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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 August 2006 (10.08.2006)

PCT

(10) International Publication Number
WO 2006/084071 A2

(51) International Patent Classification:
H04L 12/28 (2006.01)

(21) International Application Number:
PCT/US2006/003740

(22) International Filing Date: 3 February 2006 (03.02.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/650,312 4 February 2005 (04.02.2005) US

(71) Applicant *(for all designated States except US)*: **LEVEL 3 COMMUNICATIONS, INC.** [US/US]; 1025 Eldorado Boulevard, Broomfield, Colorado 80021 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **LAWRENCE, Joseph** [US/US]; 2989 Tin Cup, Boulder, Colorado 80305 (US). **EL-AAWAR, Nassar** [US/US]; 1 466 South Downing, Denver, Colorado 80209 (US). **LOHER, Darren, P.** [US/US]; 6554 Orion Court, Arvada, Colorado 80007 (US). **WHITE, Steven, Craig** [US/US]; 3711 Florentine Drive, Longmont, Colorado 80503 (US). **ALCALA, Raoul** [US/US]; 986 Monroe Way, Superior, Colorado 80027 (US).

(74) Agents: **LINDER, Walter, C.** et al.; 2200 Wells Fargo Center, 90 South Seventh Street, Minneapolis, Minnesota 55402-3901 (US).

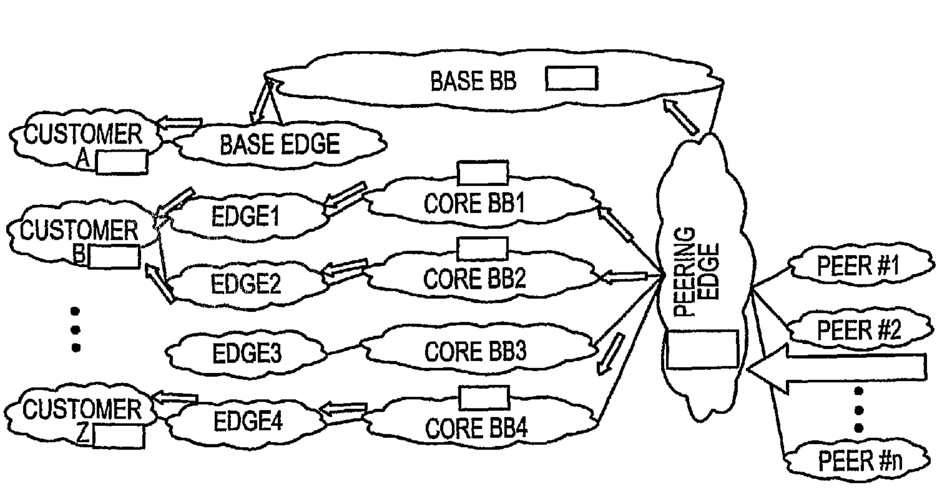
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LC,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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (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:** ETHERNET-BASED SYSTEMS AND METHODS FOR IMPROVED NETWORK ROUTING



(57) **Abstract:** Ethernet-based networks for routing Internet Protocol (IP) traffic between source and destination sites. One embodiment includes a plurality of discrete data transmission backbones between the source and destination sites. The source site includes control means for distributing IP traffic at the source site to the plurality of backbones for transmission to the destination site.

ETHERNET-BASED SYSTEMS AND METHODS FOR IMPROVED NETWORK ROUTING

Reference to Related Application

[0001] This application claims the benefit of U.S. Provisional Application Serial. No. 60/650,312, filed February 4, 2005, and entitled Systems And Methods For Improved Network Routing, which is incorporated herein in its entirety.

Field of the Invention

[0002] The present invention relates generally to network routing, and more specifically to Ethernet-based systems and methods for routing IP traffic at the edges and in the core backbone of an IP (Internet Protocol) network.

Background of the Invention

[0003] High speed internet prices continue to drop, but the underlying costs of maintaining and operating the networks remain relatively high. One of the main factors in keeping the unit costs high is the high cost for the terabit MPLS backbone routers. Accordingly, as bandwidth requirements grow, the costs will likely grow as well. Thus, a need exists for ways to scale network architectures larger (*i.e.*, higher bandwidth capacity) in a more cost effective manner.

