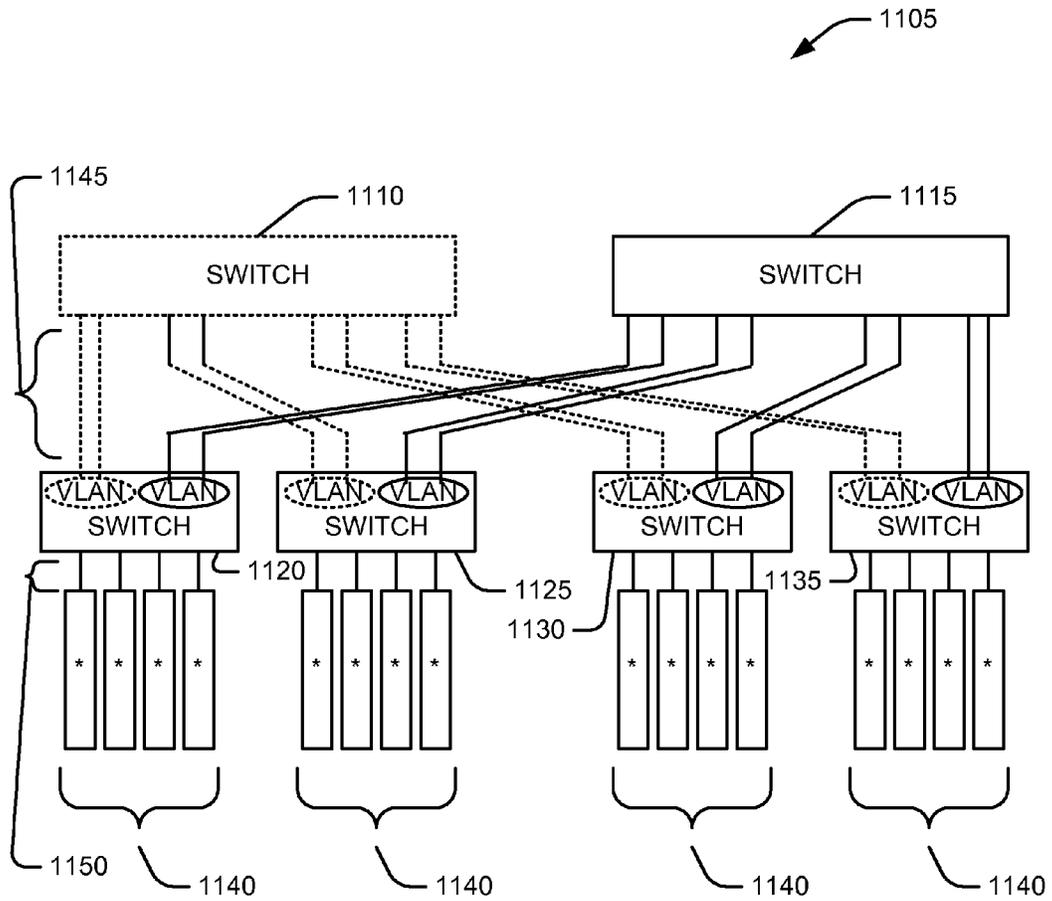
**FIG. 8**



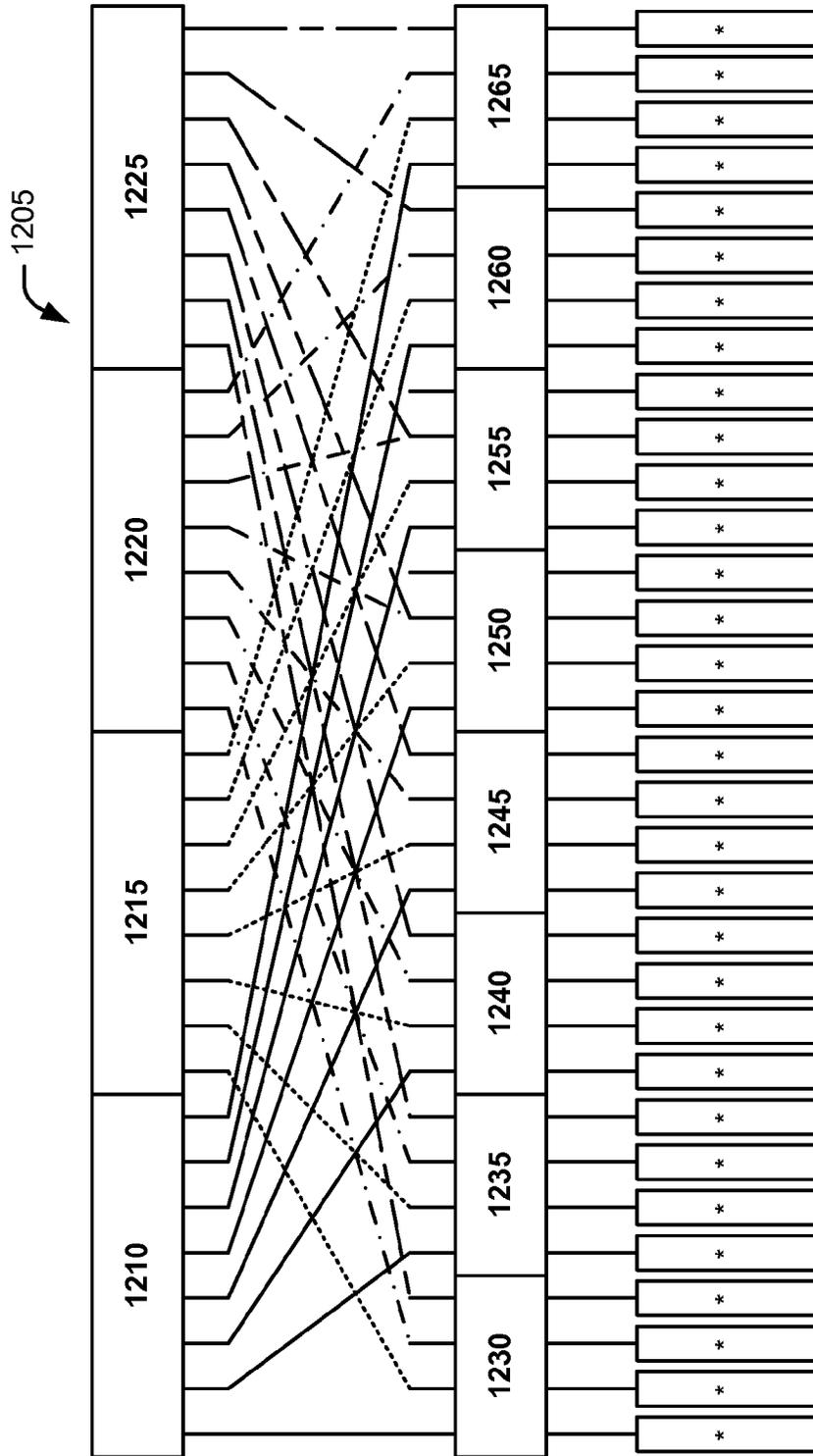
**FIG. 9**



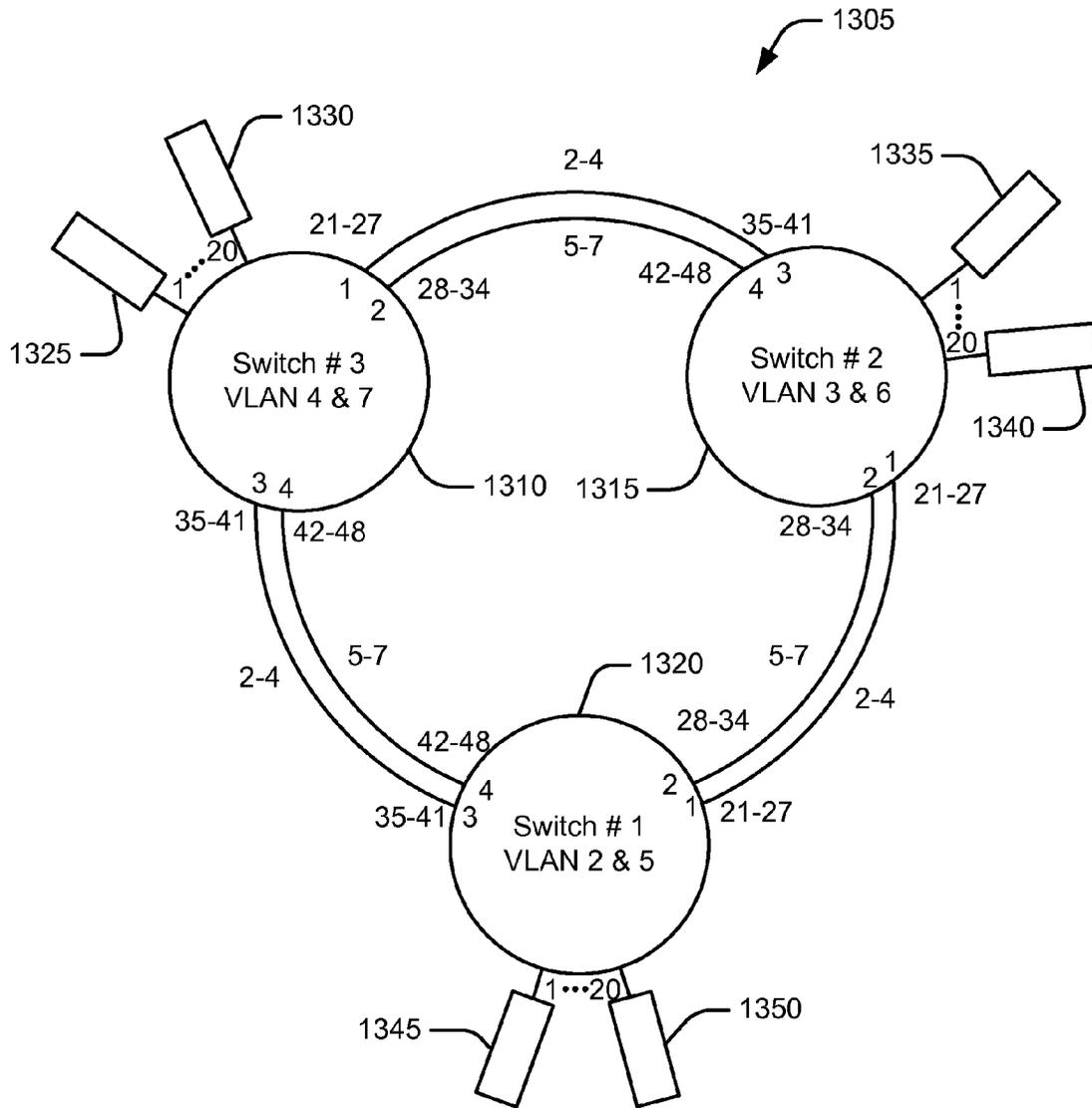
**FIG. 10**



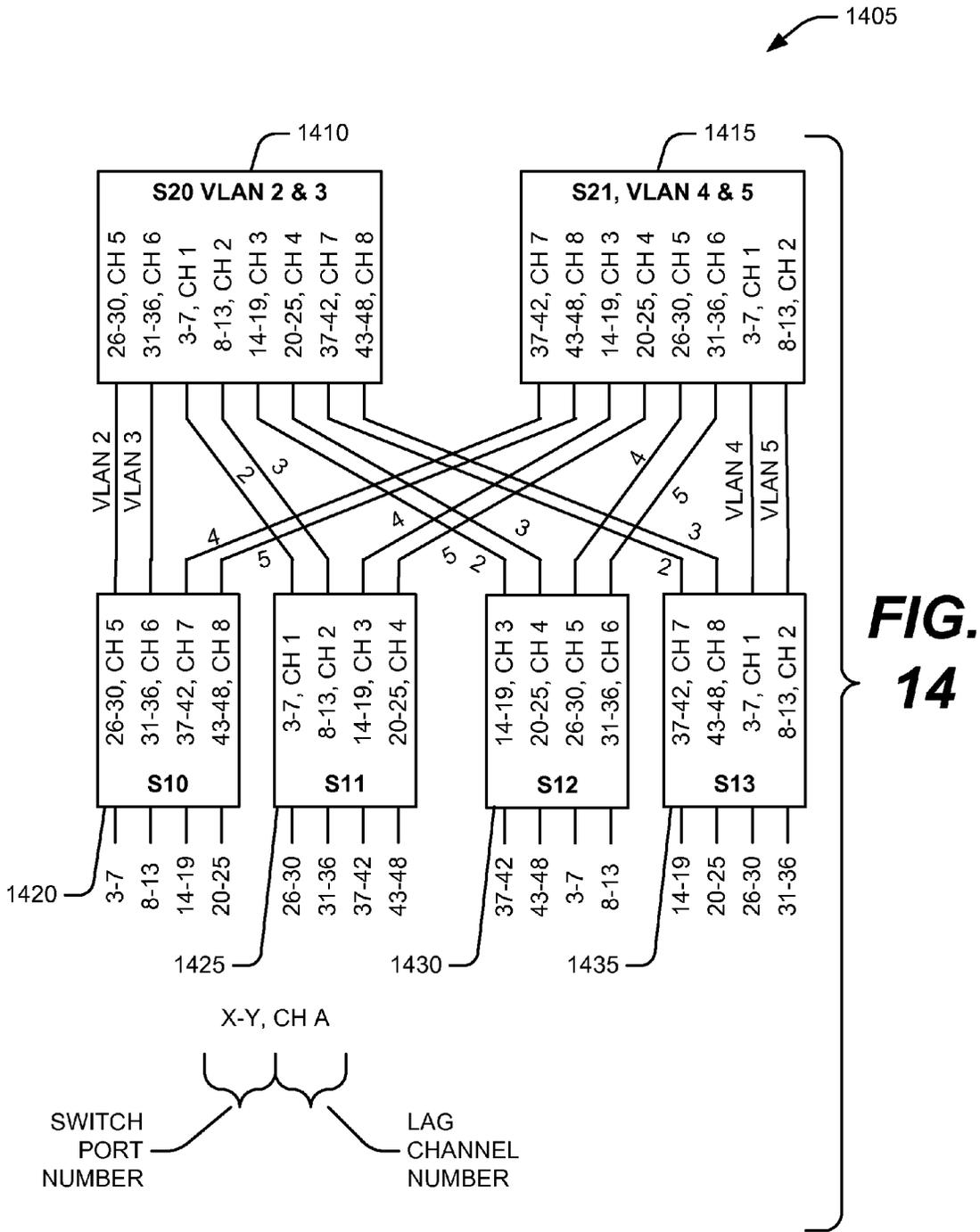
