# The Road to Carrier-Grade Ethernet

*Kerim Fouli and Martin Maier, Optical Zeitgeist Laboratory, INRS*

## ABSTRACT

Carrier-grade Ethernet is the latest step in the three-decade development of Ethernet. This work describes the evolution of Ethernet technology from LAN toward a carrier-grade operation through an overview of recent enhancements. After reviewing native Ethernet and its transport shortcomings, we introduce the major carrier-grade upgrades. We first discuss the evolution of layer-2 architectures. Then, we detail the service specifications and their QoS and traffic engineering requirements. Finally, we describe the new OAM and resilience mechanisms.

## INTRODUCTION

Ethernet has enjoyed great success as the major enabling technology for local area networks (LANs). By 1998, Ethernet accounted for 80 percent of the LAN installed base, and Ethernet port shipments exceeded 95 percent of the market share [1]. Originally set to 10 Mb/s in the 1980s, Ethernet transmission rates have evolved to higher speeds ever since, reaching 10 Gb/s upon the approval of the IEEE 802.3ae standard in 2002. Ten-gigabit Ethernet (10GbE) was the first Ethernet standard to include interoperability with carrier-grade transmission systems such as synchronous optical network/synchronous digital hierarchy (SONET/SDH). In addition to its high-speed LAN operation, 10GbE was shown to integrate seamlessly with metropolitan and wide area networks [2]. The drive toward higher transmission rates continues as the IEEE 802.3 Higher Speed Study Group (HSSG) works on the standardization of 100GbE by 2010. Ethernet passive optical network (EPON), the extension of Ethernet LANs to access environments, is poised to undergo its first tenfold, bit-rate leap from 1 Gb/s (802.3ah, 2004) to 10 Gb/s (802.3av) by 2009.

Despite increased speed and interoperability with carrier-grade technology, Ethernet has remained exclusively a LAN and access network technology. Indeed, traditional Ethernet lacks essential transport features such as wide-area scalability; resilience and fast recovery from network failures; advanced traffic engineering; and operation, administration, and maintenance (OAM) capabilities. Consequently, it falls short of delivering the quality of service (QoS) and security-guarantee levels required by typical transport-network service level agreements (SLAs).

Carrier-grade Ethernet (CGE) is an umbrella term for a number of industrial and academic initiatives that aim to equip Ethernet with the transport features it is missing. In doing so, CGE efforts aspire to extend the all-Ethernet domain beyond the first-mile and well into the metropolitan and long-haul, backbone networks.

The thrust toward CGE is driven by the promise of a reduced protocol stack. The ensuing reductions in cost and complexity are expected to be considerable. Currently installed metro and wide area networks are dominated largely by SONET; by multiprotocol label switching (MPLS); and to a lesser extent, by asynchronous transfer mode (ATM) technology, even though Internet Protocol (IP) routers lead new installations. (It is worth mentioning that having originally been conceived as a hardware switch, MPLS was subsequently revised to become a software feature.) Notwithstanding, most data traffic originates from and terminates at Ethernet LANs. In addition, most applications and services such as video and business services are migrating toward Ethernet platforms [3]. The current growth of voice over Internet Protocol (VoIP) and the expected emergence of Internet Protocol television (IPTV) imply the acceleration of that trend, leading to inhibitive costs associated with network layering and interfacing [4]. In [5], the authors show that implementing CGE in backbone networks could result in a 40 percent port-count reduction and 20–80 percent capital expenditure (CAPEX) drops compared to various non-Ethernet backbone technology alternatives.

After reviewing the evolution of currently deployed Ethernet technology in the next section, this work introduces the proposed major carrier-grade enhancements. We then present the new IEEE 802.1 hierarchical forwarding architecture; the following two sections offer an overview of the emerging service, traffic engineering, resilience, and OAM standards. We conclude in the final section.

## THE EVOLUTION TOWARD CARRIER GRADE

### NATIVE ETHERNET

Ethernet is a family of standardized networking technologies originally designed for LANs. The first experimental Ethernet was a carrier-sense multiple-access with collision detection (CSMA/CD) coaxial bus network that was

