Optical Ethernet: Making Ethernet Carrier Class for Professional Services

JIYANG WANG

Invited Paper

The existing overlaid data network architecture deployed by most service providers has shown its shortcomings in supporting professional business services due to its complexity and high cost. This paper introduces a new optical transport technology that is based on Ethernet but integrates all the required features to consolidate multiple layers below the IP layer in the existing architecture into one, thus simplifying the architecture and significantly reducing the cost both in network buildout and in network operations. The paper provides the technical details on how the limitations of traditional Ethernet are overcome and how Ethernet becomes carrier class. It also introduces the new services enabled by carrier-class Ethernet and analyzes the impact of it on Internet evolution.

Keywords*—*Carrier class, Ethernet, metropolitan area network (MAN), network, optical, services.

I. Introduction

About two-thirds of service provider revenues come from business customers, but these revenues have declined dramatically over the last three years, according to a report by RHK, South San Francisco, CA, published in September 2003 [1]. The major obstacle for service providers in offering affordable and profitable high bandwidth to business customers lies in the high cost of today's metro networks that are based on an overlaid architecture shown in Fig. 1.

The four layers in this architecture implement different functions. Wavelength-division multiplexing (WDM)/fiber is used to address the needs for high bandwidth. The optical transport is usually Synchronous Optical Network/Synchronous Digital Hierarchy (SONET/SDH) based, and it ensures reliability by automatic protection switching. SONET/SDH is also widely used to provide leased lines that may carry both data and voice traffic. The asynchronous transfer mode (ATM) layer is built on top of SONET/SDH to achieve traffic-engineering goals, increase the bandwidth utilization of SONET/SDH, and provide layer 2 virtual private network (VPN) service. The idea of converging all types of traffic, including data, voice, and video, onto ATM has been given up because of the complexity of ATM. Then comes the IP layer, which carries most of the applications today and facilitates layer 3 VPN service for business customers.

This architecture has many drawbacks. The three major ones are the complexity in management and service provisioning that involves network planning on multiple layers, very high capital expenditure (CapEx) and operational expenditure (OpEx), and the difficulty in scalability.

This would illustrate the dilemma service providers are struggling to get out of. On one side, business customers are reluctant to buy more bandwidth because they cannot afford it. On the other side, service providers cannot further lower data service price because they are already on the edge of losing money. Service providers are now exploring innovative data transport technologies and network architectures to break up the deadlock. The new data transport technology will allow service providers to simplify the network architecture, thus reducing the CapEx and OpEx. A major move taken by many service providers is to eliminate the ATM layer and connect their routers by SONET/SDH using "packet over SONET/SDH" (PoS) [15], [16]. The reason is simple and straightforward. If most applications are carried by IP, not by ATM, ATM is not so needed. Also, it is very inefficient to carry IP packets by ATM cells. It is much more efficient to map IP packets directly to SONET/SDH frames, i.e., PoS.

But PoS still has the following drawbacks in supporting data services.

- PoS is a packet-to-SONET mapping technology using high-level data link control (HDLC)-like framing that has very low transmission efficiency for IP packets due to the cost of the overhead in frame translation and the use of escape characters and byte stuffing.

Manuscript received October 24, 2003; revised March 11, 2004.
The author is with Atrica, Inc., Paris 75017, France (e-mail: Jiyang_Wang@atrica.com).
Digital Object Identifier 10.1109/JPROC.2004.832953

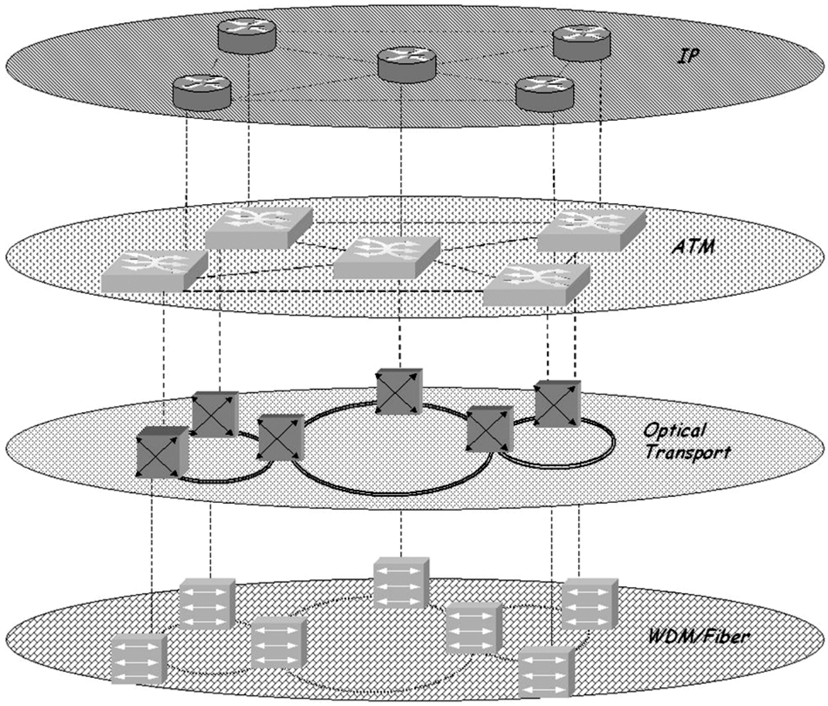

Fig. 1. Current data network architecture.