**FIG. 11**



**FIG. 12**



**FIG. 13**



1405

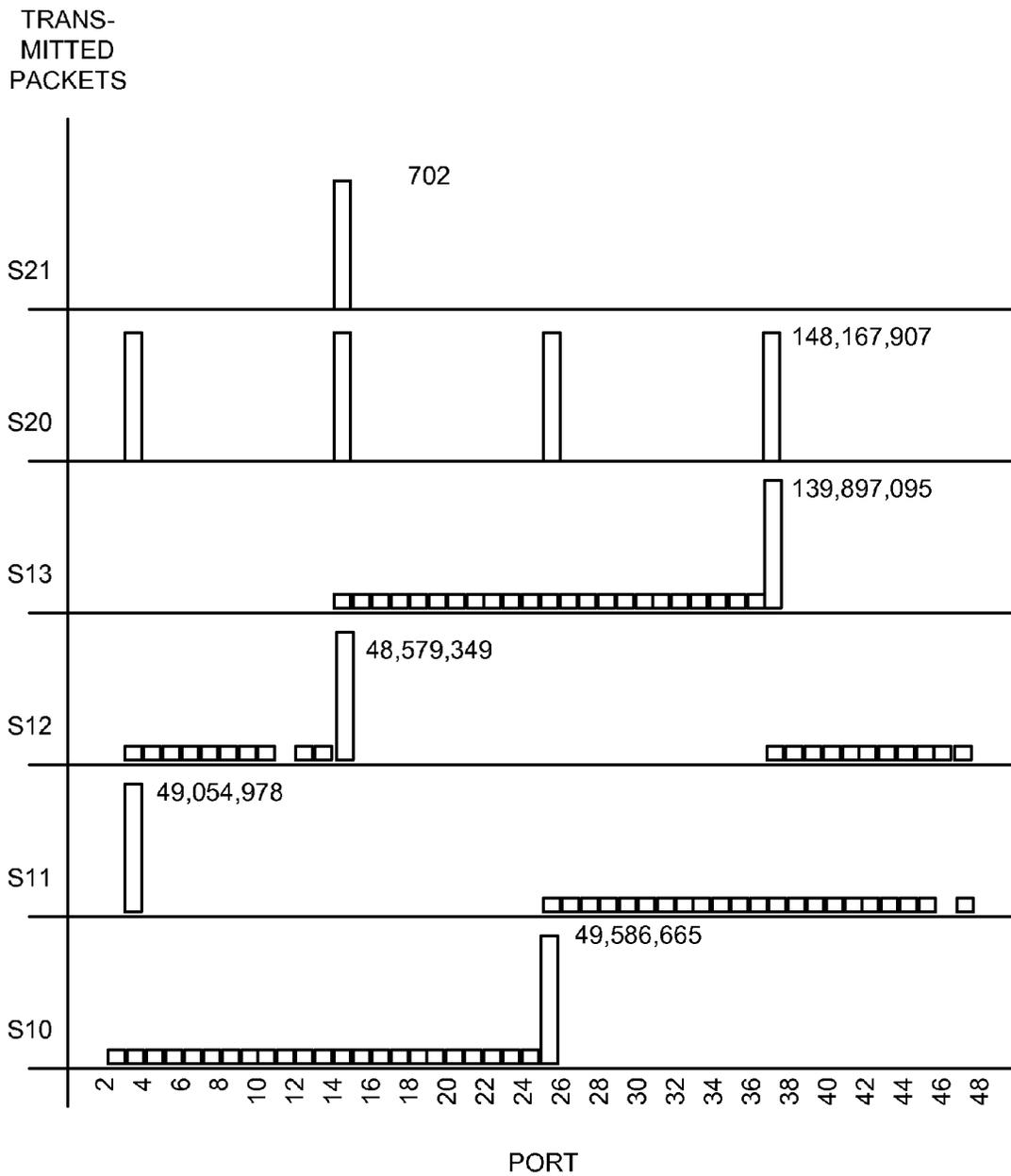


FIG. 15

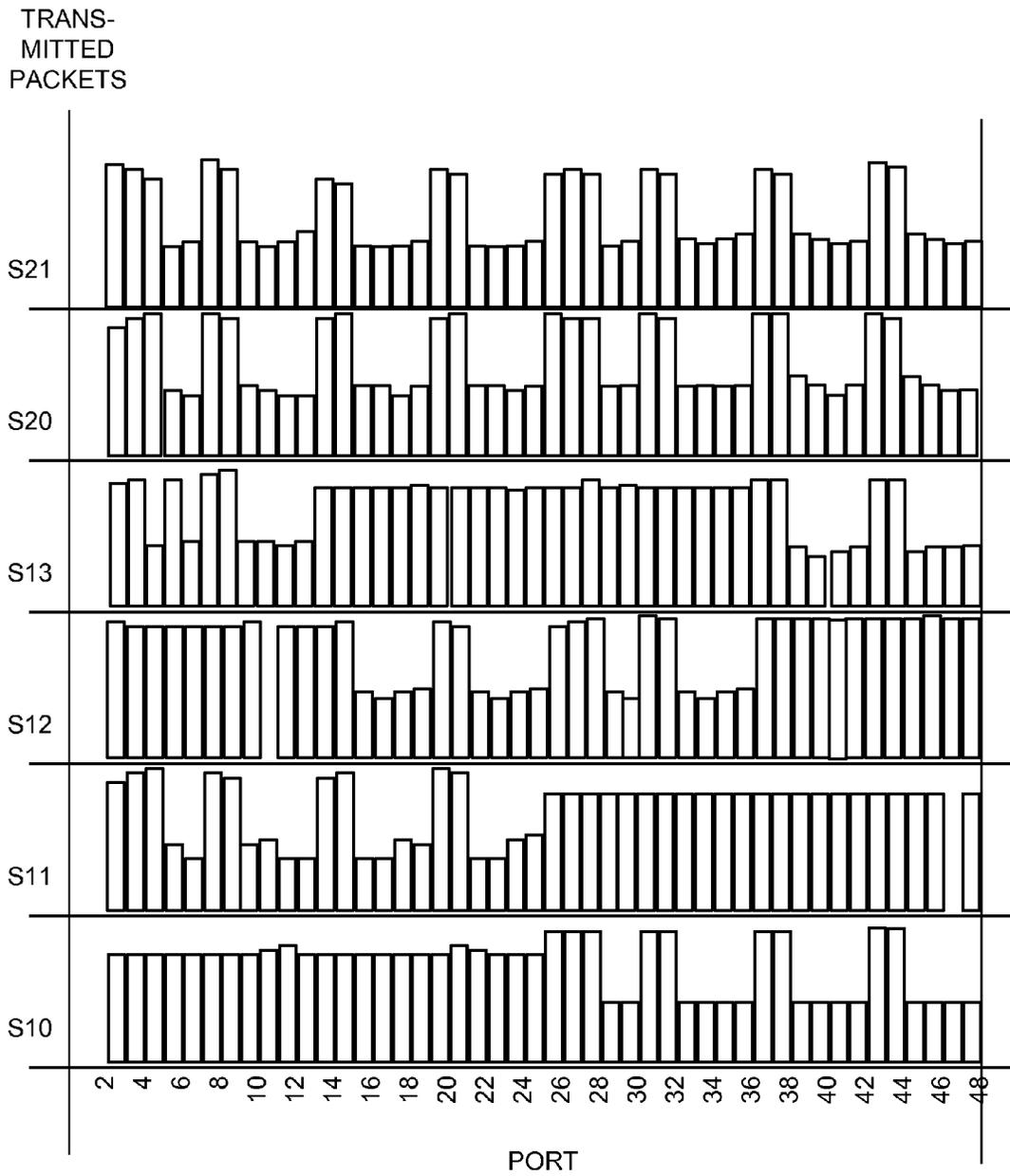
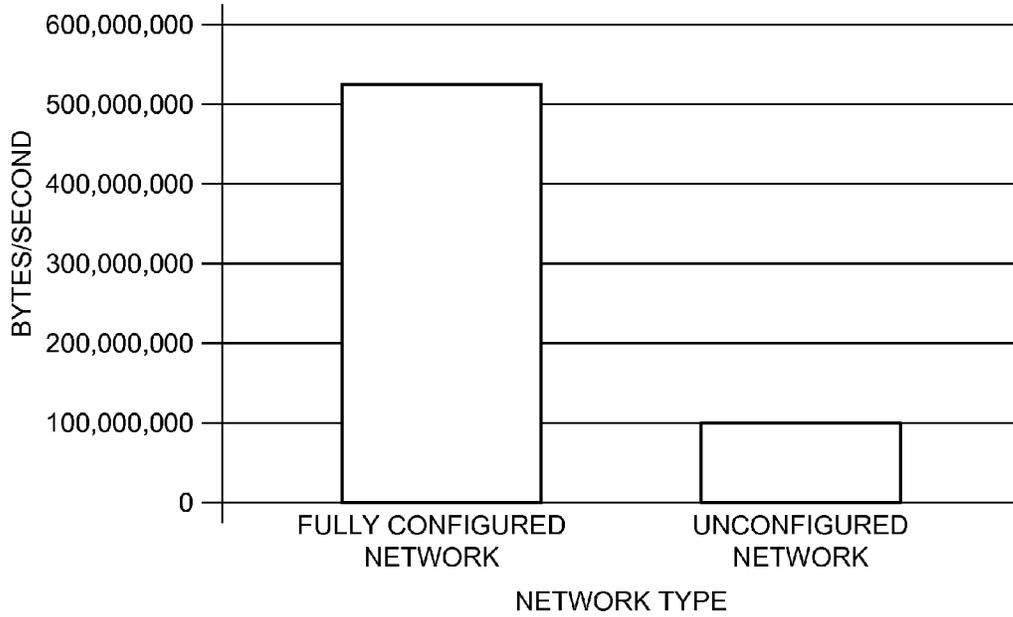
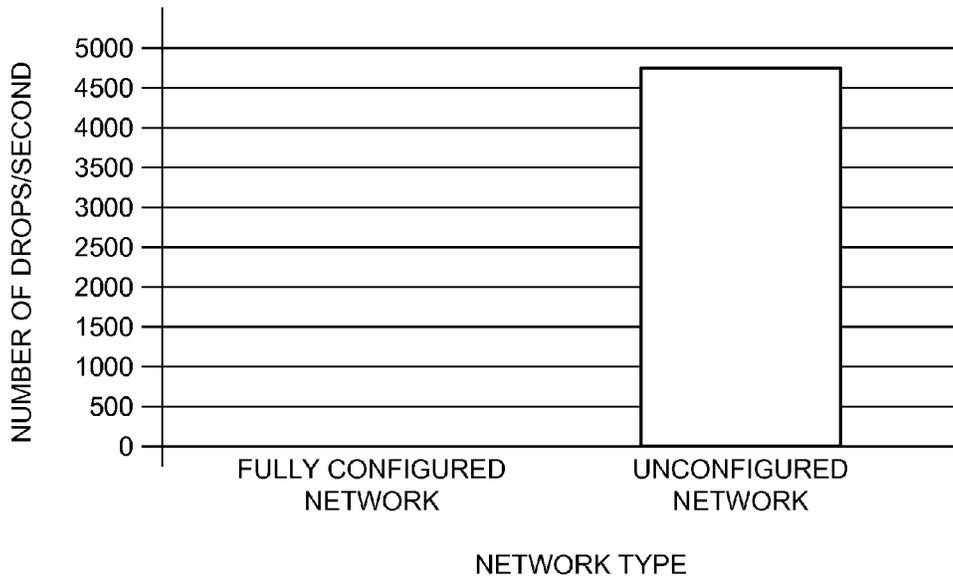


