

**Figure 1** Most current broadband deployments use a two-tier metro-aggregation-network and metro-core-network architecture.

gation router. The BRAS terminates subscriber traffic, such as PPPoX (Point-to-Point Protocol over Ethernet/ATM), VLAN (virtual local-area network), or other traffic, and applies policing and QoS (quality of service) per the subscriber profile.

The metro-aggregation router is the point at which normal residential Layer 2 traffic terminates and premium services, such as Layer 3 VPN (virtual private network), VPLS (virtual private-LAN service), EPLAN (Ethernet private LAN), and EPL (Ethernet private line), map to the metro-core layer as MPLS LSP (multiprotocol label-switching label-switched path), Layer 2/3 VPN, for example.

**TECH TRENDS** BY GOPAL GARG, CYPRESS SEMICONDUCTOR, AND R THIRUMURTHY, MIDAS COMMUNICATION TECHNOLOGIES

**SCALABILITY**

The aforementioned network architecture relies heavily on Ethernet-switching technology in the metro-access/aggregation network. However, Ethernet is a technology targeting the enterprise environment, with scalability, resilience, and other key network characteristics tuned for enterprise needs rather than for more demanding service-provider requirements. Over the past few years, the industry has introduced new protocols and management features to enable Ethernet to perform according to telephone-company standards. With respect to the scalability of metro Ethernet switches, two major issues developers must address are VLAN limitations and FDB (filtering-database) table size.

**VLAN**

One advantage of employing Ethernet in the enterprise domain is the ability to logically partition distinct user groups over the same physical network through a VLAN. This capability of Ethernet extends into the metro domain with user groups becoming individual subscribers or companies. However, the IEEE 802.1Q standard defines an address space of only 4096 available tags. With companies of-

# Optimized learning in metro switches

ETHERNET EXTENDS INTO AGGREGATION AND CORE NETWORKS.

Ethernet has emerged as an alternative access technology to SONET/SDH (synchronous optical network/synchronous digital hierarchy) for broadband deployments; its low cost and simple provisioning mechanism mainly drive this emergence. Most new broadband deployments across the globe use some form of Ethernet for their access, aggregation, and core networks. Next-generation SONET/SDH and RPR (resilient-packet ring, IEEE 802.17) have given greater impetus to the deployment of Ethernet services and extended Ethernet technology into aggregation and core networks.

Most current broadband deployments use a two-tier metro-aggregation-network and metro-core-network architecture (Figure 1). In the metro-aggregation-network portion, Layer 2 Ethernet switching aggregates traffic from access devices, such as DSLAM (digital-sub-

scriber-line-access multiplexer), MSPP (multiservice provisioning platform), and access switches. This Layer 2 traffic terminates in the metro aggregation router. In some cases, the BRAS (broadband remote-access server) collocates with (or is built into) the metro-aggre-

## AT A GLANCE

▶ Ethernet targets the enterprise environment and metro-area switches.

▶ New protocols and features enable Ethernet to comply with telephone-company standards.

▶ Major issues are virtual-local-area-network limits and filtering-database-table size.

▶ Adaptive, optimized-learning schemes can enhance performance.

fering subscribers multiple services, having only 4096 VLANs becomes a serious limitation. Providers can address this VLAN-scalability problem by increasing the VLAN available space through double-tagging (the IEEE 802.1ad draft, or QinQ) or VLAN stacking. This scheme essentially tags the packet with an outer VLAN tag, thereby expanding the address space to  $4096 \times 4096$  unique subscriber/service-identification tags.

## MAC-ADDRESS TABLE

The other scalability problem is the size of the MAC (medium-access-control)-address-learning table in switches. Because traffic in the aggregation network is Layer 2-switched, the Ethernet switches need to have a large MAC-address-learning (FDB) table that is proportional to the number of subscribers connected throughout the Layer 2 network. Proprietary methods, such as MinM (MAC in MAC), which encapsulate the subscriber Ethernet packet within the Ethernet header of the switch (using the switch's MAC address as the source), can address issues resulting from the table's large size. With MinM, intermediate metro switches need to learn only the switch addresses and not the actual subscriber MAC address. Only the switch that encapsulates the Ethernet packet needs to learn the MAC address of the subscribers that directly connect on its access links.