Brief Description of the Drawings

[0004] Fig. 1 is a diagrammatic illustration of a three-stage multichassis Ethernet router (MER) in accordance with one embodiment of the invention.

[0005] Fig. 2 is a diagrammatic illustration of multiple parallel backbones (N x BB) connected to peer and edge networks in accordance with another embodiment of the invention.

[0006] Fig. 3 is a diagrammatic illustration of a combination of the multichassis Ethernet router shown in Fig. 1 and the multiple parallel backbones shown in Fig. 2 connected between sites in accordance with another embodiment of the invention.

[0007] Fig. 4 is a diagrammatic illustration of a multichassis Ethernet router-based core in parallel with existing MPLS cores between sites in accordance with another embodiment of the invention.

[0008] Fig. 5 is a diagrammatic illustration of an alternative version of the invention shown in Fig. 4.

[0009] Fig. 6 is a diagrammatic illustration of multiple core local area networks connected in the middle of core routers and edge routers in accordance with another embodiment of the invention.

[0010] Fig. 7 is a diagrammatic illustration of an alternative LIM.

Detailed Description of the Preferred Embodiments

[0011] One way to scale these networks larger at lower costs is to use a network or matrix of Ethernet switches to perform the functions currently being performed by expensive routers. These Ethernet switch matrices can be used in place of the terabit MPLS backbone routers, as well as in place of gigabit access routers at the edge of a network backbone. By using the Ethernet switch matrices, unit costs can be lowered.

[0012] While cost is a concern, scalability (*i.e.*, the ability to grow with bandwidth demands) is also a concern when designing and implementing new systems. In fact, some forecasters are estimating a significant demand growth. Thus, the ability to scale the network at reasonable costs will be very important.

[0013] Three systems have been developed to address these issues. These systems can be used individually or together to form a cost effective, scalable core backbone network and/or edge network. The systems include a multi-chassis Ethernet router ("MER"), a multiple parallel backbone configuration ("NxBB"), and a LAN in the middle ("LIM") configuration.

Multi-Chassis Ethernet Router (MER)

[0014] In one embodiment, the MER will comprise a multi-stage CLOS matrix (*e.g.*, 3 stages) router built out of Ethernet switches. The MER will use IP protocols to distribute traffic load across multiple switch stages. This design leverages existing technology, but allows scalability by adding additional Ethernet switches, additional stages, a combination or both, or new, inexpensive MERs.

[0015] Fig. 1 is a diagrammatic illustration of one embodiment of a 3-stage MER in accordance with one embodiment of the invention. In this particular embodiment, the MER utilizes 4 Ethernet switches in each of the three stages. Again, additional switches or stages can be added. In this particular example, as illustrated by the arrows in Fig. 1, traffic destined out L34 arrives at L11. L11 equally distributes the traffic across L21-L24 using one or more load balancing or distribution methods. L21-L24 forwards traffic to L34, which combines the flows and forwards them out the necessary links. This design provides a dramatic increase in scale. For example, in the illustrated embodiment, a 4 x MER provides a 4x increase in node size. The maximum increase for a 3 stage fabric is n^2/2, where n is the number of switches used in each stage. Five stage and seven stage matrices will further increase scalability.

[0016] While CLOS matrices are known, CLOS matrices have not been implemented in a network of Ethernet switches, which is what this particular implementation provides. Further, the CLOS matrices typically implemented in the very expensive MPLS routers are implemented using proprietary software and are encompassed into a single box. In this particular implementation, multiple inexpensive Ethernet switches are formed into the matrix, and the CLOS distribution is implemented using IP protocols, not a proprietary software. Further, in this particular implementation, the CLOS matrix is implemented at each hop of the switches, instead of in a single device. Other protocols can be used in other embodiments.

[0017] After the Ethernet switches are connected together, the packets and/or packet cells can be distributed to the different stages of the matrix using flow based load balancing. Internal gateway protocols ("IGP") can be used to implement the load balancing techniques. In some embodiments, the MER can utilize equal cost load balancing, so that each third-stage box (*i.e.*, L31, L32, L33 and L34) associated with a destination receives the same amount of traffic. For example, if boxes L1, L2 and L3 all communicate with New York, each box will receive the same amount of traffic. This technique is relatively easy to implement and scales well, when new MERs are implemented.