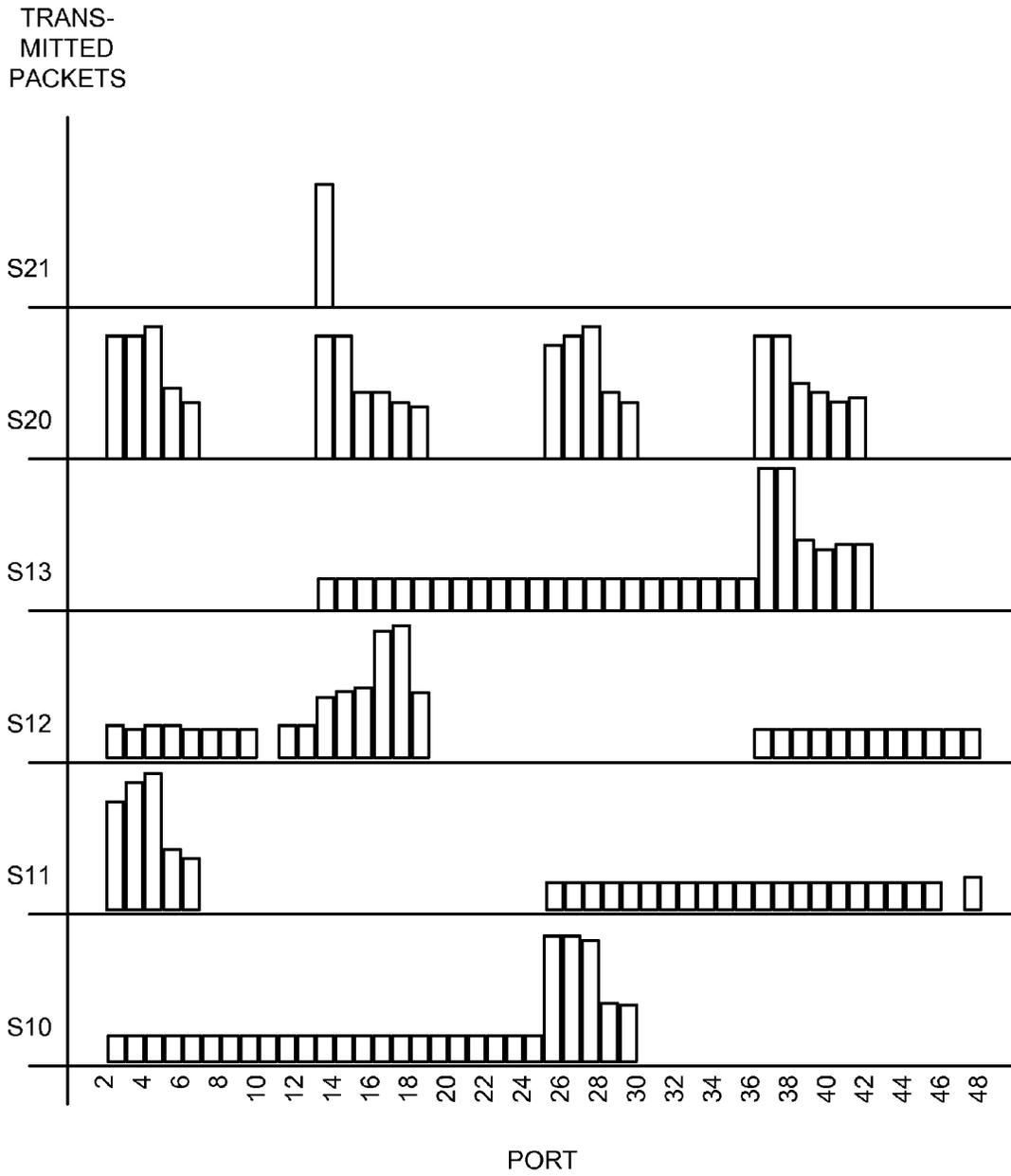
FIG. 16



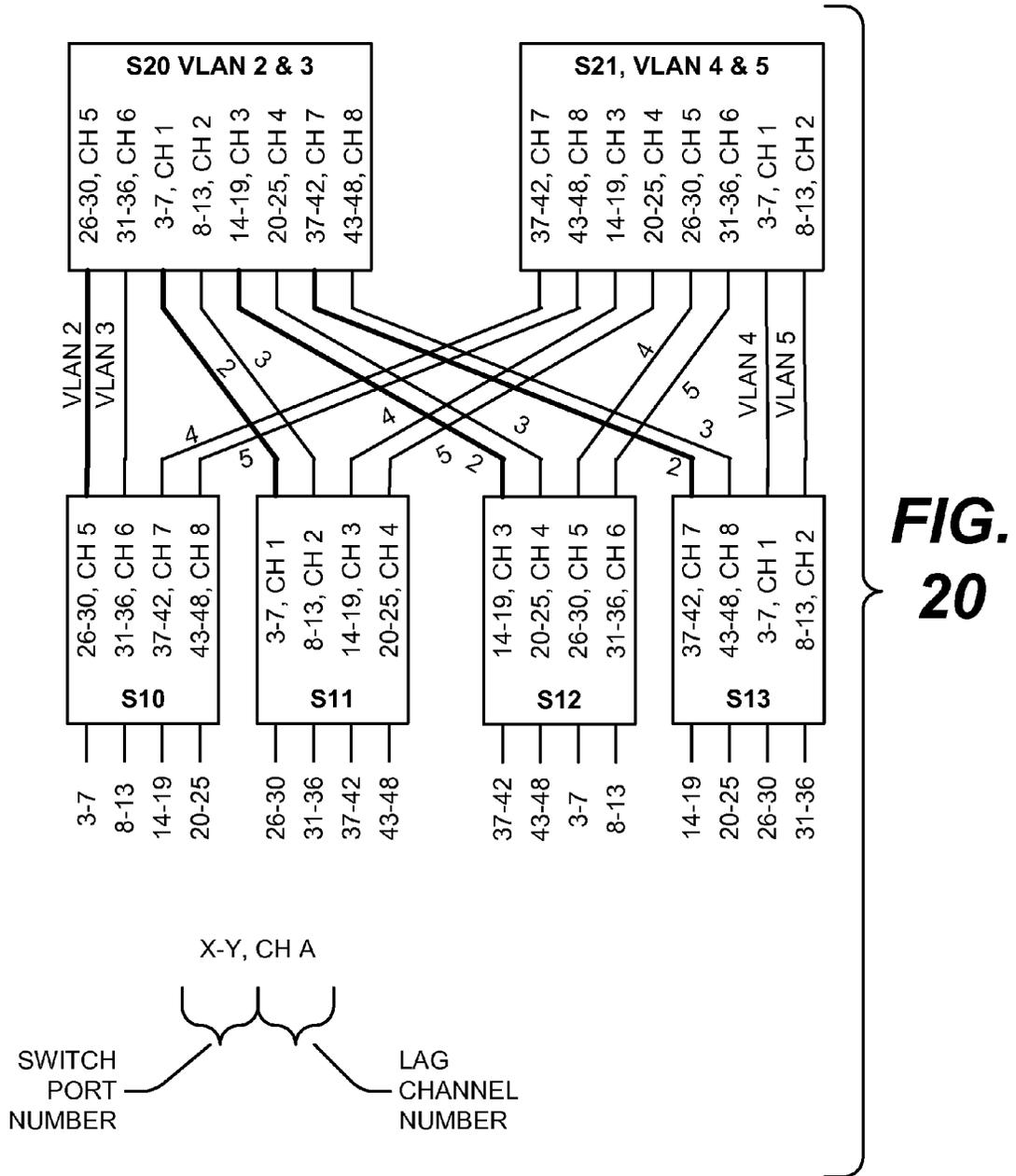
**FIG. 17**

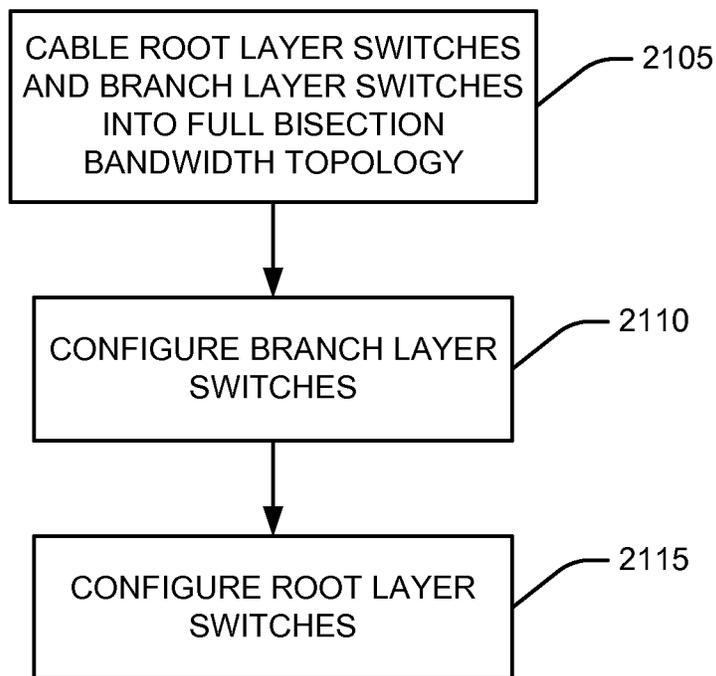


**FIG. 18**

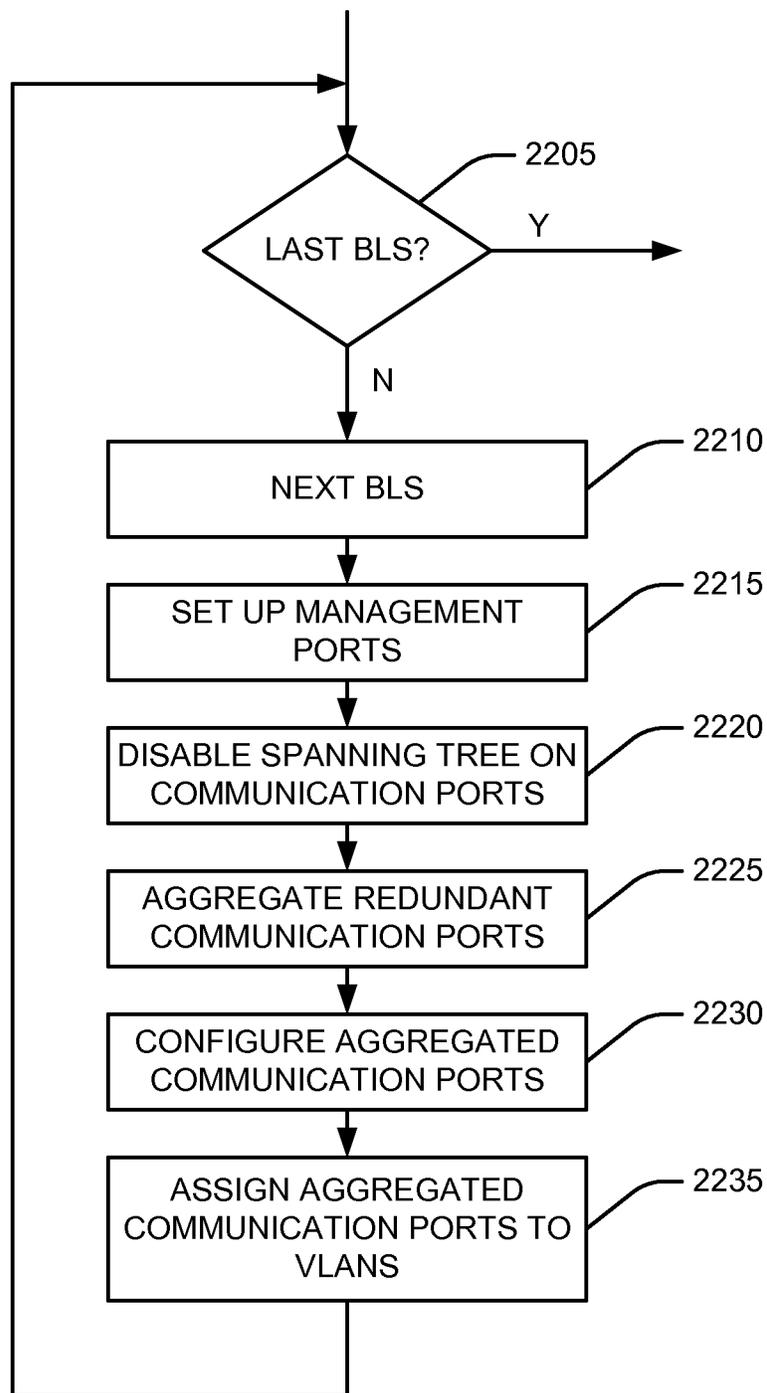


**FIG. 19**

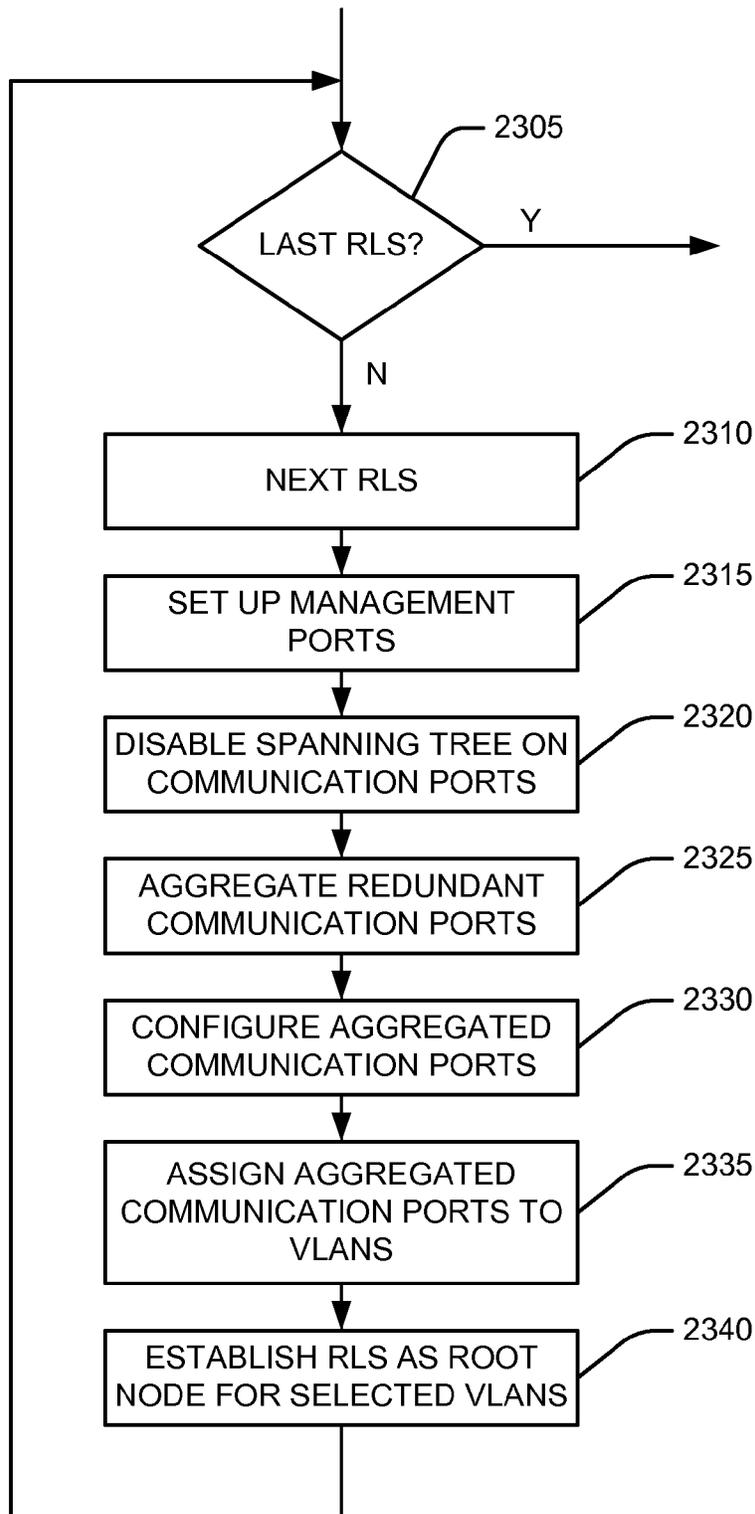




**FIG. 21**



**FIG. 22**



**FIG. 23**

**FULL BISECTION BANDWIDTH NETWORK**

## BACKGROUND

A device connected to a network, e.g., an Ethernet network, typically connects to a port on a network switch or hub. Network switches and hubs have a limited number of ports. Expanding the network to include a number of devices beyond the number of ports typically requires linking two or more switches or hubs. Redundant paths in the network are typically disabled by network protocol to prevent broadcast storms and loops in the topology. Making efficient use of such multiple-switch networks is a challenge.

## SUMMARY

In general, in one aspect, the invention features a method. A full bisection bandwidth network, having a plurality of nodes and a plurality of paths among the nodes, is divided into a plurality of Virtual Local Area Networks (“VLANs”) by assigning paths to the VLANs such that each VLAN satisfies a spanning tree protocol and all paths are active in at least one VLAN.

Implementations of the invention may include one or more of the following. The full bisection bandwidth network may include a path A connecting node X and node Y and a path B connecting node X and node Y, such that standard Ethernet protocol would treat path A and path B as redundant paths. The method may further include aggregating Path A and Path B into a single trunk group so that Path A and Path B are active. The method may further include constructing the full bisection bandwidth network to have a fat tree topology. The method may further include constructing the full bisection bandwidth network to have a fully connected mesh topology. The plurality of nodes may include a root layer of N Ethernet switches and a branch layer of M Ethernet switches. M may be greater than N. The plurality of paths may include a path from each root layer switch to each branch layer switch. Assigning paths to the VLANs may include assigning paths from a first root layer switch to a first set of VLANs, assigning paths from a second root layer switch to a second set of VLANs, the first set of VLANs not containing any VLANs belonging to the second set of VLANs, and the second set of VLANs not containing any VLANs belonging to the first set