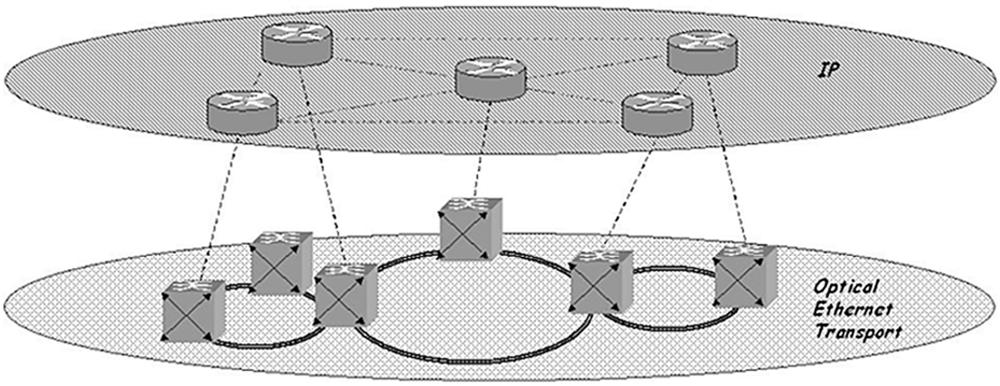

Fig. 2. New data network architecture.

- Byte stuffing of the PoS framing process interferes with the quality of service (QoS) of IP traffic, and it requires unpredictable expansion of required bandwidth.
- SONET/SDH is not efficient for IP packet transport because a big SONET/SDH pipe is sliced into small circuits by time-division multiplexing (TDM), each of which has a fixed bandwidth. Even if some of the circuits do not have any data traffic at a time, their bandwidth cannot be utilized by the data traffic in other circuits. For transporting IP packets efficiently, statistical multiplexing is needed.
- The PoS interface cards of routers are very expensive. For example, a one-port OC-48 PoS card with 1310-nm transceiver may cost $30 000, twice the price of a three-port gigabit Ethernet card for the same router. Moreover, a router may need multiple PoS interfaces to connect to other routers, because a SONET circuit is point to point. Channelized PoS is available, but it is even more expensive. For example, a one-port channelized OC-12 card costs $45 000. A higher rate channelized card is unimaginably expensive.

A new network architecture, discussed in this paper, is shown in Fig. 2.

The major change is the introduction of a new transport layer that consolidates the three layers under the IP in the old architecture shown in Fig. 1. This new transport layer must incorporate effectively all the functions and benefits that the old architecture has, which can be summarized as:

- IP-like network scalability and data friendliness;
- ATM-like QoS and traffic engineering (TE) capabilities;
- SONET/SDH-like reliability and operation, administration, and management (OAM);
- WDM-based bandwidth capacity;
- support of leased line service and layer 2 VPN service.

Last but not least, it must be much cheaper than ATM and SONET/SDH.

Optical Ethernet has emerged as the most important option for metro network transport for the following reasons.

- Ethernet is the most economically viable alternative today.
- Simplicity. Most of the IP traffic starts from Ethernet (e.g., a user's desktop) and ends at Ethernet (e.g., Web server). If there is no frame translation or mapping in between, network architecture is simplified.
- Well-defined standards provide multivendor interoperability.
- Flexible "facility-free" bandwidth provisioning from <1 Mb/s to 1 Gb/s.
- Scalability to 10-Gb/s bandwidth over greater distances from 10 to 40 km over dark fiber.

It may still sound strange that Ethernet, a technology that just passed its 30-year anniversary in 2003 and has long been perceived only as a good fit for LANs, is now becoming the hot spot of today's metro networks. Indeed, concerns and doubts are surrounding Ethernet becoming a solution for public networks even though Ethernet is 85% cheaper than SONET/SDH on a cost-per-Mb/s basis [3]. Traditional Ethernet, if used in public network, lacks some key carrier-class features and capabilities such as end-to-end QoS that is needed for professional services, scalability in supporting a large number of business customers, reliability with fast protection, integrated support for OAM functions, as well as the support for legacy TDM-based devices. Ethernet must revamp itself to be carrier class before it can be massively deployed in public networks as the new optical transport layer for IP. Much work has been done in the last three years—new standards are set, many drafts are under discussion, and an industry forum that promotes Ethernet services has been formed.

The next section focuses on the enhancements that have been developed to make Ethernet carrier class. This section is divided into seven parts, each of which covers one enhancement, respectively: connection orientation, QoS, protection, Ethernet OAM, network management and service provisioning, scalability, and circuit emulation over Ethernet. Technical details are given and the relevant standards and draft proposals are introduced. Section III summarizes the advantages of optical Ethernet. The section also makes a prediction that last mile access networks will converge to optical Ethernet in the near future. A conclusion is presented at the end of this paper.

II. What Is Carrier-Class Optical Ethernet?

Carrier-class optical Ethernet is a data-oriented technology that has kept the Ethernet media access control (MAC) layer unchanged but uses fibers as its physical media, and it implements seven major enhancements to enterprise-class Ethernet switches in order to become carrier class:

- connection orientation;
- end-to-end QoS and TE;
- fast protection and restoration;
- Ethernet OAM and manageability;
- scalability;
- fast and easy service provisioning;
- TDM circuit emulation over Ethernet.

A. *Connection Orientation*

Just like SONET/SDH, frame relay and ATM, optical Ethernet uses connection as the basic entity for customer service. An Ethernet connection or Ethernet virtual connection (EVC) is similar to an ATM or frame relay virtual connection. A connection identifies and separates traffic flows so that customer privacy is protected. Other benefits of using connection include the following.