0163-6804/09/$25.00 © 2009 IEEE

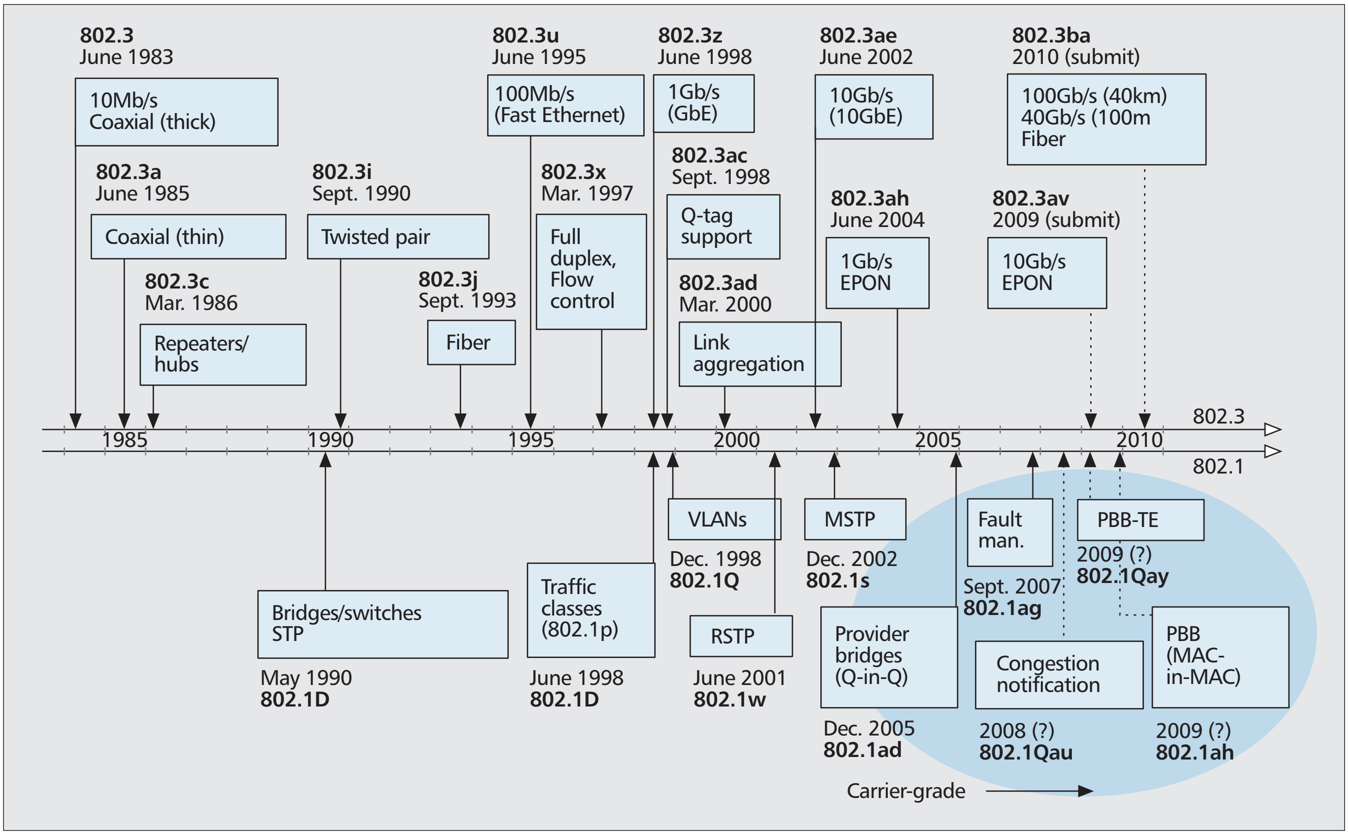


**Figure 1.** *Ethernet standardization milestones.*

designed in 1972 at Xerox and operated at a speed of 2.94 Mb/s. Eleven years elapsed before a 10-Mb/s version was proposed and endorsed as the first IEEE 802.3 standard in 1983. Figure 1 highlights some of the major standardization milestones since 1983.

Early on in the standardization process, the IEEE categorized standards by separating the Ethernet physical (PHY) layer and the medium access control (MAC) sub-layer standards (802.3, upper timeline in Fig. 1) from data-link layer bridging and management standards (802.1, lower timeline in Fig. 1).

The gradual evolution of Ethernet toward higher speeds is clear from the typical tenfold bit-rate increase characterizing the standardization of Fast Ethernet (802.3u) in 1995, Gigabit Ethernet (GbE, 802.3z) in 1998, 10GbE (802.3ae) in 2002, and the projected endorsement of 100GbE (802.3ba) by 2010. Note that the rates shown in Fig. 1 are data rates. Physical-layer clock rates are typically higher due to line-encoding schemes such as 8B/10B in GbE and 64B/66B in 10GbE.

The bit-rate leaps of Ethernet were accompanied by major qualitative technology transformations that are apparent in the physical/MAC (802.3) and the bridging and management (802.1) efforts. On the physical and MAC layers, CSMA/CD on thick coaxial cable (802.3-1983) was gradually abandoned in favor of a hub-segmented (802.3c, 1986) and then a switched (802.1D-1990) network capable of operating in full duplex mode (802.3x, 1997) over various media such as twisted pair and fiber.

In addition to implementing operations such as flow control and link aggregation (802.3ad, 2000) at the MAC level, Ethernet acquired a high degree of management functionality due to the enabling of traffic classes (802.1p, 1998), virtual LANs (VLANs, 802.1Q, 1998), and provider bridges (802.1ad, 2005). Furthermore, the virtual topology process of Ethernet shifted from Spanning Tree Protocol (STP, 802.1D-1990) to the more elaborate Rapid Spanning Tree Protocol (RSTP, 802.1w, 2001) and Multiple Spanning Tree Protocol (MSTP, 802.1s, 2002). Several of the aforementioned network-level amendments are briefly described in the following points.

***Switching*** — Switches and bridges started out as layer-2 devices connecting different LANs. Typically hardware-based, switches have a MAC layer at each port. They build and maintain a source address table (SAT) associating the source addresses of incoming frames to their corresponding ports — a process called *learning*. If the destination address of an outgoing frame is not on its SAT, a switch acts like a layer-1 repeater by sending a copy of the frame to all output ports. This is called *flooding*.

***Full Duplex*** — Full duplex operation refers to the creation of dedicated paths, rather than use of a shared medium, between nodes. This requires switching and enables nodes to transmit and receive at the same time using the full capacity of two counter-directional links.

|  | IEEE | ITU | IETF | MEF |
|---|---|---|---|---|
| Architecture and interfaces | 802.1Q<br>802.1ah<br>802.1ad<br>802.1Qay | G.8010/Y.1306<br>G.8012/Y.1308 |  | MEF-4<br>MEF-11<br>MEF-12 |
| Survivability | 802.1ag<br>802.1Qay<br>802.1aq | G.8031<br>G.8032 |  | MEF-2 |
| TE, QoS, and service specifications | 802.1Qay | G.8011/Y.1307 |  | MEF-3<br>MEF-6<br>MEF-10.1 |
| OAM and network configuration | 802.1ah<br>802.1ag<br>802.1AB<br>802.1ar<br>802.1Qau | Y.1730<br>Y.1730 | GELS<br>GMPLS control<br>— of PBT | MEF-7<br>MEF-15<br>MEF-16<br>MEF-17 |
| Security | 802.1AE/af |  |  |  |

■ **Table 1.** *Recent standardization initiatives sorted by standards body and field.*

***Flow Control*** — Although switching removed collisions, nodes could still face overflow problems, particularly when faster transmission rates are used at source nodes or intermediate switches. Flow control enables the receiver to regulate incoming traffic by sending PAUSE frames that halt transmission temporarily.

***STP*** — Parallel paths between nodes can create forwarding loops leading to excess traffic or large peaks in broadcast traffic (broadcast storms). By defining a logical tree topology, STP specifies a unique path between any pair of nodes and disables parallel paths, thus eliminating loops. Disabled links are used for backup. Switches/bridges use special frames called bridge protocol data units (BPDUs) to exchange STP information.