of VLANs. M may equal 2N. Assigning paths to the VLANs may include assigning a path from branch layer switch BLS1 to root layer switch RLS1 to a first VLAN and assigning a path from branch layer switch BLS1 to root layer switch RLS2 to a second VLAN. Assigning paths to the VLANs may include assigning a path from branch layer switch BLS1 to root layer switch RLS1 to a first VLAN and assigning a path from branch layer switch BLS2 to root layer switch RLS1 to a second VLAN. Assigning paths to the VLANs may include providing a first path PATH1 from a first branch layer switch BLS1 to a first root layer switch RLS1, providing a second path PATH2 from the first branch layer switch BLS1 to the first root layer switch RLS1, and aggregating PATH1 and PATH2 into a single trunk group. A plurality of servers may be coupled to the full bisection bandwidth network. The method further may further include providing redundant paths from one of the plurality of servers to another of the plurality of servers. The method may further include providing redundant paths from each of the plurality of servers to the others of the plurality of servers. Assigning paths to the VLANs may include assigning a first path from branch layer switch BLS1 to root layer switch RLS1 to a first VLAN and assigning a second path redundant to the first path to the first

VLAN. The plurality of paths may include a path from each root layer switch to each branch layer switch. A plurality of servers is coupled to the branch layer of Ethernet servers. Assigning paths to the VLANs may include assigning a first path from a first server to a second server and assigning a second path redundant to the first path from the first server to the second server. Assigning paths to the VLANs may include assigning redundant paths from each of the plurality of servers through the branch layer of Ethernet switches and the root layer of Ethernet switches to the others of the plurality of servers.

In general, in another aspect, the invention features a system. The system includes a full bisection bandwidth network. The full bisection bandwidth network includes a plurality of nodes. The full bisection bandwidth network includes a plurality of paths among the nodes. The full bisection bandwidth network includes a plurality of Virtual Local Area Networks (“VLANs”) incorporating the plurality of nodes and the plurality of paths. Each VLAN satisfies a spanning tree protocol. All paths are active in at least one VLAN.