[0018] In another embodiment, traffic on the MER can be distributed using bandwidth aware load balancing techniques, such as traffic engineering techniques (*e.g.*, MPLS traffic engineering) that send packets to the least busy switch. In one embodiment, the middle layer can run the traffic engineering functionality, thus making intelligent routing decisions.

[0019] In yet another embodiment, traffic awareness techniques in the middle layer (*i.e.*, L21, L22, L23, and L24) can be used to determine what the downstream traffic requirements might be. That is, the middle layer can determine demand placed on the third or last layer and then determine routing based on the capacity needs. In this embodiment, the middle layer can receive demand or capacity information from the last (*e.g.*, third) layer via traffic engineering tunnels (*e.g.*, MPLS tunnels) or via layer 2 VLANS. Alternatively, changes to IGP can be leveraged to communicate bandwidth information to the middle layer. For example, switch L31 can communicate to the middle layer (*e.g.*, via IGP or other protocols) that it is connected to New York with 30Gb of traffic. The middle layer can use this protocol information, as well as information from the other switches, to load balance the MER.

[0020] In another embodiment, an implementation of the MER can use a control box or a route reflector to manage the MER. In some embodiments, the route reflector or control box can participate in or control routing protocols, keep routing statistics, trouble shoot problems with the MER, scale routing protocols, or the like. In one embodiment the route reflector can implement the routing protocols. So, instead of a third stage in a MER talking to a third stage in another MER, a route reflector associated with a MER could talk to a route reflector associated with the other MER to determine routing needs and protocols. The route reflector could utilize border gateway protocols ("BGP") or IGP route reflection protocols could be used (*e.g.*, the route reflector could act as an area border router).

Multiple Parallel Backbones (NxBB)

[0021] Another implementation that can be utilized to scale a core backbone network is to create multiple parallel backbones. One embodiment of this type of implementation is illustrated in Fig. 2. With the NxBB configuration, traffic can be split across multiple backbones to increase scale.

[0022] As illustrated in Fig. 2, one embodiment of an implementation deploys a series of parallel backbones between core sites. The backbones can use large MPLS routers, Ethernet switches, the MERs discussed above, or any other suitable routing technology. In addition, in the illustrated embodiment, peers can connect to the backbones through a common peering infrastructure or edge connected to each backbone, and customers can connect to specific backbone edges. That is, peers are connected to the parallel backbones (BB, BB1, BB2, BB3 and BB4) through a single peering edge, and customers are connected to the backbones through separate edge networks. In Fig. 2, each backbone has is own customer edge network. In alternative embodiments, however, only one or just a couple of edge network might be utilized (similar to one peering edge). The edge network also can use different routing technologies, including the MERs discussed above. The use of MERs can help with scaling of the peering edge.

[0023] The arrows in Fig. 2 illustrate an example of traffic flows in a parallel backbone network. In this example, traffic destined for customers A-Z arrives from Peer #2. The peering edge splits traffic across the multiple backbones based on the final destination of the traffic (*e.g.*, peering edge can distribute traffic based on IP destination prefix). Then each of the backbones forwards traffic through its associated customer edge to the final customer destination.

[0024] This multiple parallel backbone network can have many advantages. For example, parallel backbones make switching needs smaller in each backbone, so Ethernet switches and/or MERs can be used. In addition, the parallel backbone configuration can leverage existing routing and control protocols, such as BGP tools like traffic engineering, confederations, MBGP, and the like. The use of the traffic engineering protocols can help steer traffic to the appropriate backbone(s). Further, with the existence of multiple backbones, fault tolerant back-up systems can be created for mission critical applications. That is, one or more backbones can be used for disaster recovery and/or back-up purposes. Further, in yet other embodiments, the parallel backbone can be organized and utilized based on different factors. For example, a peer could have one or more backbones dedicated to it. Similarly, a customer could have one or more backbones dedicated to it. In yet other embodiments, customers can be allocated across backbones based on traffic and/or services. For example, Voice Over IP (VoIP) might use one or more backbones, while other IP service might use other backbones. Thus, backbones can be provisioned by peer, customer, service, traffic volume or any other suitable provisioning parameter.