***Link Aggregation*** — In disabling parallel links between adjacent nodes, STP blocks valuable bandwidth increases. Link aggregation overcomes the STP limitation and enables nodes of exploiting parallel links.

***VLAN*** — A VLAN is essentially a logical partition of the network. VLANs were introduced to split the LAN broadcast domain, thus increasing performance and facilitating management. VLANs were initially communicated implicitly through layer-3 information such as a protocol type or an IP subnet number. The 802.1Q standard introduced the Q-tag, a new frame field that includes an explicit 12-bit VLAN identifier (VLAN ID).

***Traffic Classes*** — Besides the VLAN ID, the Q-tag also includes a three-bit field used to specify frame priority. An 802.1Q/p switch uses a queuing system capable of recognizing and processing the eight possible priority levels, as detailed in the 802.1p amendment.

***RSTP and MSTP*** — RSTP is an improved version of STP that achieves faster convergence by introducing measures such as more efficient BPDU exchanges. Rather than disabling parallel links as in STP and RSTP, MSTP exploits them by defining different spanning trees for different VLANs.

Ethernet is sometimes dubbed as "the cheapest technology that is good enough." Due to its simplicity, low cost, and high level of standardization, it exhibits excellent compatibility and interoperability with complementary technologies. However, these same trademark attributes are the source of a number of fundamental shortcomings as a transport platform.

## Shortcomings of Native Ethernet

When it comes to delivering transport services, native Ethernet suffers from the following major shortcomings.

***Architectural Scalability*** — In spite of its universal MAC addressing scheme, traditional Ethernet is not scalable to a wide-area environment because of the lack of separation between client and service domains and its address-learning method, based on flooding.

***Resilience*** — The Ethernet network response to failure is based on the reconfiguration of malfunctioning spanning trees. Although STP was superseded by RSTP and MSTP, these developments still fall short of expected wide-area network (WAN) protection speeds.

***Traffic Engineering (TE)*** — Native Ethernet requires no traffic engineering and management. Up to now, it has relied on other layers to perform basic traffic engineering operations. The enabling of a number of traffic classes through 802.1p does not allow larger networks to provision bandwidth and fine-tune traffic across the whole network. Those TE capabilities are among the fundamental aspects of carrier technologies because they enable QoS guarantees.

***OAM*** — By enabling such operations as network configuration, equipment maintenance, and performance monitoring, OAM resources exploit TE capabilities to enable the delivery of SLA guarantees. Designed for LAN operation, traditional Ethernet lacks such capabilities.

***Security*** — The concern for security grows with the number of network subscribers, making the operations of client authentication and authorization vital. For proper WAN functionality, Ethernet requires integrated security enhancements to address issues such as flooding.

Another important issue of Ethernet is synchronization. Many services and applications require synchronization of time and frequency, namely the distribution of an accurate and reliable time-of-day clock signal and/or frequency reference. The two major applications are mobile backhaul and time-division multiplexing (TDM) circuit emulation, where a frequency reference is required to derive transmission frequencies at a mobile station, or a time reference is required to recover transmitted bits at the edge-TDM emulation points. Other important applications include audio/video access applications.

TDM legacy technologies, such as SONET/SDH, naturally disseminate synchronization signaling. In contrast, reception in 802.3 Ethernet is inherently asynchronous, where preamble bits are used to achieve synchronization on a per-frame basis. Therefore, the migration from the TDM infrastructure to Ethernet means the loss of useful disseminated synchronization signaling.

Standardization bodies are proposing solutions for Ethernet synchronization at different network layers. Layer-2 and layer-3 synchronization is based on the multicasting of synchronization frames or packets containing timestamps. Receivers subsequently adjust their local time according to the received timestamp, taking into account an evaluation of the transmitter-receiver delay. The standards in this category include IEEE 1588 and its more recent carrier-grade follow-ups developed by the IEEE 802.1 Audio/Video Bridging Task Group (802.1AS, 802.1Qat, and 802.1Qav), as well as Internet Engineering Task Force (IETF) Network Time Protocol (NTP).

Higher-layer synchronization may be affected by packet-delay variation and traffic load. Hence, the true equivalent to TDM synchronization must be resident in the physical layer. A group of International Telecommunication Union-Telecommunication (ITU-T) standards (G.8261, G.8262, and G.8264) provide the ability for physical-layer dissemination of frequency synchronization information similar to SONET/SDH and may form the basis for the reliable implementation of a highly accurate physical-layer time-of-day. These physical-layer ITU-T standardization efforts form the basis of what is called Synchronous Ethernet (SyncE) [6].

### The Ethernet Carrier-Grade Roadmap

Carrier-grade Ethernet denotes an all-Ethernet backbone infrastructure enabling an end-to-end frame transport that incorporates typical carrier-grade service and quality guarantees. The following five main objectives are addressed by current standardization efforts:

- Wide-area scalability
- Network resilience and fault recovery
- Advanced traffic engineering and end-to-end QoS
- Advanced OAM
- Security

The current 802.1 standardization efforts are focused on leveraging the existing Ethernet protocols and switch architectures to perform most of those functions on the data-link layer, thus enabling CGE while maintaining backward compatibility with legacy Ethernet equipment. In Fig. 1, the shaded area represents some of the future standard amendments designed to realize CGE. The CGE efforts currently involve major telecommunications standardization bodies such as IEEE, ITU, IETF, and the Metro Ethernet Forum (MEF). Table 1 shows a classification of some existing and in-progress standards according to their carrier-grade objectives.

The latest standardization developments in the IEEE 802.1 working group aim at meeting the mentioned carrier-grade objectives. Some of the recent and future standards involved are shown within the shaded area in Fig. 1. These include the architectural modifications of provider bridges (PB, 802.1ad, 2005) and provider backbone bridges (PBB, 802.1ah), the resilience instruments of fault management (802.1ag, 2007) and congestion notification (802.1Qau), and the specifications of PBB traffic engineering (PBB-TE, 802.1Qay). Those developments are discussed in greater detail in the remainder of this work.