## OPTIMIZED LEARNING SCHEME

Most of the traffic-flow patterns in the access/aggregation network are P2P

(point-to-point) networks, such as networks in which the subscriber's Internet traffic terminates in the BRAS, the ASP (application-service-provider) server, or another area, or P2M (point-to-multipoint) networks, such as networks for multicast-video service, VPLS, and other services. In both cases, if you use QinQ, then the inside VLAN identifier identifies and classifies in the end nodes and is unseen by the intermediate switching nodes. Therefore, this discussion considers only the outside VLAN ID.

Consider a P2P case in which the traffic flows from the source to the edge switch/router and vice versa. This flow uniquely identifies with a VLAN ID. If the source node and the destination node establish a path, then in each intermediate switch, the path enters through a particular interface and exits through another specific interface. In other words, packets entering through one interface of the path will have exactly one interface through which they exit. This scenario can avoid the MAC-address learning for that flow in the transit switches.

In **Figure 2**, six switches (N1 through N6) connect in a partial mesh topology. The bold lines indicate the paths that loop-detection mechanisms, such as STP/RSTP (Spanning Tree Protocol IEEE 802.1d/Rapid STP, IEEE 802.1w), select for forwarding traffic. The P2P flow between nodes A1 and A2 passes through N1, N2, and N3. In this case, only nodes N1 and N3 need to learn the MAC addresses within the flow, and, if the ingress and egress interface of the flow is known, node N2 can pass through this flow without looking up the address.

Similarly, in the case of a multipoint flow (typically a VPLS), only certain nodes need to actually learn the MAC address; other nodes in the path can just forward the traffic. This situation is possible if you can create a tree between the nodes for each multipoint flow (identified with the VLAN ID). As **Figure 3** shows, assume that B1, B2, and B3 belong to a VPLS; the red lines in the **figure** indicate the subtree connection between the nodes for this flow. In this case, only the

switch nodes N1, N3, and N5 need to learn the MAC address for this multipoint flow. Nodes N2 and N4 can forward the traffic without learning the MAC address.

To summarize, only endpoints (for example, source and edge switches) and certain intermediate nodes should handle MAC-address learning for a P2P or P2M flow. All other intermediate nodes can blindly forward the packet from one interface to another without having to learn or look up the destination MAC address. The above optimization is possible if you can establish a path/tree for each flow (VLAN ID) in the intermediate switch.

With the aforementioned scheme, the switching nodes in the network need to learn the MAC addresses only for certain flows. This stipulation significantly reduces the FDB-table size for each node, thereby enhancing the scalability of Ethernet switching in the access/aggregation network.

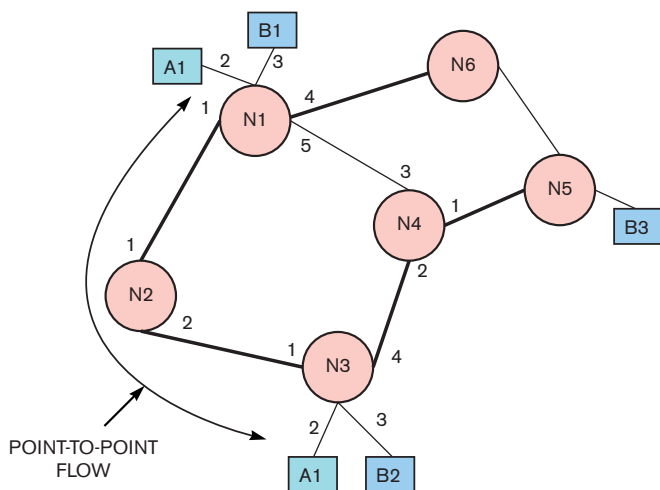
The main challenges in implementing this scheme include:

- deriving a path/tree for P2P flows, multipoint flows (QinQ ID), or both; and
- identifying the flows each node must learn.