- Many enterprises have been accustomed to the way they use frame relay services, and they want to keep this way unchanged when they use Ethernet services.
- As QoS for business customers must be end to end, the process of establishing connections naturally facilitates the end-to-end QoS mechanisms through TE.
- It is easy to implement fast protection because protection connection can be preprovisioned together with the primary connection.
- It is easy to manage services for service establishment, service troubleshooting, QoS monitoring, and billing.

EVC can be implemented in two ways, either by using VLAN or by using MPLS. The most popular way is to combine the two, i.e., VLAN is used at the network edge to keep the customer edge (CE) device as simple and cost-effective as possible, and MPLS is used in the network core to address the scalability and other issues such as 50-ms protection and TE. Although the provider edge (PE) device that sits between the CE device and the core device (the P device) needs to map a VLAN-based connection to an MPLS-based connection [i.e., label-switched path (LSP)], this mapping operation is invisible either to operators or to business users so that the connection is in fact end to end as an integral entity. Connections are also the basis for layer 2 VPN services. Fig. 3 shows a typical optical Ethernet metro network and its protocol stack evolution with an EVC established between two access devices.

While the connectionless technologies such as conventional Ethernet switching or IP routing work fine for LAN and Internet for best-effort traffic, connection-oriented technologies are definitely needed by carriers to provide professional services to business customers.

Connection-oriented technology was supposed to have scalability issues because network devices need to maintain status information of all the connections traversing them and there may be hundreds of thousands of connections in the network. But EVC, as a service entity, needs to remain static unless there are failures over it, so there is no need for updating the status of EVC dynamically or periodically. Maintaining a large number of EVCs is not a technical issue; instead the cost of building a large connection table is the major consideration.

While the data plane of optical Ethernet is on layer 2, the control and management plane can be on layer 3. This is particularly true when MPLS is used.

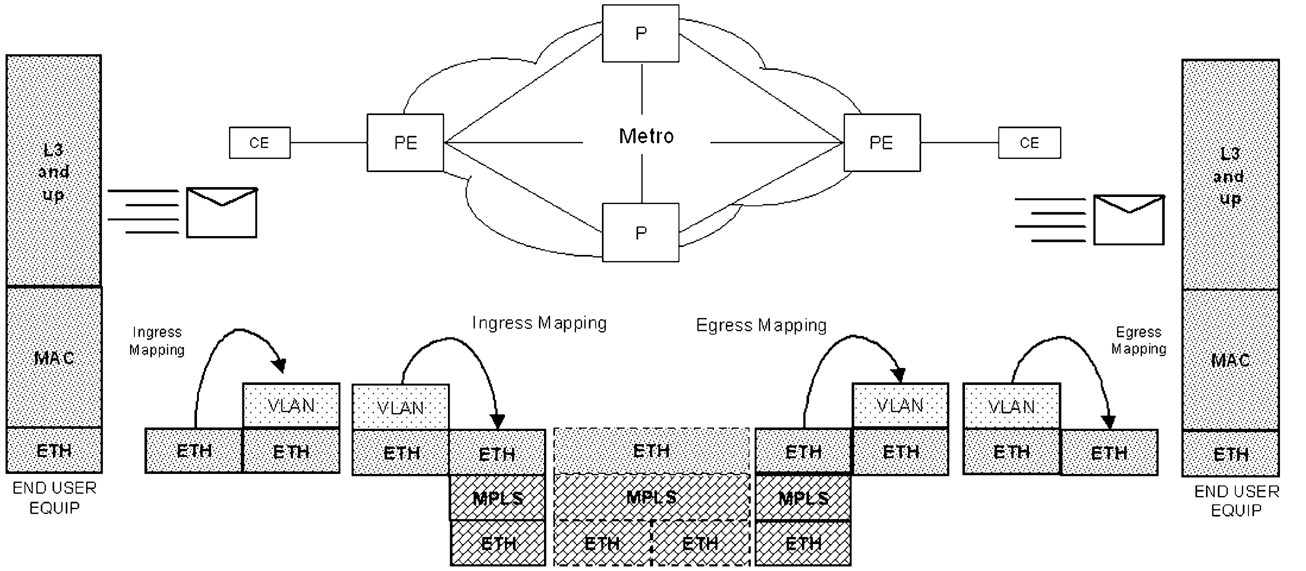

Fig. 3. Typical optical Ethernet and its protocol stack.

B. End-to-End QoS and TE

Even though it is becoming comparatively cheap to build an optical Ethernet metro network with very high bandwidth, stringent QoS control is still needed to enable multimedia and mission-critical applications for business customers because of the bursty nature of Ethernet traffic and the fact that many carriers tend to overbook the bandwidth.

In developing optical Ethernet, one advantage that equipment vendors can utilize is that many IP QoS mechanisms can be applied to optical Ethernet because of the fundamental similarities between Ethernet and IP such as variable frame/packet size, statistical multiplexing, and the store-and-forward mechanism. Unlike ATM, optical Ethernet does not need to create its own QoS architectures and mechanisms.

Despite many debates over various IP QoS architectures and mechanisms, such as IntServ versus DiffServ and Resourse Reservatuib Protocol (RSVP)-TE versus DiffServ-Aware TE (DS-TE), optical Ethernet combines the elements from all of them, and its QoS mechanism is very much service oriented.