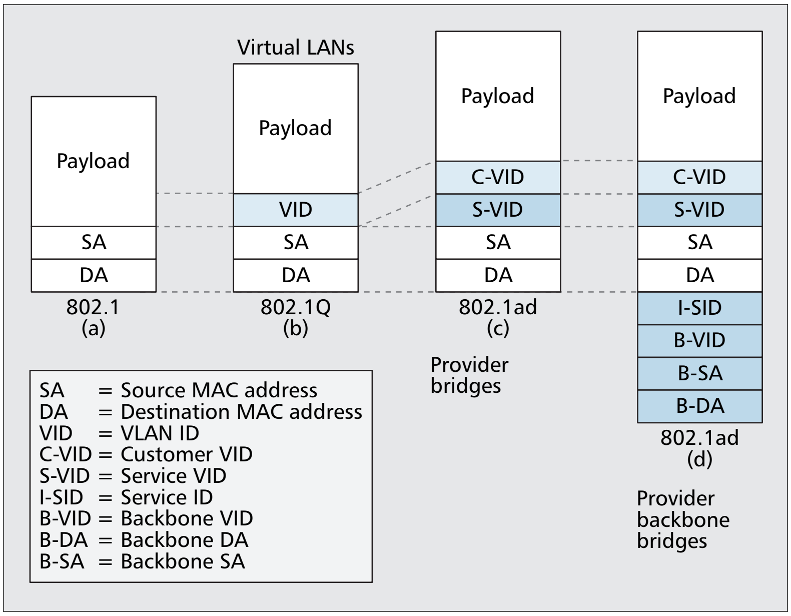


**Figure 2.** *Ethernet frame evolution [7].*

## Scalability Through Hierarchy

To fulfill carrier-grade objectives, the backbone nodes must deliver traffic that is protected, engineered, and guaranteed rather than best-effort traffic. From an architectural point of view, that implies two qualitative evolutions:

- Moving from a connectionless model supported by spanning-trees toward enabling multiple connection-oriented tunnels
- Moving from distributed-address learning to centralized path configuration

The step-by-step evolution toward such an architecture implied the introduction of hierarchical layer-2 sublayers, starting with VLANs (802.1Q), then with PBs (802.1ad), and finally with PBBs (802.1ah). Through tagging and encapsulation, the original 802.3 Ethernet frame shown in Fig. 2a underwent a consequent evolution aimed at preserving its structure for backward compatibility.

In this section, we detail the evolution of network-layer hierarchy and the associated forwarding modes designed to enable carrier-grade operation.

### VLAN Switching

The first carrier Ethernet attributes came with the emergence of VLANs. Although VLANs began as mere partitions of the customer enterprise network, they were seen by service pro-

Like MAC address learning, VLAN learning enables VLAN switches to associate new MAC addresses and VIDs dynamically with port information. To do so, VLANs maintain one or more filtering databases, depending on the learning process.

viders as the natural way to differentiate customer networks while maintaining a cheap end-to-end Ethernet infrastructure. In this setting, service providers assign a unique 12-bit VLAN ID (VID) field within the Q-tag to each customer network. VLAN switches add the Q-tag at the ingress node and remove it at the egress node (Fig. 2b).

Like MAC address learning, VLAN learning enables VLAN switches to associate new MAC addresses and VIDs dynamically with port information. To do so, VLANs maintain one or more *filtering databases*, depending on the learning process. Two VLAN learning schemes were specified in 802.1Q: independent VLAN learning (IVL) and shared VLAN learning (SVL). For a defined subset of VLANs, SVL uses one common filtering database, hence enabling the sharing of port information among VLANs. IVL, on the other hand, uses one filtering database per VLAN, thus restricting MAC learning to the VLAN space. A notable consequence of IVL is that forwarding is effectively specified by the full 60-bit combination of the destination MAC address and the VID.

IVL and SVL became instrumental in enabling service providers to separate the information of customers supported by the same provider VLAN switch. Nevertheless, this use of VLANs ran into scalability issues. The VID 12 bits limited the number of supported customers to a maximum of 4094 (excluding reserved VIDs "0" and "4095"). In addition, the customers required the same VID field to partition and manage their own networks, leading to a further reduction of the range of VIDs available to the service providers.

## Provider Bridges (Q-in-Q)

In an effort to mitigate the provider scalability problems, the VLAN plane was further split by introducing an additional Q-tag, represented in Fig. 2c by its VID field. This resulted in two separate VID fields, destined to be used by customers (C-VID) and service providers (S-VID), respectively. The VLAN stacking of two Q-tags was introduced in the 802.1ad standard and is often referred to as Q-in-Q.

In spite of service provider switches (PBs) controlling their own S-VID, scalability issues remained unresolved by 802.1ad. First, PBs still were required to learn all attached customer destination addresses (DAs), resulting in potential SAT overflows and broadcast storms. Second, control frames such as BPDUs were unrestricted to the provider or customer domains [8]. In addition, because it was designed for enterprise LAN applications, the 12-bit S-VID still was insufficient to perform two key functions simultaneously: the identification of customer service instances and forwarding within the provider network [9].

## Provider Backbone Bridges (MAC-in-MAC)

The 802.1ah standard draft introduces yet another hierarchical sublayer, this time by means of encapsulation of the customer frame within a provider frame. Backbone edge switches (PBBs) append their own source address (B-SA) and destination address (B-DA), as well as a backbone VID (B-VID). A new 24-bit field called the service ID (I-SID) also is introduced to identify a customer-specific service instance (Fig. 2d).

PBBs complete the separation between customer and provider domains by duplicating the MAC layer, hence the term MAC-in-MAC. In addition, PBBs allow up to 16 million service instances to be defined without affecting the forwarding fields (B-VID, B-SA, and B-DA).

## Provider Backbone Transport

Although PBBs create a provider infrastructure that is transparent to the customer networks, they still may use automatic best-effort techniques inherited from the LAN, such as xSTP and MAC-learning. Such techniques do not meet the configuration requirements of carrier-grade operation. Due to the modularity of Ethernet specifications, they can be turned off to pave the way for the fine-tuning of management-based processes. Moreover, the creation of a service-provider MAC layer requires a redefinition of the VLAN space to fulfill carrier-grade requirements.