Implementations of the invention include one or more of the following. The full bisection bandwidth network may include a path A connecting node X and node Y and a path B connecting node X and node Y, such that standard Ethernet protocol would treat path A and path B as redundant paths. Path A and Path B may be aggregated into a single trunk group such that Path A and Path B are active. The full bisection bandwidth network may have a fat tree topology. The full bisection bandwidth network may have a fully connected mesh topology. The plurality of nodes may include a root layer of N Ethernet switches. The plurality of nodes may include a branch layer of M Ethernet switches. M may be greater than N. The plurality of paths among the nodes may include a path from each root layer switch to each branch layer switch. Paths from a first root layer switch may be assigned to a first set of VLANs. Paths from a second root layer switch may be assigned to a second set of VLANs. The first set of VLANs may not contain any VLANs belonging to the second set of VLANs. The second set of VLANs may not contain any VLANs belonging to the first set of VLANs. M may equal N. The plurality of paths among the nodes may include a path from branch layer switch BLS1 to root layer switch RLS1 assigned to a first VLAN and a path from branch layer switch BLS1 to root layer switch RLS2 assigned to a second VLAN. The plurality of paths among the nodes may include a path from branch layer switch BLS1 to root layer switch RLS1 assigned to a first VLAN and a path from branch layer switch BLS2 to root layer switch RLS1 assigned to a second VLAN. The plurality of paths among the nodes may include a first path PATH1 from a first branch layer switch BLS1 to a first root layer switch RLS1 and a second path PATH2 from the first branch layer switch BLS1 to the first root layer switch RLS1 that are aggregated into a single trunk group. The system may further include a plurality of servers coupled to the full bisection bandwidth network. The plurality of paths among the nodes may include a plurality of redundant paths from one of the plurality of servers to another of the plurality of servers. The plurality of paths among the nodes may include a plurality of redundant paths from each of the plurality of servers to the others of the plurality of servers. The plurality of paths among the nodes may include a first path from branch layer switch BLS1 to root layer switch RLS1 assigned to a first VLAN and a second path redundant to the first path assigned to the first VLAN. The plurality of paths among the nodes may include a first path from a first server to a second server and a second path redundant to the first path from the first server to the second server. The plu-

ality of paths among the nodes may include redundant paths from each of the plurality of servers through the branch layer of Ethernet switches and the root layer of Ethernet switches to the others of the plurality of servers.

In general, in another aspect, the invention features a method. The method includes providing a full bisection bandwidth network, having a plurality of nodes and a plurality of paths among the nodes, that is divided into a plurality of Virtual Local Area Networks ("VLANs") by assigning paths to the VLANs such that each VLAN satisfies a spanning tree protocol and all paths are active in at least one VLAN. The full bisection bandwidth network carries a traffic load. The method includes balancing the traffic load among the paths.

In general, in another aspect, the invention features a method. The method includes providing a full bisection bandwidth network, having a plurality of nodes and a plurality of paths among the nodes, that is divided into a plurality of Virtual Local Area Networks ("VLANs") by assigning paths to the VLANs such that each VLAN satisfies a spanning tree protocol and all paths are active in at least one VLAN. The method further includes adding a node. The method further includes adding paths to connect the added node to the full bisection bandwidth network and adjusting the assignments of paths and the added paths to VLANs such that each VLAN satisfies a spanning tree protocol, all paths are active in at least one VLAN, and the network remains a full bisection bandwidth network.

Implementations of the invention may include one or more of the following. Adjusting the assignments may include adding a new VLAN. Adjusting the assignments may include adding the added paths to the existing VLANs.

#### BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a block diagram of a node of a parallel processing database system.

FIG. 2 is a block diagram of a parsing engine.

FIG. 3 is a block diagram of a parser.

FIG. 4 is an illustration of a full bisection bandwidth network.

FIGS. 5 and 6 are illustrations of the effects of the spanning tree protocol.

FIGS. 7-9 are illustrations of the effects of the elimination of redundant links and the application of link aggregation.

FIG. 10 is an illustration of a network using three 8-port switch elements to create a network of 12 ports.

FIG. 11 is an illustration of a network using six 8-port switch elements to create a network of 16 ports in a fat tree topology.

FIG. 12 is an illustration of a network using twelve 8-port switch elements to create a network of 32 ports.

FIG. 13 is an illustration of a fully connected mesh network.

FIG. 14 is an illustration of a fat tree network.

FIG. 15 is a graphical representation of the traffic activities of the ports in an unconfigured network.

FIG. 16 is a graphical representation of the traffic activities of the ports in a configured network.

FIG. 17 is a chart showing the improvement in throughput from an unconfigured network to a fully configured network.

FIG. 18 is a chart showing the improvement in number of drops per second from an unconfigured network to a fully configured network.

FIG. 19 is a chart showing the network activity when traffic is injected into a single Virtual Local Area Network ("VLAN").

FIG. 20 is a representation of a network illustrating traffic flowing in only a single VLAN.

FIGS. 21-23 are flow charts.

#### DETAILED DESCRIPTION

The full bisection bandwidth network technique disclosed herein has particular application, but is not limited, to large databases that might contain many millions or billions of records managed by a database system ("DBS") 100, such as a Teradata Active Data Warehousing System available from the assignee hereof. FIG. 1 shows a sample architecture for one subsystem 105<sub>1</sub> of the DBS 100. The DBS subsystem 105<sub>1</sub> includes one or more processing modules 110<sub>1...N</sub>, connected by a network 115, that manage the storage and retrieval of data in data-storage facilities 120<sub>1...N</sub>. Each of the processing modules 110<sub>1...N</sub> may be one or more physical processors or each may be a virtual processor, with one or more virtual processors running on one or more physical processors.

For the case in which one or more virtual processors are running on a single physical processor, the single physical processor swaps between the set of N virtual processors.

For the case in which N virtual processors are running on an M-processor subsystem, the subsystem's operating system schedules the N virtual processors to run on its set of M physical processors. If there are 4 virtual processors and 4 physical processors, then typically each virtual processor would run on its own physical processor. If there are 8 virtual processors and 4 physical processors, the operating system would schedule the 8 virtual processors against the 4 physical processors, in which case swapping of the virtual processors would occur.

Each of the processing modules 110<sub>1...N</sub> manages a portion of a database that is stored in a corresponding one of the data-storage facilities 120<sub>1...N</sub>. Each of the data-storage facilities 120<sub>1...N</sub> includes one or more disk drives. The DBS may include multiple subsystems 105<sub>2...N</sub> in addition to the illustrated subsystem 105<sub>1</sub>, connected by extending the network 115.

The system stores data in one or more tables in the data-storage facilities 120<sub>1...N</sub>. The rows 125<sub>1...Z</sub> of the tables are stored across multiple data-storage facilities 120<sub>1...N</sub> to ensure that the system workload is distributed evenly across the processing modules 110<sub>1...N</sub>. A parsing engine 130 organizes the storage of data and the distribution of table rows 125<sub>1...Z</sub> among the processing modules 110<sub>1...N</sub>. The parsing engine 130 also coordinates the retrieval of data from the data-storage facilities 120<sub>1...N</sub> in response to queries received from a user at a mainframe 135 or a client computer 140. The DBS 100 usually receives queries and commands to build tables in a standard format, such as SQL.

In one implementation, the rows 125<sub>1...Z</sub> are distributed across the data-storage facilities 120<sub>1...N</sub> by the parsing engine 130 in accordance with their primary index. The primary index defines the columns of the rows that are used for calculating a hash value. The function that produces the hash value from the values in the columns specified by the primary index is called the hash function. Some portion, possibly the entirety, of the hash value is designated a "hash bucket". The hash buckets are assigned to data-storage facilities 120<sub>1...N</sub> and associated processing modules 110<sub>1...N</sub> by a hash bucket map. The characteristics of the columns chosen for the primary index determine how evenly the rows are distributed.

In addition to the physical division of storage among the storage facilities illustrated in FIG. 1, each storage facility is also logically organized. One implementation divides the

storage facilities into logical blocks of storage space. Other implementations can divide the available storage space into different units of storage. The logical units of storage can ignore or match the physical divisions of the storage facilities.

In one example system, the parsing engine **130** is made up of three components: a session control **200**, a parser **205**, and a dispatcher **210**, as shown in FIG. 2. The session control **200** provides the logon and logoff function. It accepts a request for authorization to access the database, verifies it, and then either allows or disallows the access.

Once the session control **200** allows a session to begin, a user may submit a SQL query, which is routed to the parser **205**. As illustrated in FIG. 3, the parser **205** interprets the SQL query (block **300**), checks it for proper SQL syntax (block **305**), evaluates it semantically (block **310**), and consults a data dictionary to ensure that all of the objects specified in the SQL query actually exist and that the user has the authority to perform the request (block **315**). Finally, the parser **205** runs an optimizer (block **320**), which develops the least expensive plan to perform the request and produces executable steps to execute the plan. A dispatcher **210** issues commands to the processing modules  $110_1 \dots N$  to implement the executable steps.

The network **115** will continue to be described in the context of the system illustrated in FIG. 1 but it will be clear to persons of ordinary skill in the art that the network described herein is not limited to that context but can be used in any networking context.

In one embodiment, the network **115** includes a network **405**, such as that illustrated in FIG. 4. In one embodiment, the network **405** includes R switch elements, each having S ports, with each switch element having a connection to the other switch elements, leaving R(S-R+1) ports to which devices, such as the processing modules  $110_1 \dots N$  in FIG. 1, can connect. In the embodiment shown in FIG. 4, the network **405** includes six 8-port switch elements **410** (only one is labeled). Thus, R=6 and S=8, meaning that the resulting network will have  $6(8-6+1)=18$  ports. An end point device **415** ("device" or "end point": only one is labeled), such as one of the processing modules  $110_1 \dots N$  in FIG. 1, represented in FIG. 4 by an asterisk (\*), can be coupled to one of the ports **420** (only one is labeled). The network **405** illustrated in FIG. 4 can connect up to 18 devices.

The network **405** illustrated in FIG. 4 is a full bisection bandwidth network. In a full bisection bandwidth network, a device connected to one port on the network can communicate with another device connected to another port on the network at full speed, even when the network is fully populated and all ports are operating at full speed. In such a network, if every source end point wants to transmit to a different destination end point, and all source end points want to transmit at the same time, then a path exists for each one of the source end points to transmit. That is, there is no conflict or contention between source end points for paths. For example, if source end point A, shown in FIG. 4, wants to transmit to destination end point B, source end point B (which may be the same as destination end point B) wants to transmit to destination end point C, and source end point C (which may be the same as destination end point C) wants to transmit to destination end point A (which may be the same as source end point A) at the same time, the network provides paths for all of them. In a network that has the less than full bisection bandwidth, that may not be true.

A full bisection bandwidth network is realized when the network can be arbitrarily cut in half, such as by line **425**, and the number of cut links is equal to the number of end points in each half. In FIG. 4, the number of cross links cut by the line

**425** (9) is equal to the number of end points (i.e., the asterisks) (9) on either side of the line **425**. To achieve such the requirement of a full bisection bandwidth network typically requires that all available links in the network remain active.

In a typical Ethernet configuration, the network **405** illustrated in FIG. 4 would not be a full bisection bandwidth network because the Ethernet spanning tree protocol among switch elements disables redundant paths to avoid broadcast storms. By disabling redundant paths, however, the network loses valuable connections between end points.

FIG. 5 illustrates a typical scenario in which a redundant path is disabled by the spanning tree protocol. Before the spanning tree protocol is applied, nodes **405**, **410**, and **415** are connected by paths **420**, **425**, and **430**. In this configuration, multiple redundant paths are available between any two nodes. For example, nodes **405** and **410** are connected by (a) path **420**, and (b) paths **425** and **430** through node **415**. Under the typical implementation of Ethernet networks, such redundant paths create the possibility of loops and broadcast storms.

In the typical Ethernet network, one of the paths shown in FIG. 5 would be disabled under the spanning tree protocol. For example, after the spanning tree protocol is applied, path **430** is disabled, as indicated by the dashed line representing path **430** on the right side of FIG. 5, thereby eliminating redundant paths between nodes **405**, **410**, and **415**.

In one embodiment of a network **405**, the IEEE 802.1q protocol is applied to divide a network subject to the spanning tree protocol into VLANs in order to keep all network paths active and available for traffic. For example, as shown in FIG. 6, the network in FIG. 5 is configured to have three VLANs. The first VLAN, illustrated by tree **1** in FIG. 6, has path **430** disabled. The second VLAN, illustrated by tree **2** in FIG. 6, has path **425** disabled. The third VLAN, illustrated by tree **3** in FIG. 6, has path **420** disabled. Thus, as can be seen, in the configuration illustrated in FIG. 6, all three of the paths **420**, **425**, and **430** are active and each is active in two VLANs (e.g., path **420** is active in VLANs tree **1** and tree **2**).