To define Ethernet-based professional services that can be acknowledged and implemented by the whole telecom industry, the Metro Ethernet Forum (MEF) was founded in July 2001. MEF has 62 members that include many service providers and almost all the major equipment vendors who share common interests in metro Ethernet. MEF has defined two types of Ethernet services [4], i.e., point-to-point Ethernet line (E-Line) and multipoint Ethernet LAN (E-LAN). Both E-Line and E-LAN are layer 2 VPN service because they are transparent to layer 3 protocols. These services should have the traffic parameters of committed information rate (CIR) and its corresponding burst size (BS), and excess information rate (EIR) and its corresponding burst size (EBS), and a traffic class that can be classified by physical interface, customer VLAN ID, IEEE 802.1p bits, IP types of service (TOS) or DiffServ differentiated service code point (DSCP) bits, or MPLS experimental bits (EXP bits) in the packets.

To implement the traffic attributes defined by MEF, the QoS mechanism of optical Ethernet consists of the following three essential building blocks.

1) Call Admission Control (CAC) and TE: CAC ensures the availability of required bandwidth for CIR over the selected path for an EVC. In optical Ethernet, CIR is not overbooked, but EIR can be and usually is overbooked. If no paths in the network that meet customer's bandwidth (CIR) requirement exist, the request for establishing a new EVC is rejected by CAC. The service provisioning system will inform the operator of the reasons for the rejection.

The Open Shortest Path First (OSPF)-TE [5] is a major protocol widely deployed by optical Ethernet to discover bandwidth information on each link in an MPLS-based core network and to calculate the best path that meets the bandwidth requirement (constraint-based routing). The LSP is established either by the two PE devices at both ends of the LSP using signaling protocols on control plane or configured by the network management system (NMS) on the management plane (NMS usually knows the topology and bandwidth information and is able to calculate the path). The most widely used signaling protocol is RSVP-TE [6].

So CAC and TE deployed by an optical Ethernet core network is rarely different from those deployed by IP networks. But there is one thing that optical Ethernet particularly needs to address. The CAC between CE and PE cannot utilize MPLS-based TE protocols because CE devices do not support MPLS for the sake of low cost. One solution is, of course, to extend MPLS to CE, which will make a CE device much more expensive than a conventional Ethernet switch. Another solution is to use NMS to perform CAC between CE and PE. This is a viable solution because anyway NMS needs to maintain all the information about the whole network and connections and that information may include bandwidth used and still available on the links between CE and PE.

8	7	6	5	4	3	2	1
User Priority			CFI	VLAN Identifier (4 bits)			
VLAN Identifier (8 bits)							

Fig. 4. IEEE 802.1Q VLAN tag format.

Table 1
802.1p Bits Assignment With Integral DE Marking

Value	Assignment	Application Example	Note
111	Delay and Jitter Sensitive	Voice and CES	No frame dropping
110	Control and Management	Signaling messages	No frame dropping
101	Delay Sensitive DE=1	Real-time video	Subject to discard
100	Delay Sensitive DE=0	Real-time video	Not discarded
011	Mission Critical DE=1	ERP	Subject to discard
010	Mission Critical DE=0	ERP	Not discarded
001	Normal DE=1	E-mail	Subject to discard
000	Normal DE=0	E-mail	Not discarded

2) Traffic Policing and Marking: Traffic policing meters the average rate of an ingress traffic flow against the BS for each incoming Ethernet frame belonging to that traffic flow and marks the frames based on the results of the metering. The dual token buckets algorithm is widely used for traffic metering. Based on the metering result, a frame will be manipulated in the following way.

- If Traffic Rate<=CIR, the frame will be forwarded without experiencing congestion.
- If EIR>=Traffic Rate > CIR, the frame will be marked as discard eligible (DE). This frame will pass through the network if there is no congestion. When there is congestion, it may or may not be discarded, depending on its dropping precedence discussed below.
- If Traffic Rate > EIR, the frame is dropped.

Although the token bucket algorithm is very straightforward and easy to implement by software, it has to be implemented by hardware because there may be thousands of EVCs per CE interface and each EVC needs a policer.

Another challenge is that unlike ATM, Ethernet has not defined DE bit(s) in its frame. There are now discussions in the MEF and the IEEE 802.1 working group on a standard way of DE marking. One idea is to use some of the eight values of 802.1p when not all the eight priorities are used. Another idea is to use the canonical format indicator (CFI) bit.

The format of the IEEE 802.1Q VLAN tag is shown in Fig. 4.

The CFI bit was intended to encapsulate token ring frames in Ethernet, and Ethernet devices do not use this bit. So it is possible to use the CFI bit to indicate DE marking. The advantage of using the CFI bit is that all the eight priorities are kept. But there are two disadvantages. One is that very few Ethernet vendors think about processing the CFI bit, so most existing products need a hardware change if they have to process the CFI bit. The other is that in MPLS over an Ethernet cloud, the link layer header (i.e., the Ethernet header in this paper) that precedes the MPLS shim header may or may not contain the VLAN tag, according to Internet Engineering Task Force (IETF) Request for Comments 3032 [9]. In reality, that VLAN tag usually does not exist because it is useless and simply introduces overhead. This is also the reason for the provider VLAN tag being stripped by the ingress PE device (refer to Fig. 3). In order to use the CFI bit for DE marking, a VLAN tag needs to be added to the link layer header, or the provider VLAN tag needs to be kept in the network core. This will introduce more overhead to the frames and does not justify using two octets of the VLAN tag (the tag control word) for just one bit.

As such, many vendors tend to use 802.1p bits for DE marking. One more reason is that 802.1p bits can be easily mapped to the EXP bits in the MPLS shim header because they have the same number of bits and, very importantly, MPLS EXP bits are also used for indicating priority classes of traffic. Table 1 shows an example of the arrangement of 802.1p/MPLS EXP bits for various classes of traffic and DE marking.