In 802.1Qay, provider backbone transport (PBT) is defined by a backbone architecture implemented together with a set of measures to enable traffic engineering. PBT relies on PBBs at the edge of the provider network and PBs to perform forwarding within its core. Rather than multicast trees, VLANs represent connection-oriented, point-to- point (PtP) or multipoint-to-point (MPtP) tunnels (Ethernet switched paths [ESPs]) traversing the core from one PBB to another.

A range of B-VIDs is reserved to identify the ESPs. Rather than having global significance, these B-VIDs are tied to destination PBB addresses (B-DA) and can be reused. At each provider switch, the egress port of a frame is determined by the 60-bit combination of B-VID and B-DA. For instance, reserving 16 out of the 4094 B-VIDs implies a theoretical maximum of $16 \times 2^{48}$ available ESPs. This eliminates the scalability limits of VLAN stacking and is considered sufficient for transport purposes [7].

To enable PBT, the following measures are required at the provider switches and within the range of B-VIDs allocated for ESPs:

- Disable automatic MAC learning and flooding to enable the configuration of forwarding tables at the management layer.
- Filter (remove) unknown-destination, multicast, and broadcast frames.
- Disable xSTP to allow for loops and alternate path-oriented resilience mechanisms.
- For any given destination PBB, assign a unique B-VID to each ESP.
- Activate IVL.

The latter measure disables MAC address sharing between VLANs. This prevents the egress port associated to one ESP from being altered by the configuration of an alternate ESP toward the same destination PBB. Consequently, at the provider switches, the egress port is determined locally by the full 60-bit B-VID/B-DA sequence.

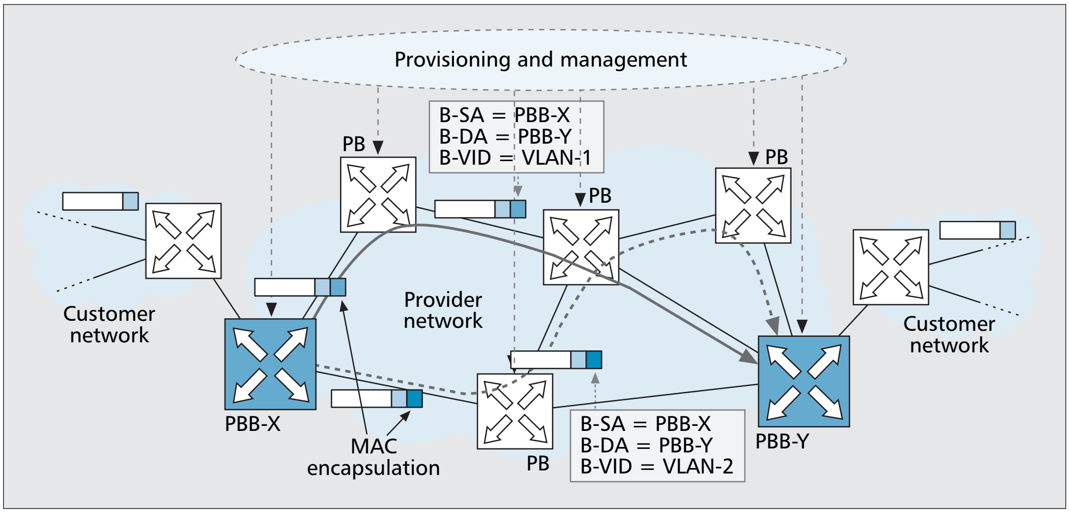


**Figure 3.** *Provider backbone transport (PBT) example [7].*

*Due to the modularity of Ethernet specifications, they can be turned off to pave the way for the fine-tuning of management-based processes. Moreover, the creation of a service-provider MAC layer requires a redefinition of the VLAN space to fulfill carrier-grade requirements.*

Note that provider switches can revert to normal VLAN switching outside the prescribed set of B- VIDs. Moreover, ESPs must be created in both directions to establish connection symmetry, a feature that is required for the proper operation of Ethernet customer networks [7].

The PBT architecture is illustrated in the example of Fig. 3. At the management plane, two paths are configured from PBB-X to PBB-Y to interconnect two customer networks. The forwarding tables of all the traversed PBB and PB switches are updated accordingly. A distinct B-VID is assigned to each path.

At PBB-X, customer frames are encapsulated. PBB-X, PBB-Y, and either VLAN-1 or VLAN-2 are entered in the B-SA, B-DA, and B-VID fields, respectively. The customer frame is recovered and forwarded to the customer layer at the destination edge node (PBB-Y). In the provider network core, enabling IVL ensures that PBs maintain distinct routes for different B-VIDs although the same destination address appears, as is the case when VLAN-1 and VLAN-2 cross in Fig. 3. (Note that in Fig. 3, the switches within the provider network may be connected either directly or by intermediate optical cross-connects.)

The PBT architecture allows for total path configuration from source to destination. Multiple ESPs can be created between PBBs for traffic engineering, load balancing, protecting connections, and separating service/customer instances.

## Toward Traffic Engineering and QoS-Enabled Services

The CGE architectures and switching technology described in the previous section serve to deliver Ethernet virtual connections (EVCs) between user-network interfaces (UNIs). The MEF defines a UNI as an interface between the equipment of the subscriber and the equipment of the service provider. The UNI runs service-level data-, control-, and management-plane functions on both client and network sides. An EVC is described as a set of frame streams sharing a common forwarding treatment and connecting two or more UNIs [10].

In its *Ethernet services definitions*, the MEF advances three EVC connections: PtP, multipoint-to-multipoint (MPtMP), and point-to-multipoint (PtMP) [10]. These correspond to the three service types shown in Fig. 4 and described below.

An E-LINE is a PtP service connecting two UNIs. Two implementations are proposed by the MEF: Ethernet private line (EPL) and Ethernet virtual private line (EVPL). An EPL replaces a TDM private line and uses dedicated UNIs for PtP connections, whereas an EVPL uses UNIs with EVC-multiplexing capabilities to replace services such as frame relay (FR) and ATM.