[0025] Further, as illustrated in Fig. 3, a combination of multi-chassis Ethernet routers (MER) and parallel backbones (NxBB) can be used for even greater scaling. For example, as illustrated in the example in Fig. 3, a 300G Ethernet switch capacity could be increased 64x to 19,200G using a combination of MER and parallel backbones. In this example, an 8x MER and an 8x parallel backbone is combined to get 64x scalability. Scalability can be even larger if larger MERs (*e.g.*, 16x or 32x) and/or more parallel backbones are used. Thus, these technologies used alone and/or together can help scale capacity greatly.

[0026] Further, as illustrated in Fig. 4, an Ethernet-based core (*e.g.*, a core based on MERs) can be added as a parallel core to existing MPLS cores, thus adding easy scalability at a reasonable price without having to replace existing cores. In this implementation, some existing customers as well as new customers could be routed to the new Ethernet-core backbone. Alternatively, specific services, such as VoIP could be put on the new backbone, while leaving other services on the MPLS. Many different scenarios of use of the two cores could be contemplated and used.

[0027] Fig. 5 is another illustration of the Ethernet-based parallel core in parallel with an existing MPLS core. BGP techniques can be used to select which backbone to use on a per destination basis. Candidate routes are marked with a BGP community string (and IP next hop) that forces all traffic to the destination address to the second backbone. The selection can be done on a route by route basis and could vary based on source. Alternatively, a customer-based global policy can be used so that all traffic exiting a specific set of customer parts would use the same backbone. Route selection and route maps can be automatically generated by capacity planning tools.

LAN in the Middle (LIM)

[0028] Another network implementation that could used to scale backbone cores is the LIM. One embodiment of a LIM is illustrated in Fig. 6. In the illustrated embodiment, core routers are connected to edge routers through Ethernet switches. This is a similar configuration to the MERs discussed above, except existing core routers and edge routers are used in stages 1 and 3, instead of all stages using Ethernet switches. The benefit of this configuration is that the existing routers can be scaled larger without having to replace them with Ethernet switches. Using Ethernet switches in the middle layer and using CLOS matrices, as discussed above, will increase capacity of the existing core and edge routers. In one embodiment, the core and edge routers will be responsible for provisioning the traffic through the matrix.

[0029] Fig. 7 is a diagrammatic illustration of an alternative LIM. Customer facing provider edges (PE) can, for example, have 4 x 10G to the LIM. With a 1+1 protection, this would allow 20G customer facing working traffic. On the WAN facing side, each provider or core router (P) has 4 x 10 G to the LIM. With 1+1 protection, this allows at least 20 G of WAN traffic.

[0030] Although the present invention has been described with reference to preferred embodiments, those skilled in the art will recognize that changes can be made in form and detail without departing from the spirit and scope of the invention.

WHAT IS CLAIMED IS:

1. A router for routing Internet Protocol (IP) traffic between source and destination backbones, including:

an N x M IP-implemented CLOS matrix of Ethernet switches, where N>1 is the number of stages in the matrix and M>1 is the number or switches in each stage; and

routing protocol control means for distributing IP traffic between the switches.

2. The router of claim 1 wherein the routing protocol control means includes load balancing means for balancing the flow of traffic between two or more switches of each of one or more stages.

3. The router of claim 2 wherein the routing protocol control means includes flow-based load balancing means for balancing traffic.

4. The router of claim 2 wherein the routing protocol control means includes internal gateway protocol (IGP)-implemented means for balancing traffic.

5. The router of claim 2 wherein the routing protocol control means includes equal cost-based load balancing means for balancing traffic by causing each switch of a final stage associated with a common destination to receive about the same amount of traffic.

6. The router of claim 2 wherein the routing protocol control means includes bandwidth-aware load balancing means for balancing traffic.

7. The router of claim 2 wherein the routing protocol control means includes traffic-awareness load balancing means for balancing traffic.

8. The router of claim 1 wherein the routing protocol control means includes a controller or route reflector coupled to the matrix of switches.