3) Traffic Queuing and Congestion Control: The design of queuing and congestion control mechanism is determined by traffic policing and the DE marking mechanism. Take Table 1 as an example. Each network device should have five priority output queues on each interface. The queue scheduling must be consistent with the frame dropping policy. An example of such a scheduling mechanism may consist of three schedulers using different scheduling policies [Strict Priority Queuing (SPQ) and Weighted Round Robin (WRR)], as shown in Fig. 5.

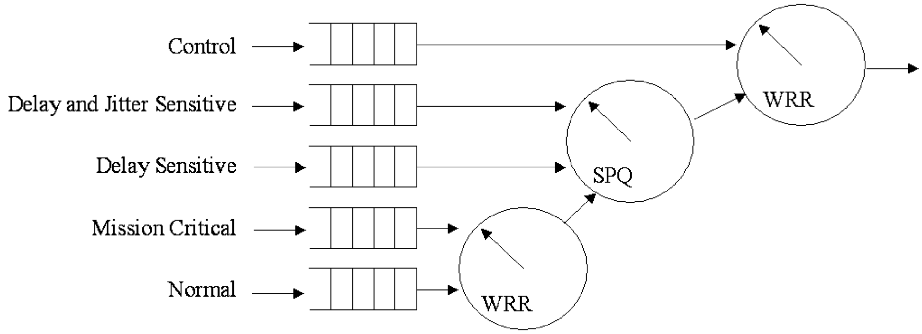

Fig. 5. Example of traffic queuing mechanism.

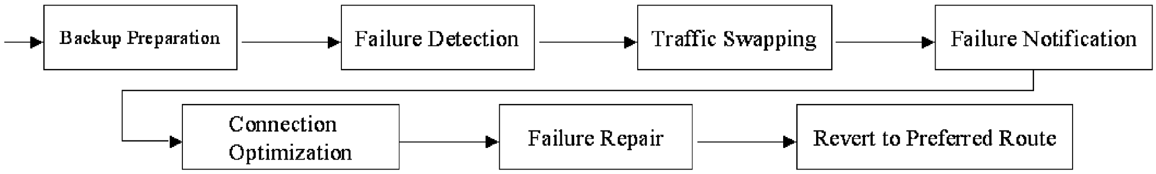

Fig. 6. Procedures/actions of a protection mechanism.

The last WRR scheduler is used to allocate to control messages a small amount (say, 1% of the whole bandwidth on a link) but guaranteed bandwidth. The SPQ scheduler that takes effect on all the other four classes of traffic ensures that such traffic as voice and circuit emulation service (CES) experiences minimum delay and jitter and no frame loss. It also ensures that delay-sensitive traffic is forwarded in precedence to mission-critical and normal traffic, while the last two classes of traffic can be configured with a ratio of the bandwidth left for them by another WRR scheduler.

As strict priority queues may make other queues starve, service provisioning system can enforce a bandwidth bound to the classes of traffic that enjoy SPQ. The effect is similar to that of DS-TE that aims at CAC on a per-class basis. CAC will use this bandwidth bound as the criterion to determine whether a request for a new connection to be used by high-priority traffic should be granted or rejected.

Various queue management algorithms can be used to control congestion. Congestion is suggested by the fullness of the queues. Since the incoming frames have been colored by the traffic policers at the entry point (i.e., a CE device) of the network, the frames with the DE bit marked are subject to being discarded when their corresponding queues are going to be full. Instead of tail dropping, weighted random early detection (WRED) can be used to improve system performance. In this case, the priority bits value in Table 1 also implies dropping precedence.

C. *Fast Protection*

Fast protection is critical for business continuity and should be provided by optical Ethernet for professional services. Nowadays optical Ethernet runs at 1 Gb/s to 10 Gb/s on each link and a short period of service downtime would lead to a tremendous frame loss that significantly affects service quality, especially for voice and Transmission Control Protocol (TCP)-based applications [10].

The target of the protection mechanism of optical Ethernet is to achieve SONET-like fast protection and restoration within 50 ms over any network topology. Spanning tree protocol over VLAN or reestablishing LSP by routing and signaling protocols after failure involve software implementations and their convergence time is far more than 50 ms and is not deterministic, which depends on the complexity of network topology.

A good protection mechanism should reach the following goals:

- fast recovery speed;
- efficient resource utilization;
- no or little human intervention and manual configuration;
- The service level agreement (SLA) can be retained when protection takes effect.

Protection has many aspects, and their designs affect the achievement of the above goals. A protection mechanism may include the following procedures/actions, illustrated in Fig. 6.

The arrangement of each procedure can be different depending on how the protection mechanism will be implemented. For example, Fig. 7 depicts another mechanism.

If the first step is to prepare the backup of primary connection, the protection connection (or backup LSP in an MPLS cloud) will be preprovisioned. If it is after failure detection and/or failure notification, as shown in Fig. 7, the backup LSP is established dynamically either by an ingress node or by an intermediate node local or remote to the point of failure after the failure occurs. The last two steps in Fig. 7 may not exist in some protection mechanisms. More extensive discussions on various protection mechanisms are beyond the scope of this paper and can be found in [11].

To reach the four objectives described previously, carrier-class optical Ethernet usually implements MPLS Fast Reroute [12] in hardware to protect against link and/or node failure in less than 50 ms, as well as end-to-end protection switching to retain the SLA and TE. While there are several flavors of MPLS Fast Reroute with various capabilities and characteristics, the mechanism of optical Ethernet has the following characteristics.