An E-LAN is an MPtMP service offering full transparency to customer control protocols and VLANs (transparent LAN service [TLS]). As for E-LINEs, the two E-LAN categories are Ethernet private LAN (EPLAN) and Ethernet virtual private LAN (EVPLAN). Similar to EPON, the E-TREE service offers PtMP connectivity from a root UNI to the leaf UNIs and MPtP connectivity from the leaves to the root.

The MEF specifications further associate several service attributes to UNIs and EVCs [10]:

- The **Ethernet physical interface** determines the PHY/MAC sublayer features such as speed and physical interface.
- The **bandwidth profile** is a set of five traffic parameters that characterize the connection, namely committed information rate (CIR), committed burst size (CBS), excess information rate (EIR), excess burst rate (EBS), and color mode (CM). The CM is a binary parameter indicating whether a UNI employs the MEF color-marking system discussed below.
- MEF-10.1 includes definitions of frame delay, jitter, loss, and service availability for PtP and multipoint EVCs. Together, these quantities form the **performance parameter** attributes of EVCs.
- The **class of service (CoS)** attribute is a frame prioritization scheme based on the physical port, 802.1p priority bits, or higher-layer service-differentiation methods.

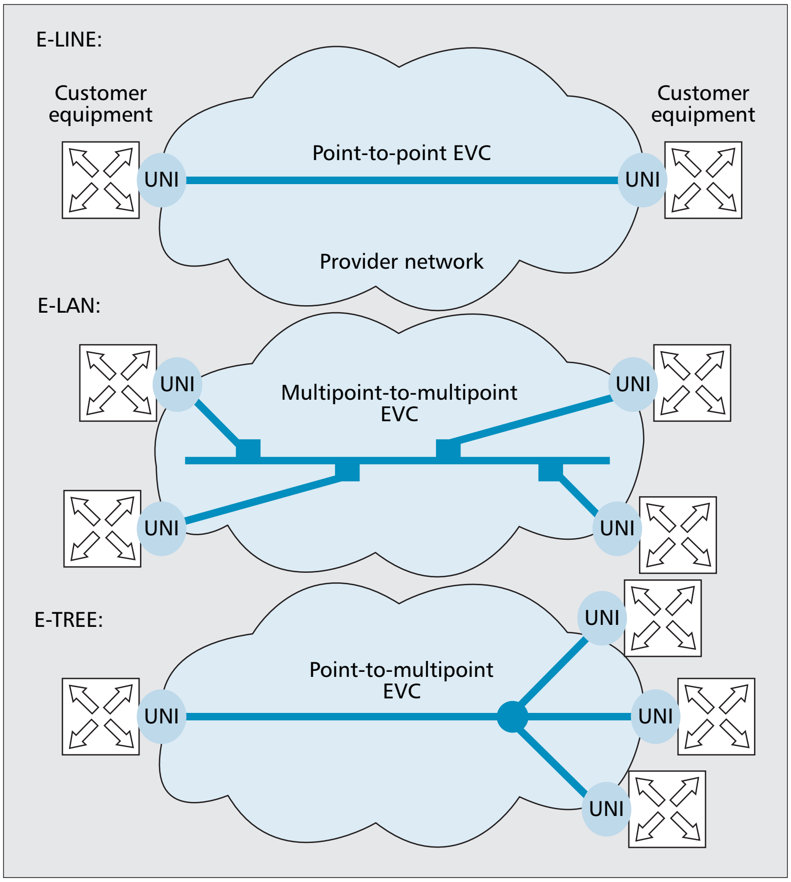


**Figure 4.** *Metro Ethernet Forum (MEF) service definitions: E-LINE, E-LAN, and E-TREE.*

- The **service frame delivery** attribute determines whether client data and service frames transmitted over an EVC are unicast, multicast, or broadcast. In addition, this attribute specifies whether client layer-2 control frames are processed, forwarded, or discarded.
- **VLAN tag support** determines whether the UNI supports the various Q-tag fields.
- The multiplexing capability of a UNI is indicated by the **service multiplexing** attribute.
- **Bundling** establishes a mapping between customer VLAN IDs (C-VIDs) and EVCs whereby a single EVC can carry more than one VLAN. All-to-one bundling is further defined as a binary parameter mapping all customer VLANs to one EVC.
- **Security filters** represent the frame filtering attributes of a UNI. For instance, a UNI may restrict access to source MAC addresses within an access control database [10]. A description of the current security vulnerabilities of STP-based Ethernet networks can be found in [8].

The QoS requirements for CGE are detailed within the bandwidth profile attribute specification in MEF-10.1. The set of traffic parameters defining the bandwidth profile (CIR, CBS, EIR, EBS, and CM) are controlled and enforced by a two-rate, three-color marker (trTCM) algorithm that is run at the UNI. Input frames are marked green, yellow, or red, using a token bucket model. Whereas green frames are delivered and red frames are discarded, yellow frames are delivered only if excess bandwidth is available [10, 11]. Further possible QoS and fairness control techniques based on layer-2 frame marking are discussed in [11].

In conjunction with security filters and service-frame delivery attributes, the bandwidth profile algorithm enables traffic engineering through traffic policing. Other traffic-engineering mechanisms such as traffic shaping and load balancing are enabled by the described attributes. By introducing connection-oriented networking models, the IEEE 802.1Qay (PBT) standard provides a concrete embodiment of MEF traffic-engineering requirements. Nevertheless, the standardization of the CGE traffic-engineering infrastructure still is ongoing. According to [12], the current standardization initiatives still do not fully address the stringent traffic management requirements of future applications such as IPTV.