In some embodiments, multiple point-to-point paths between nodes are used to achieve full bisection bandwidth. Typically, Ethernet protocol disables redundant links to prevent loops in the network. For example, consider the nodes **705** and **710** connected by redundant paths **715**, **720**, and **725** in FIG. 7. Typical Ethernet protocol will disable two of the paths (e.g., paths **715** and **720**), as shown in FIG. 8.

In one embodiment of a network **405**, the IEEE 802.2ad protocol is applied, as shown in FIG. 9, to aggregate multiple links into a single trunk group (represented by wrapper **905**), such that all paths remain active during normal operation.

FIG. 10 shows one embodiment of a full configuration of a network **1005** using 3 switch elements **1010**, **1015**, and **1020** of 8 ports each to expand from a single 8-port switch to a full bisection bandwidth network of 12 ports. As can be seen, the embodiment shown in FIG. 10 provides two paths **1025** and **1030** between switch element **1010** and switch element **1015**, two paths **1035** and **1040** between switch element **1010** and switch element **1020**, and two paths **1045** and **1050** between switch element **1015** and switch element **1020**. In addition, each switch element provides connections to four end points **1055** (such as, for example, the processing modules  $110_1 \dots N$  in FIG. 1)(only one is labeled). Both protocols 802.1q and 803.2ad are used in this example to achieve full bisection bandwidth. Each source end point has 2 paths available to reach any destination end point. For example, a source end point attached to switch element **1010** can reach a destination end point attached to switch element **1015** through path **1025** or through path **1030**. Further, if (a) network **1005** has been

divided into 3 VLANs, as described above in connection with the description of FIG. 6, (b) a source end point attached to switch element 1010 desires to communicate to a destination end point attached to switch element 1015, and (c) paths 1025 and 1030 are disabled in the VLAN being used for the communication, multiple paths are still available through paths 1035, 1040, 1045 and 1050. Alternatively, such communication could be accomplished via a VLAN in which paths 1025 and 1030 are enabled.

FIG. 11 shows one embodiment of a fully configured network with 6 switch elements of 8 ports each to expand from a single 8-port switch to a network of 16 ports in a full bisection bandwidth network having a fat tree topology. Two of the switch elements 1110 and 1115 are at a root layer of the fat tree topology. The other switches 1120, 1125, 1130, and 1135 are at a branch layer of the fat tree topology. The end points 1140 are at a leaf level of the fat tree topology. Thus, the network in FIG. 11 has the appearance of an inverted tree with the root at the top and the leaves at the bottom. The network illustrated in FIG. 11 is known as a "fat tree" because there are more paths between the root layer and the branch layer (those paths are labeled generally as 1145) than between the branch layer and the leaf layer (those branches are labeled generally as 1150). In the network 1105 shown in FIG. 11, there are 16 paths between the root layer and the branch layer and 16 paths between the branch layer and the leaf layer. Both protocols 802.1q and 803.2ad are used to achieve full bisection bandwidth. Each source end point has 2 paths to reach any destination end point. Any source end point can reach any destination end point through one of the dashed paths or through one of the solid paths. This multiplicity of paths provides the ability to load balance traffic across paths.

FIG. 12 shows a fat tree topology network 1205 with four 8-port switch elements 1210, 1215, 1220, and 1225 at the root layer and eight 8-port switch elements 1230, 1235, 1240, 1245, 1250, 1255, 1260, and 1265 at the branch layer, which produces a 32-port network. In this configuration, the Link Aggregation Protocol (803.2ad) is not needed to achieve full bisection bandwidth. There are four VLANs in the network, a first represented by the solid paths, a second represented by the dashed (i.e., "----") paths, a third represented by the dash-dot paths ("-•••"), and a fourth represented by the long-dash/short-dash ("- - - -") paths. Any source end point (represented by the asterisks) in the topology shown in FIG. 12 can reach any destination end point (also represented by the asterisks) through one of the four path sets.

FIG. 13 shows one embodiment of a fully connected mesh network 1305 using 3 Dell 6248 48-port Gigabit Ethernet switches 1310, 1315, 1320. The switches in the network are cabled together as shown:

- (a) ports 1-20 of switch 1310 are connected to end points 1325 and 1330;
- (b) ports 1-20 of switch 1315 are connected to end points 1335 and 1340;
- (c) ports 1-20 of switch 1320 are connected to end points 1345 and 1350;
- (d) ports 42-48 of switch 1310 are connected to ports 42-48 of switch 1320;
- (e) ports 35-41 of switch 1310 are connected to ports 35-41 of switch 1320;
- (g) ports 21-27 of switch 1310 are connected to ports 35-41 of switch 1315;
- (h) ports 28-34 of switch 1310 are connected to ports 42-48 of switch 1315;
- (i) ports 28-34 of switch 1315 are connected to ports 28-34 of switch 1320;

- (j) ports 21-27 of switch 1315 are connected to ports 21-27 of switch 1320;
- (k) switch 1310 is configured to be the root of VLANs 4 and 7;
- (l) switch 1315 is configured to be the root of VLANs 3 and 6; and
- (m) switch 1320 is configured to be the root of VLANs 2 and 5.

The switches are configured using Multiple Spanning Tree Protocol (802.1q) and Link Aggregation Protocol (803.2ad) to provide a 60-port network with full bisection bandwidth. All paths in the network are active at all times to carry traffic and there are 6 distinct paths for each source end point to inject traffic into the network to reach any destination end point.

FIG. 14 shows a fat tree network using six Dell 6248 48-port Gigabit Ethernet switches. The switches are cabled together as shown:

- (a) ports 1 and 2 of each switch 1410, 1415, 1420, 1425, 1430, 1435 are used for system management by a monitor element (not shown);
- (b) ports 26-30 of switch 1410 are aggregated together as trunk channel 5 and are connected to ports 26-30 of switch 1420, which are aggregated together as trunk channel 5;
- (c) ports 31-36 of switch 1410 are aggregated together as trunk channel 6 and are connected to ports 31-36 of switch 1420, which are aggregated together as trunk channel 6;
- (d) ports 3-7 of switch 1410 are aggregated together as trunk channel 1 and are connected to ports 3-7 of switch 1425, which are aggregated together as trunk channel 1;
- (e) ports 8-13 of switch 1410 are aggregated together as trunk channel 2 and are connected to ports 8-13 of switch 1425, which are aggregated together as trunk channel 2;
- (f) ports 14-19 of switch 1410 are aggregated together as trunk channel 3 and are connected to ports 14-19 of switch 1430, which are aggregated together as trunk channel 3;
- (g) ports 20-25 of switch 1410 are aggregated together as trunk channel 4 and are connected to ports 20-25 of switch 1430, which are aggregated together as trunk channel 4;
- (h) ports 37-42 of switch 1410 are aggregated together as trunk channel 7 and are connected to ports 37-42 of switch 1435, which are aggregated together as trunk channel 7;
- (i) ports 43-48 of switch 1410 are aggregated together as trunk channel 8 and are connected to ports 43-48 of switch 1435, which are aggregated together as trunk channel 8;
- (j) ports 37-42 of switch 1415 are aggregated together as trunk channel 7 and are connected to ports 37-42 of switch 1420, which are aggregated together as trunk channel 7;
- (k) ports 43-48 of switch 1415 are aggregated together as trunk channel 8 and are connected to ports 43-48 of switch 1420, which are aggregated together as trunk channel 8;
- (l) ports 14-19 of switch 1415 are aggregated together as trunk channel 3 and are connected to ports 14-19 of switch 1425, which are aggregated together as trunk channel 3;
- (m) ports 20-25 of switch 1415 are aggregated together as trunk channel 4 and are connected to ports 20-25 of switch 1425, which are aggregated together as trunk channel 4;

- (n) ports **26-30** of switch **1415** are aggregated together as trunk channel **5** and are connected to ports **26-30** of switch **1430**, which are aggregated together as trunk channel **5**;
- (o) ports **31-36** of switch **1415** are aggregated together as trunk channel **6** and are connected to ports **31-36** of switch **1430**, which are aggregated together as trunk channel **6**;
- (p) ports **3-7** of switch **1415** are aggregated together as trunk channel **1** and are connected to ports **3-7** of switch **1435**, which are aggregated together as trunk channel **1**;
- (q) ports **8-13** of switch **1415** are aggregated together as trunk channel **2** and are connected to ports **8-13** of switch **1435**, which are aggregated together as trunk channel **2**;
- (r) ports **3-25** of switch **1420** are available for connection to end points;
- (s) ports **26-48** of switch **1425** are available for connection to end points;
- (t) ports **3-13** and **37-48** of switch **1430** are available for connection to end points;
- (u) ports **14-36** of switch **1435** are available for connection to end points;
- (v) switch **1410** is configured to be the root of VLANs **2** and **3**;
- (w) switch **1415** is configured to be the root of VLANs **4** and **5**;
- (x) channels **1, 3, 5, and 7** on switch **1410**, channel **5** on switch **1420**, channel **1** on switch **1425**, channel **3** on switch **1430**, and channel **7** on switch **1435** are configured to be paths in VLAN **2**;
- (y) channels **2, 4, 6, and 8** on switch **1410**, channel **6** on switch **1420**, channel **2** on switch **1425**, channel **4** on switch **1430**, and channel **8** on switch **1435** are configured to be paths in VLAN **3**;
- (z) channels **1, 3, 5, and 7** on switch **1415**, channel **7** on switch **1420**, channel **3** on switch **1425**, channel **5** on switch **1430**, and channel **1** on switch **1435** are configured to be paths in VLAN **4**; and
- (aa) channels **2, 4, 6, and 8** on switch **1415**, channel **8** on switch **1420**, channel **4** on switch **1425**, channel **6** on switch **1430**, and channel **2** on switch **1435** are configured to be paths in VLAN **5**;

The network is configured using Multiple Spanning Tree Protocol (802.1q) and Link Aggregation Protocol (803.2ad) to provide a 92-port network with full section bandwidth. All links in the network are active at all time to carry traffic and there are 4 distinct paths for each source node to inject traffic into the network to reach any destination node. Two ports in each switch are dedicated for system management.