9. The router of claim 1 wherein the routing protocol control means includes control means incorporated into the Ethernet switches.

10. A network system for routing Internet Protocol (IP) traffic between a source site and a destination site, including:

a plurality of discrete data transmission backbones between the source and destination sites;

source site control means for distributing IP traffic at the source site to the plurality of backbones for transmission to the destination site.

11. The network system of claim 10 and further including destination site control means for routing IP traffic received over the plurality of backbones at the destination site.

12. The network system of claim 11 wherein:

one or more of the backbones are dedicated to IP traffic from one or more customers;

the source site control means distributes IP traffic to the plurality of backbones as a function of the customer originating the traffic; and

the destination site control means includes control means dedicated to one or more of the backbones for routing IP traffic as a function of the customer originating the traffic.

13. The network system of claim 11 wherein:

the source site control means includes a peering edge common to all the backbones; and

the destination site control means includes a plurality of peering edges, each connected to one or more of the plurality of backbones, for routing IP traffic as a function of the customer originating the traffic.

14. The network system of claim 10 wherein the source site control means includes means for distributing IP traffic as a function of traffic volume on the plurality of backbones.

15. The network system of claim 10 wherein the source site control means includes means for distributing IP traffic as a function of the nature or type of the IP traffic.

16. A system for routing Internet Protocol (IP) traffic between a core backbone and edge, including:

an N x M IP-implemented CLOS matrix of switches, wherein:

N>1 is the number of stages in the matrix;

M>1 is the number or switches in each stage;

the M switches of the first and last stages are Multi-Protocol Label Switching (MPLS) switches; and

the M switches of at least one stage between the first and last stages are Ethernet switches; and

routing protocol control means for distributing IP traffic between the switches.