## RESILIENCE AND OAM

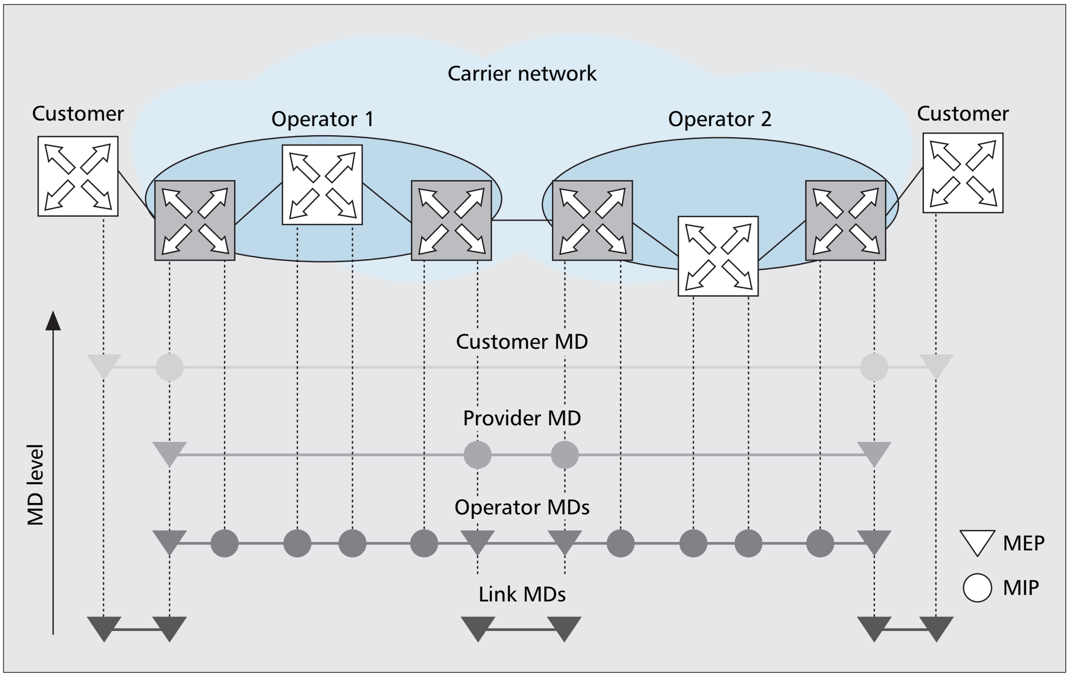
Resilience is defined as the ability of a network to detect faults and recover from them [8]. The carrier-grade resilience reference is set by SONET/SDH technology, with sub-50–ms recovery times. Such performance levels require a strong OAM framework comprising fault and performance management. Fault management includes failure detection, localization, notification, and recovery mechanisms, whereas performance management aims at monitoring and reporting network performance metrics such as throughput, frame loss, and bit error rates.

Through coordinated initiatives, IEEE, ITU-T, and MEF developed a number of standards addressing carrier-grade resilience and OAM. Whereas IEEE 802.3ah (Ethernet in the first mile [EFM]) specifies link-level OAM processes such as automatic discovery, IEEE 802.1ag (connectivity fault management [CFM]) defines end-to-end VLAN-level OAM functions (described below). At the ITU-T, Study Group 13 introduces a broader set of OAM functions within Y.1731 (OAM functions and mechanisms for Ethernet-based networks). In addition, the completed G.8031 (Ethernet protection switching) of Study Group 15 and the ongoing G.8032 (Ethernet ring protection switching) initiatives focus on VLAN-level protection mechanisms. Focusing on the service-level, the MEF specifications (MEF-15, 16, and 17) highlight OAM requirements related to SLA performance and edge-node management functions.

CFM and Y.1731 specify layer-2 OAM functions designed for connection-oriented settings (e.g., PBT) and are compatible with existing link-level EFM processes. The remainder of this section describes those functions, as well as the proposed Ethernet restoration mechanisms.

### OAM ARCHITECTURE AND MECHANISMS

CFM and Y.1731 introduce a hierarchical architecture where eight management levels enable customers, service providers, and network operators to run OAM processes in their own maintenance domains (MDs). The edge nodes of the various nested MDs are called maintenance end points (MEPs) and initiate OAM mechanisms,



■ **Figure 5.** *Illustration of OAM MDs in CFM.*

*The OAM frame format defined by CFM and Y.1731 is an Ethernet frame with a data field partitioned into OAM-specific fields. The latter include an MD-level field and an operation code (OpCode) associating an OAM frame with one OAM function. Unless specified, an OAM frame does not pass its domain boundaries.*

whereas intermediate nodes (maintenance intermediate points [MIPs]) respond to them. The example of Fig. 5 shows a PtP EVC service delivered by a provider over two adjacent operators. Although the customer and provider MDs support the PtP connection from end to end, each operator establishes a distinct MD over its segment of the EVC below the provider MD-level. The nested MDs run their OAM functions independently.

The OAM frame format defined by CFM and Y.1731 is an Ethernet frame with a data field partitioned into OAM-specific fields. The latter include an MD-level field and an operation code (OpCode) associating an OAM frame with one OAM function. Unless specified, an OAM frame does not pass its domain boundaries.

CFM specifies three fault management functions:

- **Continuity check (CC)**: MEPs within an MD multicast to each other periodic CC frames. CC messages can be used to detect loss of connectivity or network misconfiguration and to measure frame loss.
- **Link trace (LT)**: An LT request frame sent by a MEP toward a target node triggers LT reply frames from all intermediate nodes back to the source MEP. This procedure enables fault localization and the monitoring of network configuration.
- **Loopback (LB)**: An LB message (or MAC ping) sent by a MEP to any node triggers an LB reply. This process is used to check the responsiveness of intermediate nodes and verify bidirectional connectivity.

The following are some of the functions introduced within Y.1731:

- **Alarm indication signal (AIS)** messages are used to notify nodes that a fault was reported to the network management system (NMS). AIS suppresses further alarms within the MD and at higher MD levels.
- Due to their configurable test data, **test** frames can be used to measure throughput, frame loss, and bit error rates.
- A **locked (LCK)** signal indicates to higher MD levels that maintenance operations are taking place at a MEP, thus suppressing false alarms.
- The **maintenance communication channel (MCC)** function sets up a channel for vendor-specific OAM applications such as remote maintenance.
- The **experimental/vendor specific OAM (EXP/VSP)** frame types are unspecified and reserved for temporary or vendor-specific OAM extensions.

In addition, Y.1731 defines loss, delay, and delay-variation (jitter) measurements using appropriate OAM functions.

### ETHERNET PROTECTION AND RESTORATION

In Ethernet LANs, the family of STP protocols performs protection and restoration functions through topology reconfiguration. The associated recovery times vary from 1 s to 60 s [8] and fall short of carrier- grade requirements. Besides, the loop prevention mechanisms of xSTP are no longer required in the CGE connection-oriented settings.