- Protection is a service, so an EVC may or may not have automatic protection upon the customer's request. Customers can choose bronze protection (manual reestablishment of the connection from NMS, which may take minutes or hours, depending on the SLA), silver protection (less than 10 s), or golden protection (less than 50 ms).
- Silver protection is implemented by end-to-end protection switching, which means a protection connection is established together with the primary connection, and the resource is reserved for the protection connection with the same QoS profile as the primary connection. Protection switching-based protection is also called end-to-end protection.
- Golden protection is implemented by Fast Reroute in addition to end-to-end protection. Fast Reroute can switch traffic over to the detour path of the failed link or node in less than 50 ms. Fast Reroute-based protection must be used together with end-to-end protection to retain the SLA.

The protection procedure of silver protection and golden protection follows the one shown in Fig. 6.

Backup preparation includes the preprovisioning of the end-to-end protection connection, the discovery of the detour path for each link and node that is on the primary path, and the establishment of the protection tunnel via RSVP-TE [13]. The protection connection is set up in the same way as the primary connection but on the residue topology after excluding the paths that do not have enough available bandwidth for the requested CIR. In order to have the least overlapping with the primary path, the secondary path is calculated with the cost of the primary path being increased significantly.

The protection tunnel on the detour path is not allocated any resources for the following reasons.

- Resource saving. Otherwise, less CIR can be allocated to subscribers and resource utilization will be very low when there is no failure (99.9% of the time).
- Keeping track of TE. Otherwise, since the protection tunnel is set up by the local nodes themselves, it may be troublesome to inform NMS of the bandwidth allocated for protection tunnel on the detour path.
- Failure is short lived, so it should be acceptable if the QoS of low-priority traffic is not in conformance to SLA for a short period (about 2 s).

When a failure is detected, traffic of all the affected connections will first be swapped to the protection tunnel (if a detour path exists) within 50 ms by Fast Reroute. Then the detour path may be overloaded and the SLA of all the connections on the detour path may be violated. In order to guarantee the SLA for high-priority traffic (which is usually delay and frame loss sensitive), the SLA of low-priority traffic will have to be compromised. To do this, a bound is enforced on the maximum overall bandwidth for high-priority traffic on each link. This is consistent with the traffic queuing policy described in Section II-B3. If this bandwidth is below the half of the link bandwidth, say, 450 Mb/s (for CIR) on a gigabit Ethernet link and the 550 Mb/s left for low-priority traffic (also for CIR), high-priority traffic swapped from a failed link will preempt the bandwidth taken by low-priority traffic on the detour path because of the queue scheduling scheme illustrated by Fig. 5. After a while, as soon as the ingress node and the egress node are notified of the failure, they will start sending traffic to the protection connection for which resource is already reserved, and the SLA for all classes of traffic will be retained.

Failure notification is done by sending OAM messages from the ingress node (the sender) to the egress node (the receiver). When the nodes local to the point of failure receive an OAM message, they will set the object of failure notification in the OAM message which will be sent to the other side of the failed point via the detour path, and the OAM message continues on the original primary path until it gets to the egress node.

D. Ethernet OAM

The original Ethernet has few OAM capabilities. This is acceptable for a LAN, but not for a metropolitan area network (MAN) that spans a large area and supports a large number of users. In a MAN, troubleshooting is more difficult and OAM becomes a necessity. There has been a lot of work on defining Ethernet OAM in the IEEE 802.3ah [Ethernet in First Mile (EFM)] Working Group and the MEF. ITU-T SG13 also initiated study on Ethernet OAM in February 2002. Some vendors have already implemented prestandard OAM functions such as Ethernet loop-back, MAC ping, and SLA measurements into their products, as well as alarms for critical problems.

OAM should be implemented for each layer of network, from the physical layer (PHY) to the transport layer, because these layers run independently. For all the layers, OAM has the same objectives:

- defection detection and positioning;
- remote failure indication/reporting;
- connectivity verification;
- performance monitoring/measurement.

Ethernet (or IEEE 802.3) has two major layers, the PHY and the MAC layer, depicted by Fig. 8. The MAC control sublayer is optional, and its main function is to process MAC control frames such as PAUSE, defined by IEEE 802.3x. In implementing Ethernet OAM, it should introduce minimum change to PHY and no change to the MAC Layer.

The EFM defines MAC OAM functions as: 1) detailed failure indication and 2) MAC layer ping and loop-back control, and PHY OAM functions as: 1) PHY ping and loop-back control; 2) alarm indications and immediate fault signaling; and 3) local/remote fault indication.

Ethernet OAM can be implemented by the local device sending Ethernet OAM frames to a remote device at a regular interval or when needed. The remote device will respond to the local device with either another OAM frame or by

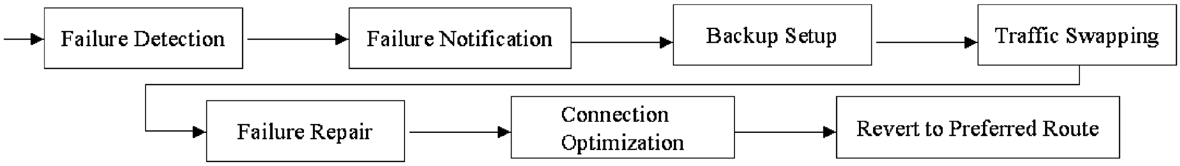

Fig. 7. Another example of protection procedures/actions.