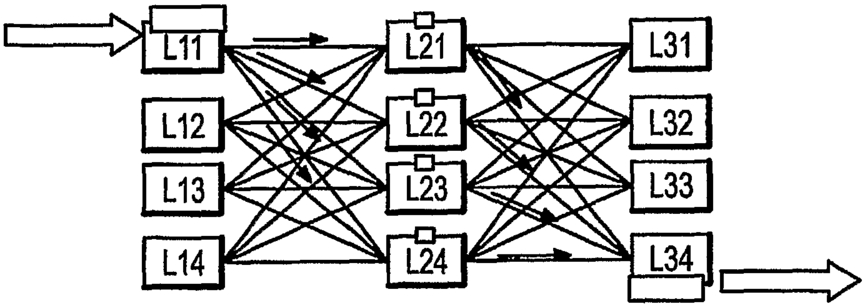

L11
L12
L13
L14
L21
L22
L23
L24
L31
L32
L33
L34

FIG.1

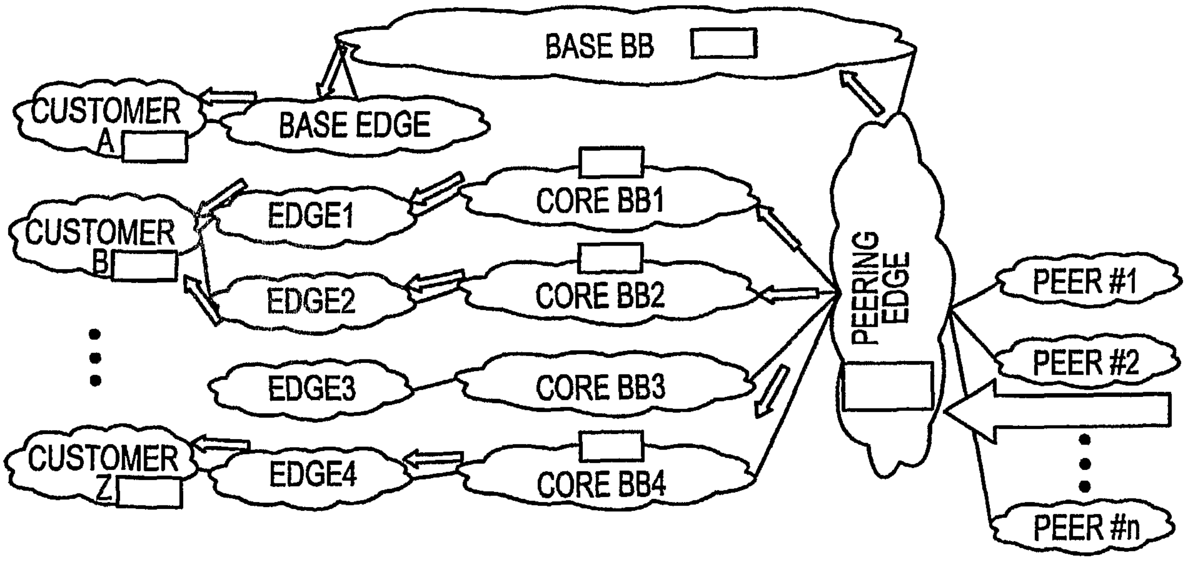

BASE BB
CUSTOMER A
BASE EDGE
CUSTOMER B
EDGE1
CORE BB1
EDGE2
CORE BB2
EDGE3
CORE BB3
CUSTOMER Z
EDGE4
CORE BB4
PEERING EDGE
PEER #1
PEER #2
PEER #n