G.8031 specifies SONET/SDH-style 1 + 1 unidirectional and 1 + 1 or 1:1 bidirectional protection switching for PtP paths or segments. To coordinate bidirectional protection switching, G.8031 includes an Automatic Protection Switching (APS) protocol. APS uses OAM frames identified by a specific OpCode.

In PBT, end-to-end protection switching between edge PBB nodes is accomplished by provisioning a protection path for each working path. Loss of connectivity along one path automatically triggers the source PBB to replace the working-path B-VID with the protection-path B-VID in outgoing frames [7]. The pre-configuration of adequate protection paths is left to the NMS.

A number of alternative proprietary and studied mechanisms are mentioned in [8, 11], usually based on redundancy mechanisms or STP enhancements.

The co-existence of various protection and restoration mechanisms requires careful definition and prioritization. Network failures usually are resolved faster and more efficiently within the layer where they occur [5]. In CFM, accordingly, cross-level fault notifications from lower MD levels have higher priority.

*As long as Ethernet dominates LAN technology, native transport enhancements will translate into cost and complexity reductions compared to hybrid solutions.*

## Conclusion

The targets of the current CGE evolution are the connection-oriented architectures and control tools established through SONET/SDH, MPLS, and IP. Although the complexity of such amendments departs from the trademark simplicity of Ethernet, powerful drivers back the transition. Indeed, as long as Ethernet dominates LAN technology, native transport enhancements will translate into cost and complexity reductions compared to hybrid solutions.

Whether that economic advantage outweighs leveraging deployed MPLS equipment is a legitimate issue. Nonetheless, the effort to advance CGE solutions is merely starting. The recent IETF generalized-MPLS Ethernet label switching framework draft, for example, aims at giving Ethernet the advanced control features of GMPLS, such as fast reroute (FRR). In contrast to the protection-based resilience procedures described above, FRR represents a powerful restoration mechanism designed to achieve carrier-grade restoration times (tens of milliseconds) in a more bandwidth-efficient fashion through the activation of pre-computed alternate routes.

## References

[1] R. Breyer and S. Riley, “Switched, Fast, and Gigabit Ethernet,” *New Riders*, 3rd ed., 1998.

[2] J. Hurwitz and W. Feng, “End-to-End Performance of 10-Gigabit Ethernet on Commodity Systems,” *IEEE Micro*, vol. 24, no. 1, Jan. 2006, pp. 10–22.

[3] R. Ramaswami, “Optical Networking Technologies: What Worked and What Didn’t,” *IEEE Commun. Mag.*, vol. 44, no. 9, Sept. 2006, pp. 132–39.

[4] P. A. Bonenfant and S. M. Leopold, “Trends in the U.S. Communications Equipment Market: A Wall Street Perspective,” *IEEE Commun. Mag.*, vol. 44, no. 2, Feb. 2006, pp. 141–47.

[5] A. Kirstädter et al., “Carrier-Grade Ethernet for Packet Core Networks,” *Proc. Asia Pacific Optical Commun. Conf. (SPIE Conf. Series)*, South Korea, Oct. 2006, vol. 6354.

[6] J.-L. Ferrant et al., “Synchronous Ethernet: A Method to Transport Synchronization,” *IEEE Commun. Mag.*, vol. 46, no. 9, Sept. 2008, pp. 126–34.

[7] D. Allan et al., “Ethernet as Carrier Transport Infrastructure,” *IEEE Commun. Mag.*, vol. 44, no. 2, Feb. 2006, pp. 134–40.

[8] M. Huynh and P. Mohapatra, “Metropolitan Ethernet Network: A Move from LAN to MAN,” *Comp. Net.*, vol. 51, no. 17, Dec. 2007, pp. 4867–94.

[9] G. Parsons, "Ethernet Bridging Architecture," *IEEE Commun. Mag.*, vol. 45, no. 12, Dec. 2007, pp. 112-19.

[10] A. Kasim, "Carrier Ethernet," Ch. 2, *Delivering Carrier Ethernet: Extending Ethernet beyond the LAN*, McGraw-Hill, 2007, pp. 45–104.

[11] A. Iwata, "Carrier-Grade Ethernet Technologies for Next Generation Wide Area Ethernet," *IEICE Trans.*, vol. 89-B, no. 3, Mar. 2006, pp. 651-60.

[12] S. Vedantham, S.-H. Kim, and D. Kataria, "Carrier-Grade Ethernet Challenges for IPTV Deployment," *IEEE Commun. Mag.*, vol. 44, no. 7, July 2006, pp. 24-31.

## Biographies

KERIM FOULI (fouli@emt.inrs.ca) is a Ph.D. student at INRS. He received his B.Sc. degree in electrical engineering at Bilkent University in 1998 and his M.Sc. degree in optical communications at Laval University in 2003. He was a research engineer with AccessPhotonic Networks, Quebec City, Canada, from 2001 to 2005. His research interests are in the area of optical access and metropolitan network architectures with a focus on enabling technologies. He is the recipient of a two-year doctoral NSERC Alexander Graham Bell Canada Graduate Scholarship for his work on the architectures and performance of optical coding in access and metropolitan networks.

MARTIN MAIER (maier@ieee.org) was educated at the Technical University of Berlin, Germany, and received M.Sc. and Ph.D. degrees, both with distinction (summa cum laude), in 1998 and 2003, respectively. In the summer of 2003 he was a post-doctoral fellow at the Massachusetts Institute of Technology, Cambridge. Since May 2005 he has been an associate professor at Institut National de la Recherche Scientifique (INRS), Montreal, Canada. He was a visiting professor at Stanford University, California, October 2006 through March 2007. He is the founder and creative director of the Optical Zeitgeist Laboratory (http://www.zeitgeistlab.ca). His research aims at providing insights into technologies, protocols, and algorithms that are shaping the future of optical networks and their seamless integration with broadband wireless access networks. He is the author of the books *Optical Switching Networks* (Cambridge University Press, 2008) and *Metropolitan Area WDM Networks: An AWG-Based Approach* (Kluwer Academic, 2003).