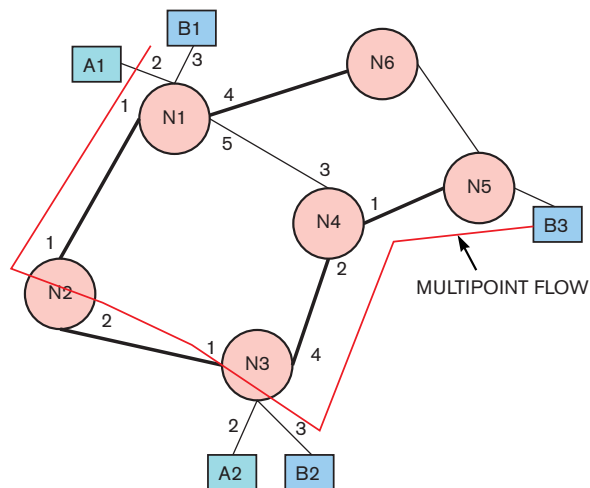
Establishing a path or tree for each flow requires a mechanism that selects the membership ports for each flow in the node. Because you use the VLAN for flow identification, you can use the GVRP (GARP VLAN Registration Protocol) for deriving the port membership. (GARP stands for Generic Attribute Registration Protocol.) IEEE 802.1p and 802.1Q define GVRP as providing a mechanism for dynamic maintenance of VLAN-registration entries for each VLAN and for propagating the information to other nodes. This information allows nodes to dynamically establish and update their knowledge of the set of VLANs that are active and the ports through which you can reach them.

Considering **Figure 2**, to provide P2P connectivity between A1 and A2, the static VLAN (say, V) entry is configured in port 2 of node N3 for user A2. Then:

- The GVRP registers the VLAN on



**Figure 2** Most of the traffic-flow patterns in the access/aggregation network are P2P (point-to-point) networks.



**Figure 3** In a multipoint flow (typically a virtual private-LAN service), only certain nodes need to actually learn the MAC address.

the port connected to A2 and propagates it on all other ports.

- Node N2 receives the registration and registers the VLAN on the port (port 2) received, then propagates the registration on port 1.
- N1 receives the registration in port 1 and includes port 1 as a member of VLAN V, then propagates the registration on all other ports.

In the above sequence, port 2 of N2 and port 1 of N1 have become members of the VLAN V (flow).

Because A1 belongs to the same flow, the same VLAN entry will be statically configured in N1 on port 2. Hence:

- N1 propagates the VLAN membership on ports 1, 3, and 4.
- N2 receives the registration message on port 1 and forwards it to port 2.
- N3 receives the registration on port 1 and includes port 1 as a member of VLAN V.

This sequence includes ports 1 and 2 of N1, ports 1 and 2 of N2, and ports 1 and 2 of N3 as members of VLAN V, creating a bidirectional path. You can easily extend this procedure to create a tree for a multipoint flow.

The GVRP runs within the context of the spanning tree created through RSTP/STP. This step ensures that the VLAN membership updates only on the primary path that the RSTP/STP creates. When

the primary path breaks, GVRP updates VLAN membership on the nodes in the secondary path. The system automatically recalculates the path or tree created on link or node failure.

### FDB LEARNING RULE

Once you establish the tree/path, a simple learning rule enables implementation of the above scheme. In the switch, MAC-address learning occurs selectively for each VLAN; in other words, for each VLAN, the following rules enable or disable learning.

For a particular VLAN, learning is enabled if:

1. one of the members of the VLAN is an access port, or
2. the number of members of the VLAN is greater than two.