The scripts used to accomplish this configuration with the network **1405** shown in FIG. **14** is repeated below (using Dell script language; comments are in italics):

---

Switch element S10:

```

configure
vlan database
vlan 2-5           All VLANs are declared
exit
interface range ethernet 1/g1-1/g2
spanning-tree disable
spanning-tree portfast
exit
interface range ethernet 1/g3-1/g48
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit

```

-continued

---

```

interface range ethernet 1/g3-1/g25
spanning-tree disable
spanning-tree portfast
exit
5 interface range ethernet 1/g26-1/g30 Link aggregation setting
channel-group 5 mode auto
exit
interface range ethernet 1/g31-1/g36
channel-group 6 mode auto
10 exit
interface range ethernet 1/g37-1/g42
channel-group 7 mode auto
exit
interface range ethernet 1/g43-1/g48
channel-group 8 mode auto
15 exit
interface port-channel 5           Attach VLAN to LAG group
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
20 interface port-channel 6
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
25 interface port-channel 7
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
30 interface port-channel 8
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
35 spanning-tree mode mstp
spanning-tree mst configuration
instance 1 add vlan 1           Assign unique MSTP instance
instance 2 add vlan 2
instance 3 add vlan 3
instance 4 add vlan 4
40 instance 5 add vlan 5
exit
interface port-channel 5
spanning-tree mst 2 port-priority 0 Set up priority to guide routes
spanning-tree mst 3 port-priority 16
exit
45 interface port-channel 6
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
exit
interface port-channel 7
spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 16
50 exit
interface port-channel 8
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0
exit
spanning-tree mst configuration   Declare a single MSTP region
55 name "teradata"
exit
exit
Switch element S11:


---


60 configure
vlan database
vlan 2-5
exit
interface range ethernet 1/g1-1/g2
65 spanning-tree disable
spanning-tree portfast
exit

```

```

interface range ethernet 1/g3-1/g48
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface range ethernet 1/g26-1/g48
spanning-tree disable
spanning-tree portfast
exit
interface range ethernet 1/g3-1/g7
channel-group 1 mode auto
exit
interface range ethernet 1/g8-1/g13
channel-group 2 mode auto
exit
interface range ethernet 1/g14-1/g19
channel-group 3 mode auto
exit
interface range ethernet 1/g20-1/g25
channel-group 4 mode auto
exit
!
interface port-channel 1
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 2
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 3
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 4
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
spanning-tree mode mstp
spanning-tree mst configuration
instance 1 add vlan 1
instance 2 add vlan 2
instance 3 add vlan 3
instance 4 add vlan 4
instance 5 add vlan 5
exit
interface port-channel 1
spanning-tree mst 2 port-priority 0
spanning-tree mst 3 port-priority 16
exit
interface port-channel 2
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
exit
interface port-channel 3
spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 16
exit
interface port-channel 4
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0
exit
spanning-tree mst configuration
name "teradata"
exit
exit
exit
Switch element S12:

```

---

```

configure
vlan database

```

```

vlan 2-5
exit
5 interface range ethernet 1/g1-1/g2
spanning-tree disable
spanning-tree portfast
exit
interface range ethernet 1/g3-1/g48
switchport mode general
no switchport general acceptable-frame-type tagged-only
10 switchport general allowed vlan add 2-5 tagged
exit
interface range ethernet 1/g3-1/g13
spanning-tree disable
spanning-tree portfast
exit
15 interface range ethernet 1/g37-1/g48
spanning-tree disable
spanning-tree portfast
exit
interface range ethernet 1/g14-1/g19
channel-group 3 mode auto
exit
20 interface range ethernet 1/g20-1/g25
channel-group 4 mode auto
exit
interface range ethernet 1/g26-1/g30
channel-group 5 mode auto
exit
25 interface range ethernet 1/g31-1/g36
channel-group 6 mode auto
exit
interface port-channel 3
hashing-mode 5
switchport mode general
30 no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
!
interface port-channel 4
hashing-mode 5
35 switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 5
hashing-mode 5
40 switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 6
hashing-mode 5
switchport mode general
45 no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
spanning-tree mode mstp
spanning-tree mst configuration
instance 1 add vlan 1
50 instance 2 add vlan 2
instance 3 add vlan 3
instance 4 add vlan 4
instance 5 add vlan 5
exit
interface port-channel 3
55 spanning-tree mst 2 port-priority 0
spanning-tree mst 3 port-priority 16
exit
interface port-channel 4
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
60 exit
interface port-channel 5
spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 16
exit
65 interface port-channel 6
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0

```

```

exit
spanning-tree mst configuration
name "teradata"
exit
exit
exit
Switch element S13:
-----
configure
vlan database
vlan 2-5
exit
interface range ethernet 1/g1-1/g2
spanning-tree disable
spanning-tree portfast
exit
interface range ethernet 1/g3-1/g48
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface range ethernet 1/g14-1/g36
spanning-tree disable
spanning-tree portfast
exit
interface range ethernet 1/g37-1/g42
channel-group 7 mode auto
exit
interface range ethernet 1/g43-1/g48
channel-group 8 mode auto
exit
interface range ethernet 1/g3-1/g7
channel-group 1 mode auto
exit
interface range ethernet 1/g8-1/g13
channel-group 2 mode auto
exit
!
interface port-channel 7
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 8
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 1
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 2
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
spanning-tree mode mstp
spanning-tree mst configuration
instance 1 add vlan 1
instance 2 add vlan 2
instance 3 add vlan 3
instance 4 add vlan 4
instance 5 add vlan 5
exit
interface port-channel 7
spanning-tree mst 2 port-priority 0
spanning-tree mst 3 port-priority 16
exit
interface port-channel 8
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
exit
interface port-channel 1

```

```

spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 16
exit
5 interface port-channel 2
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0
exit
spanning-tree mst configuration
name "teradata"
10 exit
exit
exit
Switch element S20:
-----
configure
vlan database
15 vlan 2-5
exit
!
interface range ethernet 1/g1-1/g2
spanning-tree disable
spanning-tree portfast
20 exit
interface range ethernet 1/g3-1/g48
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
25 interface range ethernet 1/g3-1/g7 Link aggregation setting
channel-group 1 mode auto
exit
interface range ethernet 1/g8-1/g13
channel-group 2 mode auto
exit
30 interface range ethernet 1/g14-1/g19
channel-group 3 mode auto
exit
interface range ethernet 1/g20-1/g25
channel-group 4 mode auto
exit
35 interface range ethernet 1/g26-1/g30
channel-group 5 mode auto
exit
interface range ethernet 1/g31-1/g36
channel-group 6 mode auto
exit
40 interface range ethernet 1/g37-1/g42
channel-group 7 mode auto
exit
interface range ethernet 1/g43-1/g48
channel-group 8 mode auto
exit
!
45 interface port-channel 1 Attach VLAN to LAG group
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
50 !
interface port-channel 2
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
55 exit
interface port-channel 3
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
60 exit
interface port-channel 4
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
65 exit
interface port-channel 5

```

15

-continued

```

hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 6
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 7
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 8
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
spanning-tree mode mstp
spanning-tree mst configuration      Assign unique MSTP instance
instance 1 add vlan 1
instance 2 add vlan 2
instance 3 add vlan 3
instance 4 add vlan 4
instance 5 add vlan 5
exit
interface port-channel 1          Set up priority to guide routes
spanning-tree mst 2 port-priority 0
spanning-tree mst 3 port-priority 16
exit
interface port-channel 2
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
exit
interface port-channel 3
spanning-tree mst 2 port-priority 0
spanning-tree mst 3 port-priority 16
exit
interface port-channel 4
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
exit
interface port-channel 5
spanning-tree mst 2 port-priority 0
spanning-tree mst 3 port-priority 16
exit
interface port-channel 6
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
exit
interface port-channel 7
spanning-tree mst 2 port-priority 0
spanning-tree mst 3 port-priority 16
exit
interface port-channel 8
spanning-tree mst 2 port-priority 16
spanning-tree mst 3 port-priority 0
exit
spanning-tree mst 1 priority 0      MSTP instance priority for VLAN root
spanning-tree mst 2 priority 0
spanning-tree mst 3 priority 0
spanning-tree mst 4 priority 16384
spanning-tree mst 5 priority 16384
spanning-tree mst configuration    Assign MSTP region for network
name "teradata"
exit
exit
exit
Switch element S21:

```

---

```

configure
vlan database
vlan 2-5
exit

```

16

-continued

```

interface range ethernet 1/g1-1/g2
spanning-tree disable
spanning-tree portfast
5 exit
interface range ethernet 1/g3-1/g48
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
10 interface range ethernet 1/g3-1/g7
channel-group 1 mode auto
exit
interface range ethernet 1/g8-1/g13
channel-group 2 mode auto
exit
15 interface range ethernet 1/g14-1/g19
channel-group 3 mode auto
exit
interface range ethernet 1/g20-1/g25
channel-group 4 mode auto
exit
20 interface range ethernet 1/g26-1/g30
channel-group 5 mode auto
exit
interface range ethernet 1/g31-1/g36
channel-group 6 mode auto
exit
25 interface range ethernet 1/g37-1/g42
channel-group 7 mode auto
exit
interface range ethernet 1/g43-1/g48
channel-group 8 mode auto
exit
30 interface port-channel 1
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
!
35 interface port-channel 2
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
40 interface port-channel 3
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
45 interface port-channel 4
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
50 interface port-channel 5
hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 6
55 hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 7
60 hashing-mode 5
switchport mode general
no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
interface port-channel 8
65 hashing-mode 5
switchport mode general

```

-continued

```

no switchport general acceptable-frame-type tagged-only
switchport general allowed vlan add 2-5 tagged
exit
spanning-tree mode mstp
spanning-tree mst configuration
instance 1 add vlan 1
instance 2 add vlan 2
instance 3 add vlan 3
instance 4 add vlan 4
instance 5 add vlan 5
exit
interface port-channel 1
spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 16
exit
interface port-channel 2
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0
exit
interface port-channel 3
spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 20
exit
interface port-channel 4
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0
exit
interface port-channel 5
spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 25
exit
interface port-channel 6
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0
exit
interface port-channel 7
spanning-tree mst 4 port-priority 0
spanning-tree mst 5 port-priority 30
exit
interface port-channel 8
spanning-tree mst 4 port-priority 16
spanning-tree mst 5 port-priority 0
exit
spanning-tree mst 1 priority 16384
spanning-tree mst 2 priority 16384
spanning-tree mst 3 priority 16384
spanning-tree mst 4 priority 0
spanning-tree mst 5 priority 0
spanning-tree mst configuration
name "teradata"
exit
exit
exit
    
```

The network in FIG. 14 was used to collect relevant statistics for the invention. Fifteen server end points were used to emulate a 90-end-point fully connected, fully configured network. A diagnostic driver was designed to allow the ability to inject selectively one, many, or all VLAN traffic into the network. Using the diagnostic driver the network was characterized.

Tables 1 and 2 contain statistics collected when the network 1405 is not configured as described above. That is, the statistics shown in Table 1 show the number of bytes transmitted through one of the end points and the number of dropped packets when the network is configured with a single VLAN and is allowed to self-configure what it considers to be its best topology:

TABLE 1

Throughput	Drops
100,669,496	4,711
101,644,964	5,197

TABLE 1-continued

Throughput	Drops
96,521,132	4,561
93,187,456	4,557
97,508,640	4,812

FIG. 15 shows the graphical representation of the traffic activities of all ports in the network. The figure shows that uplink activities are essentially reduced to one path and only one switch of the root layer switches (S20) is actively carrying traffic.

Table 2 shows network statistics collected when the network is configured into a full bisection bandwidth topology, as described herein.

TABLE 2

Throughput	Drops
527,064,660	6
566,409,564	6
524,742,348	9
539,036,720	12
522,375,176	8

FIG. 16 shows the graphical representation of the traffic activities of all ports in the network. The figure shows that all paths in the network are actively carrying traffic and that all root level switches are actively carrying traffic.

FIGS. 17 and 18 show the difference in bandwidth and the number of packet drops per end point depending on whether the invention is used ("Fully configured network") or not ("Unconfigured network").

FIG. 19 shows that when only VLAN 2 traffic is injected into the network, only relevant ports that were configured for VLAN 2 carry the traffic. The point of this experiment is to show that there is a distinct path for any source to any destination for a particular VLAN. Traffic for each VLAN is completely isolated from all other traffic. Consequently, each source end point can decide how best to load balance traffic based on VLAN. For instance, one VLAN could be used for high priority traffic, one VLAN could be used as a broadcast only path, etc.

FIG. 20 depicts the same concept in a different way. In FIG. 20, the bold lines are paths that are dedicated to VLAN 2. Injecting packets in different VLANs improves isolation and increases network throughput by avoiding congestion.

In order to achieve a topology with full bisection bandwidth and predictable routes between sources and destinations, the network is cabled with strict connectivity for full section bandwidth, and each switch element is configured to achieve a complete network. Each switch element is configured to meet Multiple Spanning Tree Protocol (802.1q) ("MSTP") and Link Aggregation Protocol (803.2ad) ("LAG") requirements. The configuration allows the MSTP and LAG protocols to automatically produce the desired network. The following are the principle configuration settings used to meet MSTP and LAG requirements:

All switches in the network are declared with the same number of VLANs and allowed accessible by all declared VLAN traffic.

All links that connect between 2 switch elements have Link Aggregation groups declared for their corresponding connections. All LAG groups are attached to all of the VLANs tags.