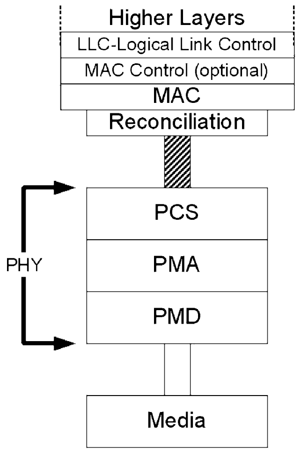

Fig. 8. Ethernet layers.

looping back the received OAM frame. This is called OAM in frames because OAM information is carried in standard Ethernet frames. The format of an Ethernet OAM frame complies with a standard Ethernet MAC control frame. The format of an Ethernet OAM frame may look like the one in Fig. 9.

As 802.3ah EFM only deals with last-mile access, as such, its OAM is just for a single link (device to device); the MEF decided to fill the gap in Ethernet OAM for end-to-end connection intra- and interproviders. Since a connection may be identified by VLAN at the network edge, the format of the Ethernet OAM frame in Fig. 9 must be changed by inserting an IEEE 802.1Q VLAN tag (VLAN EtherType + VLAN tag) right after the source MAC address so that OAM MAC layer functions such as MAC ping, EVC trace, and EVC loop-back can be applied to a specific connection that goes across multiple network nodes and links.

The above MAC layer OAM functions can be complemented by PHY OAM. PHY OAM can be used for low-level error checking and for the so-called "last gasp" immediate fatal error signaling (e.g., when a demarcation device loses its electrical power, it sends an alarm to NMS before it completely goes down). A draft proposal in IEEE 802.3ah uses the 64-b preamble sequence in Ethernet frames to carry OAM information/commands (the preamble sequence in Ethernet frame is used by the receiver to lock and synchronize its receiving clock with the clock of the transmit signal). This is called OAM in preambles. OAM in preambles requires a new sublayer called the preamble handler sublayer (PHS) to be inserted between PHY and the reconciliation sublayer.

PHY OAM is important for demarcation devices that only have PHY functions and do not have a Simple Network Management Protocol (SNMP) agent. With PHY OAM, the access/aggregation device can know the status of the demarcation devices it connects to and discover the errors all the way to the customer's premises. For more sophisticated devices such as optical Ethernet switch, PHY OAM is really just an option. While PHS is transparent to the MAC layer, it certainly needs a new PHY device. On the other hand, Ethernet has already had basic PHY error checking functions on such things as encoding error and operation status of an Ethernet interface. So network operators may not need PHY OAM everywhere. More details on PHY OAM can be found in [14].

E. Scalability

Enterprise-class Ethernet has intrinsic limitations on scalability when used as a public network. These limitations include the number of VLANs per network, the number of MAC addresses that need to be learned and stored in the device, and the long and nondeterministic convergence time of the spanning tree protocol, which is associated with the number of network elements, the number of VLANs in each element, and the complexity of the network topology. The usage of MPLS by optical Ethernet with the right system architecture enables operators to address the scalability of the network and that of the services such as E-LAN.

As described in Section II-A, VLAN is only used at the network edge, and the VLAN tag has only local significance to an interface of PE devices. This VLAN tag is the service provider's VLAN tag and is independent from the customer's VLAN. In the network core, the provider VLAN is mapped to an MPLS LSP that has a 20-b label significant to a single MPLS device. So theoretically there is no limitation on the number of point-to-point EVCs in a network.

While a conventional Ethernet switch needs to learn MAC addresses to build the forwarding table, optical Ethernet uses VLAN IDs or labels to make the forwarding decision. Therefore, there is no need to learn MAC addresses if the EVC is point to point.

Spanning Tree is also avoided because an EVC is established intentionally without loops and its protection does not rely on Spanning Tree Protocol.

MAC address learning is needed in core devices, though, in supporting multipoint E-LAN service (or Virtual Private LAN Services (VPLS), in the IETF's terms). This does not constitute any practical concern because a core device usually supports millions of MAC address, and if business customers use routers to connect to optical Ethernet, there are not many MAC addresses to be learned. The major concern comes from the full-mesh LSPs needed by a network for VPLS service, which is known as the $O(n^2)$ problem, in

MAC DA	MAC SA	Length/Type	Subtype	OAM Code	OAM Data	CRC
6	6	2	1	1	108	4

Field	Description	Value
MAC DA	Well-known Multicast Address	01-80-c1-00-00-02
MAC SA	Station's MAC Address	48-bit individual address of the station (egress port) sending the frame
Length/type	Protocol_Type	88-09
Subtype	Protocol Subtype value for EFM OAM	03 is the next available
OAM Code	01 = Ping Request 02 = Ping Response 03 = Link Monitor etc.	
OAM Data	Up to 108 octets	Data/Pad
FCS	Frame Check Sequence	32-bit CRC

Fig. 9. IEEE 802.3ah EFM Ethernet OAM frame.

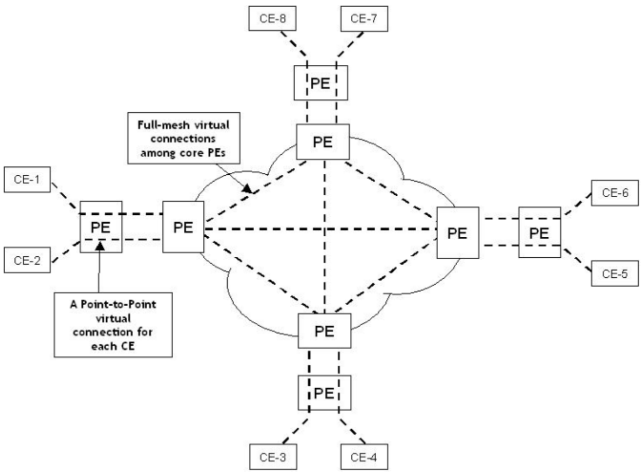

Fig. 10. HVPLS.

which n is the number of customer sites on the same VPLS. To address this issue, optical Ethernet adopts hierarchical VPLS (HVPLS). HVPLS just needs to set up full-mesh LSPs (also called LSP tunnels) among the PE devices to which the customer sites are connected. From a customer site to its connected PE device, only one point-to-point connection is needed. So as long as any one of the PE devices connects more than one customer site, the number of LSPs needed in the core is reduced. As shown in Fig. 10, the number of LSPs needed is reduced from $O(8^2)$ to $O(4^2)$.