FIG.2

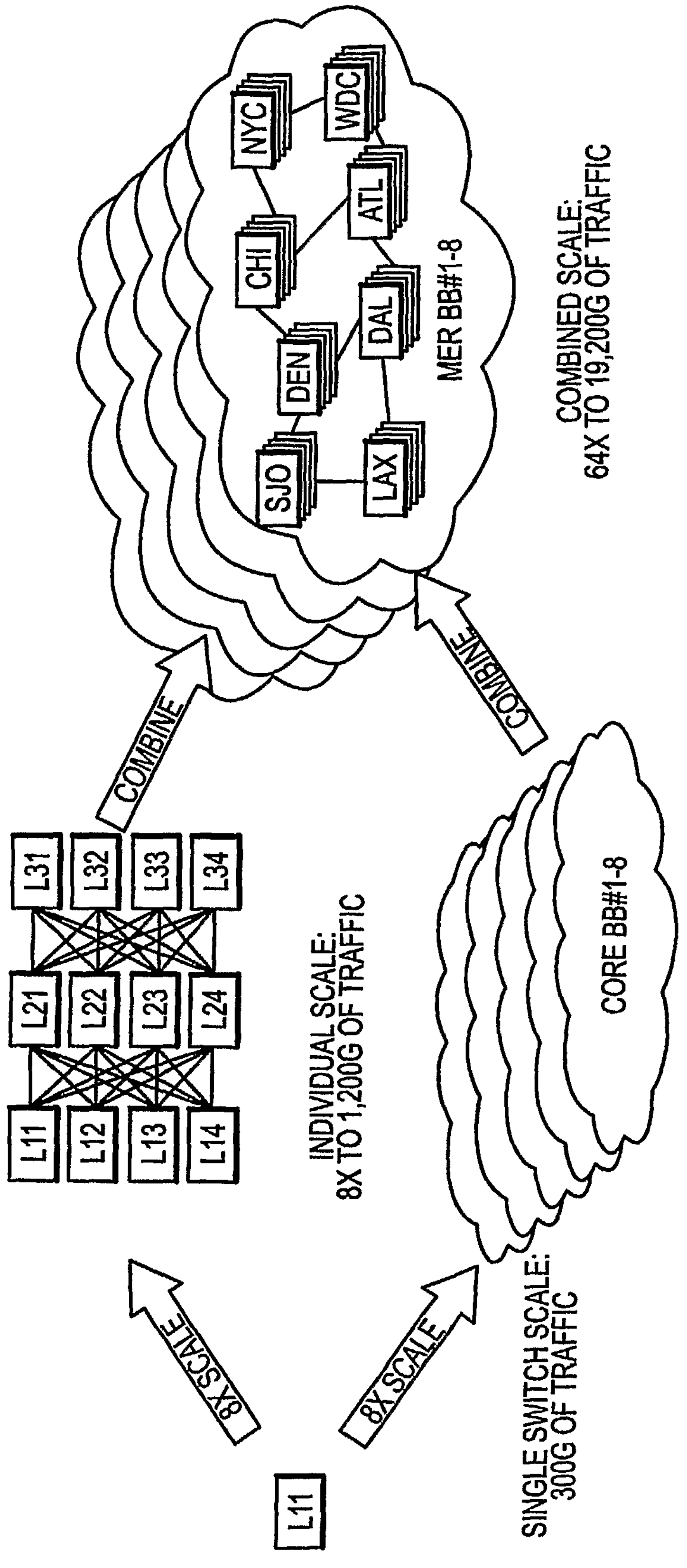

L11
8X SCALE
8X SCALE
SINGLE SWITCH SCALE:
300G OF TRAFFIC
L11
L12
L13
L14
L21
L22
L23
L24
L31
L32
L33
L34
INDIVIDUAL SCALE:
8X TO 1,200G OF TRAFFIC
CORE BB#1-8
COMBINE
COMBINE
SJO
LAX
DEN
DAL
CHI
ATL
NYC
WDC
MER BB#1-8
COMBINED SCALE:
64X TO 19,200G OF TRAFFIC