Each VLAN is given a unique MSTP instance.

Each root level switch element is set up to make it a root switch for one MSTP instance.

All switch elements in the network are declared to be in a single MSTP region.

Priority is set up to guide the MSTP and traffic into the desired paths.

The resulting configuration:

allows expansion of Ethernet networks beyond one switch element without loss of full bisection bandwidth;

allows multiple low cost switches to be cascaded to achieve high port count networks at a lower cost than big iron networks;

provides multiple redundant paths in the network, resulting in greater network resiliency in the face of link failures (this can be further enhanced by connecting more than one port per end point to more than one switch element) (redundancy further helps prevent loss of connectivity when switch elements fail);

can be incrementally grown as needed;

allows load balancing through multiple paths to reduce congestion and packet loss.

As illustrated in FIG. 21, network configuration of a fat tree network according to the principles described herein begins by cabling root layer switches and branch layer switches into a full bisection bandwidth topology (block 2105), such as that shown in FIG. 14. The branch layer switches are configured (block 2110) and the root layer switches are configured (block 2115).

Configuration of the branch layer switches, as illustrated in FIG. 22, begins by determining if the last branch layer switch has been configured (block 2205). If it has ("Y" branch out of block 2205), then branch layer configuration is complete. If it has not ("N" branch out of block 2205), then the next branch layer switch is configured (block 2210). The management ports are set up (block 2215). The spanning tree protocol is disabled on the communication ports (i.e., the non-management ports) that are or can be connected to end points (i.e., not ports used to connect two switches) (block 2220). Redundant communication ports are aggregated (block 2225). The aggregated communication ports are configured (block 2230). Aggregated communication ports are assigned to VLANs (block 2235).

Configuration of the root layer switches, as illustrated in FIG. 23, begins by determining if the last root layer switch has been configured (block 2305). If it has ("Y" branch out of block 2305), then root layer configuration is complete. If it has not ("N" branch out of block 2305), then the next root layer switch is configured (block 2310). The management ports are set up (block 2315). Spanning tree is disabled on the communication ports (i.e., the non-management ports) that are or can be connected to end points (i.e., not ports used to connect two switches) (block 2320). Redundant communication ports are aggregated (block 2325). The aggregated communication ports are configured (block 2330). Aggregated communication ports are assigned to VLANs (block 2335). Root level switches are established as root nodes for selected VLANs (block 2340).

The foregoing description of the preferred embodiment of the invention has been presented for the purposes of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. It is intended that the scope of the invention be limited not by this detailed description, but rather by the claims appended hereto.

What is claimed is:

1. A system comprising:

a full bisection bandwidth network comprising:

a plurality of nodes;

a plurality of paths among the nodes;

a plurality of Virtual Local Area Networks ("VLANs") incorporating the plurality of nodes and the plurality of paths, wherein the plurality of VLANs comprises a first VLAN and a second VLAN and the plurality of paths comprises a first path, wherein the first path is between two nodes and does not pass through any intervening nodes;

wherein the first path is assigned to the first VLAN and the second VLAN;

wherein the first VLAN would not satisfy a spanning tree protocol if the first path is enabled in the first VLAN and the first VLAN satisfies the spanning tree protocol with the first path disabled in the first VLAN; and

wherein the first path is enabled in the second VLAN and the second VLAN satisfies the spanning tree protocol with the first path enabled in the second VLAN;

adding a node;

adding paths to connect the added node to the full bisection bandwidth network;

adjusting the assignments of the paths and the added paths to VLANs such that:

each VLAN satisfies a spanning tree protocol;

each of the plurality of paths is active in the full bisection bandwidth network; and

the network remains a full bisection bandwidth network;

wherein:

the plurality of nodes comprises:

a root layer of N Ethernet switches;

a branch layer of M Ethernet switches,  $M > N$ ;

the plurality of paths among the nodes comprises:

a path from branch layer switch BLS1 to root layer switch RLS1 assigned to a first VLAN; and

a path from branch layer switch BLS2 to root layer switch RLS1 assigned to a second VLAN.

2. The system of claim 1 wherein:

the full bisection bandwidth network comprises a path A connecting node X and node Y and a path B connecting node X and node Y, such that standard Ethernet protocol would treat path A and path B as redundant paths; and Path A and Path B are aggregated into a single trunk group such that Path A and Path B are active.

3. The system of claim 1 wherein:

the full bisection bandwidth network has a fat tree topology.

4. The system of claim 1 further comprising:

the full bisection bandwidth network has a fully connected mesh topology.

5. The system of claim 1 wherein:

the plurality of paths among the nodes comprises:

a path from each root layer switch to each branch layer switch;

paths from a first root layer switch are assigned to a first set of VLANs;

paths from a second root layer switch are assigned to a second set of VLANs;

the first set of VLANs does not contain any VLANs belonging to the second set of VLANs; and

the second set of VLANs does not contain any VLANs belonging to the first set of VLANs.

## 21

6. The system of claim 5 wherein  $M=2N$ .

7. The system of claim 1 wherein:  
the path from branch layer switch BLS1 to root layer switch RLS1 and a second path from branch layer switch BLS1 to root layer switch RLS1 are aggregated into a single trunk group. 5

8. The system of claim 1 further comprising:  
a plurality of servers coupled to the full bisection bandwidth network; and  
the plurality of paths among the nodes comprises a plurality of redundant paths from one of the plurality of servers to another of the plurality of servers. 10

9. The system of claim 1 further comprising:  
a plurality of servers coupled to the full bisection bandwidth network; and 15  
the plurality of paths among the nodes comprises a plurality of redundant paths from each of the plurality of servers to the others of the plurality of servers.

10. The system of claim 1 wherein:  
the plurality of paths among the nodes comprises: 20  
a second path redundant to the path from branch layer switch BLS1 to root layer switch RLS1 assigned to the first VLAN.

11. The system of claim 1 wherein:  
a plurality of servers is coupled to the branch layer of Ethernet servers; 25  
the plurality of paths among the nodes comprises:  
a first path from a first server to a second server; and  
a second path redundant to the first path from the first server to the second server. 30

12. The system of claim 1 wherein:  
a plurality of servers coupled to the branch layer of Ethernet servers;  
the plurality of paths among the nodes comprises:  
redundant paths from each of the plurality of servers 35  
through the branch layer of Ethernet switches and the root layer of Ethernet switches to the others of the plurality of servers.

13. A method comprising:  
providing a full bisection bandwidth network, having a 40  
plurality of nodes and a plurality of paths among the nodes, that is divided into a plurality of Virtual Local Area Networks ("VLANs"), wherein the plurality of VLANs comprises a first VLAN and a second VLAN and the plurality of paths comprises a first path, wherein 45  
the first path is between two nodes and does not pass through any intervening nodes, by assigning the first path to the first VLAN and to the second VLAN;  
disabling the first path in the first VLAN, wherein the first VLAN with the first path enabled in the first VLAN 50  
would not satisfy a spanning tree protocol and the first VLAN with the first path disabled in the first VLAN satisfies the spanning tree protocol;  
enabling the first path in the second VLAN, wherein the second VLAN satisfies the spanning tree protocol with 55  
the first path enabled in the second VLAN;  
the full bisection bandwidth network carrying a traffic load;  
balancing the traffic load among the paths;  
adding a node; 60  
adding paths to connect the added node to the full bisection bandwidth network;  
adjusting the assignments of the paths and the added paths to VLANs such that:  
each VLAN satisfies a spanning tree protocol; 65  
each of the plurality of paths is active in the full bisection bandwidth network; and

## 22

the network remains a full bisection bandwidth network;  
wherein the plurality of nodes comprises a root layer of N Ethernet switches and a branch layer of M Ethernet switches,  $M>N$ , the plurality of paths comprises a path from each root layer switch to each branch layer switch, and wherein assigning the first path to the first VLAN and to the second VLAN comprises:  
assigning a path from branch layer switch BLS1 to root layer switch RLS1 to the first VLAN; and  
assigning a path from branch layer switch BLS2 to root layer switch RLS1 to the second VLAN.

14. A method comprising:  
providing a full bisection bandwidth network, having a plurality of nodes and a plurality of paths among the nodes, that is divided into a plurality of Virtual Local Area Networks ("VLANs"), wherein the plurality of VLANs comprises a first VLAN and a second VLAN and the plurality of paths comprises a first path, wherein the first path is between two nodes and does not pass through any intervening nodes, by assigning the first path to the first VLAN and to the second VLAN;  
disabling the first path in the first VLAN, wherein the first VLAN with the first path enabled in the first VLAN would not satisfy a spanning tree protocol and the first VLAN with the first path disabled in the first VLAN satisfies the spanning tree protocol; and  
enabling the first path in the second VLAN, wherein the second VLAN satisfies the spanning tree protocol with the first path enabled in the second VLAN;  
adding a node;  
adding added paths to connect the added node to the full bisection bandwidth network;  
adjusting the assignments of the paths and the added paths to VLANs such that:  
each VLAN satisfies a spanning tree protocol;  
each of the plurality of paths is active in the full bisection bandwidth network; and  
the network remains a full bisection bandwidth network;  
wherein the plurality of nodes comprises a root layer of N Ethernet switches and a branch layer of M Ethernet switches,  $M>N$ , the plurality of paths comprises a path from each root layer switch to each branch layer switch, and wherein assigning the first path to the first VLAN and to the second VLAN comprises:  
assigning a path from branch layer switch BLS1 to root layer switch RLS1 to the first VLAN; and  
assigning a path from branch layer switch BLS2 to root layer switch RLS1 to the second VLAN.

15. A method of claim 14 wherein adjusting the assignments comprises:  
adding a new VLAN.

16. A method of claim 14 wherein adjusting the assignments comprises:  
adding the added paths to the existing VLANs.

17. The method of claim 14 wherein the full bisection bandwidth network comprises a path A connecting node X and node Y and a path B connecting node X and node Y, such that standard Ethernet protocol would treat path A and path B as redundant paths, the method further comprising:  
aggregating Path A and Path B into a single trunk group so that Path A and Path B are active.

18. The method of claim 14 further comprising:  
constructing the full bisection bandwidth network to have a fat tree topology.

**23**

- 19. The method of claim **14** further comprising:  
constructing the full bisection bandwidth network to have  
a fully connected mesh topology.
- 20. The method of claim **14** further comprising:  
assigning paths from a first root layer switch to a first set of  
VLANs;  
assigning paths from a second root layer switch to a second  
set of VLANs;  
the first set of VLANs not containing any VLANs belong-  
ing to the second set of VLANs; and  
the second set of VLANs not containing any VLANs  
belonging to the first set of VLANs.
- 21. The method of claim **20** wherein  $M=2N$ .
- 22. The method of claim **14** further comprising:  
aggregating the path from the first branch layer switch  
BLS1 to a first root layer switch RLS1 with the path from  
the first branch layer switch BLS1 to the first root layer  
switch RLS1 into a single trunk group.
- 23. The method of claim **14** wherein a plurality of servers  
is coupled to the full bisection bandwidth network and the  
method further comprises:

**24**

- providing redundant paths from one of the plurality of  
servers to another of the plurality of servers.
- 24. The method of claim **14** wherein a plurality of servers  
is coupled to the full bisection bandwidth network and the  
method further comprises:  
providing redundant paths from each of the plurality of  
servers to the others of the plurality of servers.
- 25. The method of claim **14** further comprising:  
assigning a path redundant to the path from branch layer  
switch BLS1 to root layer switch RLS1 to the first  
VLAN.
- 26. The method of claim **14** further comprising:  
assigning a first path from a first server to a second server;  
and  
assigning a second path redundant to the first path from the  
first server to the second server.
- 27. The method of claim **14** further comprising:  
assigning redundant paths from each of the plurality of  
servers through the branch layer of Ethernet switches  
and the root layer of Ethernet switches to the others of  
the plurality of servers.

\* \* \* \* \*