Multipoint flows require the first rule to prevent flooding of downstream traffic on the access ports and for access control of P2P flow, such as EIA (Ethernet Internet access). The second rule enables learning only if the number of member ports for any particular VLAN is greater than two and disables learning for VLANs whose members are just two (intermediate switches on the VLAN path). In metro switches, for most transit VLANs, the members for the VLAN are two; hence, learning is disabled for those VLANs, and one

interface can efficiently forward the packet to another.

### REQUIREMENT IN SWITCHES

Implementing the above scheme in a switch requires minor modifications in the protocol stack and switching fabric: a capability in the switching fabric to enable/disable learning for each VLAN and a control-plane implementation of the above-mentioned rules. Most switching fabrics available today provide the capability to enable/disable learning for each VLAN indirectly, through filtering rules, forwarding rules, or other mechanisms.

The control-plane implementation requires only a simple software modification that can be a special configuration for GVRP. GVRP switches off learning for a VLAN depending on the number of members in that VLAN, along with the rules mentioned in the previous section.

### TCAM-BASED ARCHITECTURE

You can most efficiently implement the route look-up mechanism in a network processor and TCAM (ternary-content-addressable-memory)-based architecture. You can implement the control function in the forwarding plane; in this way, when the system encounters an ingress/egress interface failure, it can detect the alternate interface for a flow and transmit the packet with little or no dis-

ruption. You can easily accomplish the mapping of a failed interface to the flow (VLAN) if you maintain the VLAN table in TCAM. You can optimize this type of implementation with an extra procedure to provide fast convergence for flows.

Such an implementation is straightforward in metro switches and optimizes learning without affecting any other protocol behavior of the switch, enabling switches with optimized learning to in-

teroperate seamlessly with other switches. The scheme works effectively in metro-access/aggregation networks with any topology, including ring or mesh. Because the technique is based on GVRP, the optimization procedure can work with VLAN STP and MSTP (multiple STP).

### SCALABILITY

Without optimized-learning mechanisms, intermediate switches in the ac-

cess/aggregation network must learn the source MAC address of all traffic that flows through it. Hence, if  $N$  is the number of switches in the Layer 2 access/aggregation network, and  $M$  is the number of stations connected in each node, then the FDB-table size that each node requires is  $O(M \times N)$ .

Using an optimized-learning approach, switches learn the MAC addresses of the nodes in the access interfaces and for some multipoint transit flows for which the switch has more than two ports as members. Hence, if  $f$  is the number of flows with more than two ports as members, and  $K$  is the number of nodes for each flow, then each switch requires an FDB-table size of  $O(N + f \times K) \rightarrow O(N)$ .

Few flows require address learning, but if you ignore the ones that do, the required FDB-table size is  $O(N)$ . Hence, you can significantly reduce the large FDB-table-size requirement in metro Ethernet switches. The aforementioned optimization procedure reduces the MAC-address learning-table size from  $O(M \times N)$  to  $O(N)$ .

The optimized-learning scheme that this article describes reduces the MAC-address learning-table size necessary in metro Ethernet switches. You can implement this scheme on most common switch fabrics and network-processor-based architectures. The implementation of optimized learning reduces switch complexity as well as the required capital expenditure. The technique simplifies network management and troubleshooting, because learning occurs only at the node level, and traffic tunnels through the rest of the network. **EDN**

### AUTHORS' BIOGRAPHIES

*Gopal Garg is a senior technology-marketing director at Cypress Semiconductor, where he markets networking semiconductor products. He holds a BS from Birla Institute of Technology and Science (Pilani, India) and an MBA from PU India.*

*R Thirumurthy is the head of R&D at Midas Communication Technologies, where he is responsible for broadband-product architecture, design, and product positioning. He holds a BE in computer science from College of Engg, Guindy (Chennai, India), and an MS in computer science from the Indian Institute of Technology, Madras.*