FIG.3

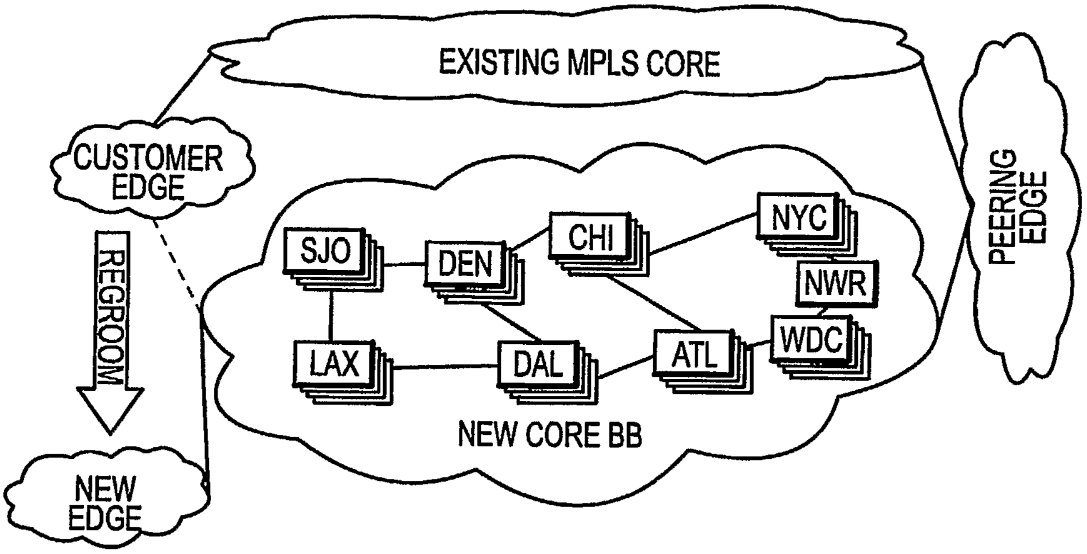

EXISTING MPLS CORE
CUSTOMER EDGE
REGROOM
NEW EDGE
SJO
DEN
CHI
NYC
NWR
LAX
DAL
ATL
WDC
NEW CORE BB
PEERING EDGE

FIG.4

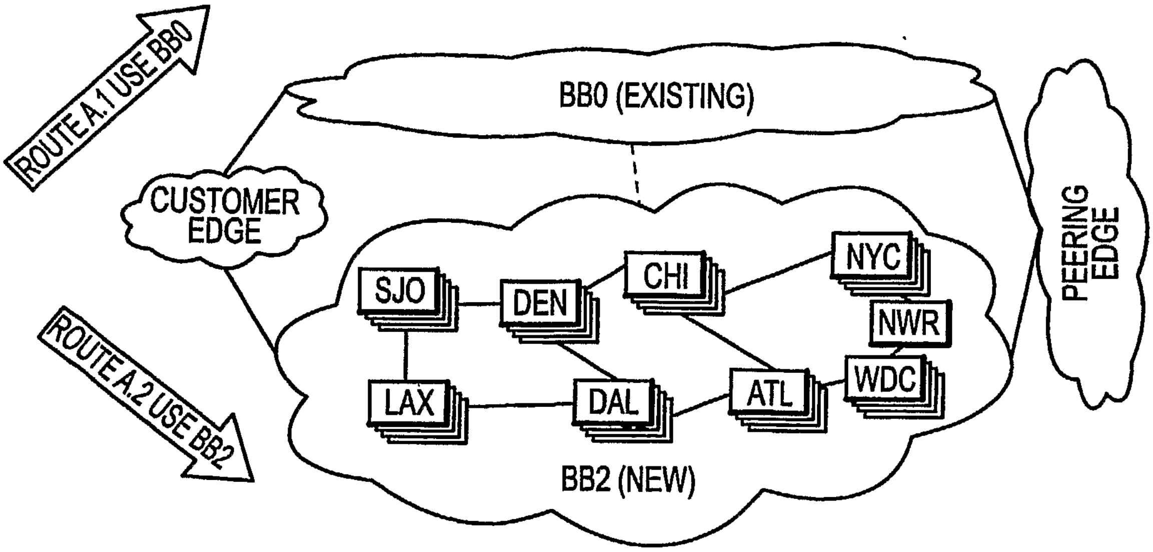

ROUTE A.1 USE BB0
BB0 (EXISTING)
CUSTOMER EDGE
ROUTE A.2 USE BB2
SJO
DEN
CHI
NYC
NWR
LAX
DAL
ATL
WDC
BB2 (NEW)
PEERING EDGE

FIG.5

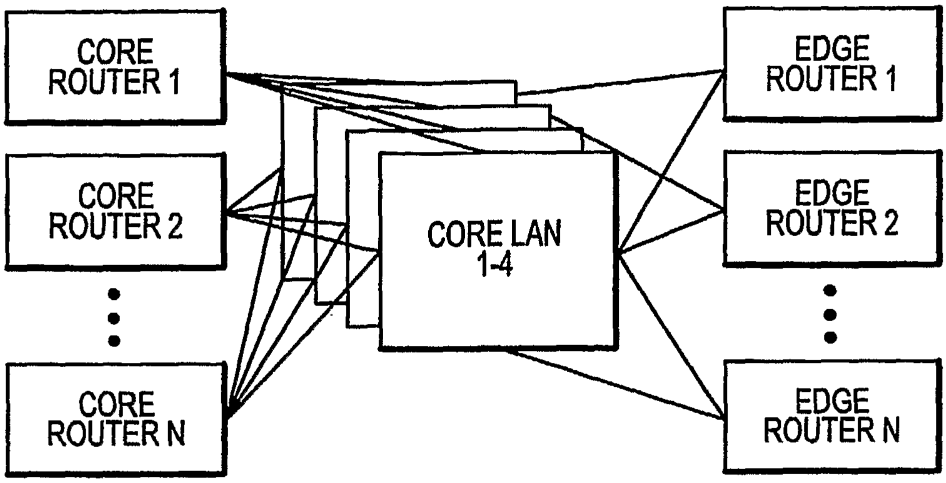

CORE ROUTER 1
CORE ROUTER 2
CORE ROUTER N
CORE LAN 1-4
EDGE ROUTER 1
EDGE ROUTER 2
EDGE ROUTER N

FIG.6

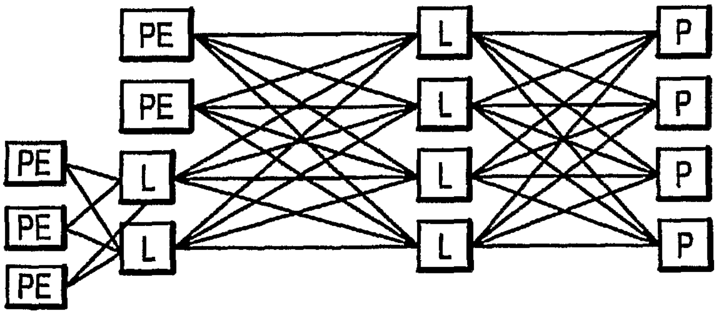

PE
PE
PE
PE
PE
L
L
L
L
L
L
P
P
P
P

FIG.7