The management and control plane of optical Ethernet is on layer 3; it has the same scalability as IP network in terms of the number of network elements supported.

The bandwidth of optical Ethernet is scalable from 100 Mb/s to 1 Gb/s and up to 10 Gb/s. Some equipment vendors also integrate dense wavelength-division multiplexing (DWDM) with ten GE modules in the same chassis so that a pair of fibers can support as much as 320 Gb/s bandwidth.

F. Fast and Easy Service Provisioning

Carriers are concerned about the network management of Ethernet network and how Ethernet services are provisioned. In fact, many carriers are still using inventory-based service provisioning, i.e., the network planners use a spreadsheet or database to maintain information on the channels, used or unused, between any two nodes, and to plan the path and allocate a channel. Operators can then log into each network element and configure the channel manually. It would be a nightmare if carriers were required to manage an Ethernet network and provision services in the same way.

Some optical Ethernet equipment vendors provide point-and-click service provisioning systems which enable operators to establish an E-Line or E-LAN service by simply choosing the end points of E-Line or E-LAN service on the NMS screen in just a few seconds. Operators simply need to point and click on the endpoints (the customer sites) of an E-Line or E-LAN service on the management station screen and set the SLA parameters. Either the provisioning system determines the best path in the network that meets the SLA requirement of an E-Line or E-LAN service or it sends the parameters of the service to the relevant PE devices and the PE devices will establish the LSP in the network core and NMS just needs to configure the CE devices. In any case, the service is established automatically.

Network operators can also explicitly choose the path on the screen and force the service to pass through this path provided that it meets the SLA.

Optical Ethernet service provisioning systems also maintain inventories for network elements, physical links, and bandwidth allocated and still available on each link, subscribers' names, locations, and services, as well as the status of each service. The system also displays the path of a service in an easy-to-understand diagram.

Optical Ethernet also provides a common object request broker architecture (CORBA) interface to operations support systems (OSS), integration with billing systems, fault management, SLA management, third-party network management, etc., which are very important for service providers to operate all their networks in a consistent and convenient way.

G. Circuit Emulation Service

Circuit emulation service (CES) is a technology that emulates a TDM circuit over Ethernet. CES can be used to interconnect private branch exchange (PBX) and existing SDH/SONET networks over Ethernet or to provide TDM-based private lines to interconnect routers with legacy interfaces. The CES implementation is based on the IETF PWE3 Working Group draft standard.

III. Conclusion

Based on the above introduction on carrier-class optical Ethernet, we can draw the following conclusions.

A. Optical Ethernet Is an Ideal Solution for Service Providers to Offer Professional Services in MANs

Carrier-class optical Ethernet is an ideal and cost-effective solution for supporting professional services in MANs. It has stringent ATM-like end-to-end QoS, SONET/SDH-like reliability with fast protection, IP-like scalability, as well as comprehensive OAM and easy-to-use service provisioning. It perfectly matches IP packets and QoS architecture.

Many service providers have started offering Ethernet services such as E-Line and E-LAN. E-Line is a natural replacement of frame relay service. E-LAN is a comparatively new type of service that had to be implemented before by multiple point-to-point virtual leased lines, which led to severe scalability issues, thus preventing it from widely being deployed. Now with MPLS-based optical Ethernet, E-LAN service becomes very scalable, and it can have the same SLA insurance as E-Line. E-LAN service is now the primary choice for LAN-to-LAN interconnections for business customers who have multiple offices or buildings in a city.

To service providers, another benefit of using optical Ethernet is that the architecture of their data network is significantly simplified. This simplification helps service providers reduce the operation complexity and cost, and shortens the time of service establishment, thus giving them the competitive edge they need.

B. Optical Ethernet Forms the Best Transport Layer for IP

Today, various access technologies are being used depending on the media (copper, fiber, or wireless) available to customers. Notably, these access technologies use different backhaul networks such as ATM or SONET/SDH. This situation is changing now. Many access technologies have begun to use Ethernet as the backhaul interface. Examples are the Ethernet-based digital subscriber line access multiplexer (DSLAM), the Ethernet-based passive optical network (EPON), and the Ethernet-based storage area network (iSCSI, for example). The IEEE EFM Working Group (802.3ah) has also developed the standards that will make Ethernet run over any media. Compared with ATM and SONET/SDH, optical Ethernet is not only cheaper, but is also more scalable in bandwidth and the number of interfaces, and it supports IP traffic more efficiently. It gets clearer that Ethernet will become the convergence layer of various services.

The implication of Ethernet transport for IP networks is significant. For the first time, the transport network has the same fundamental characteristics and attributes as the IP network, such as variable packet size and statistical multiplexing. For the first time, IP QoS can be mapped to layer 2 QoS so that end-to-end QoS can really be achieved across a WAN. For the first time, the transport network can evolve together with the IP network in the same way because they can share many technological advances.

C. Summary

In this paper, the issues of today's data network architecture, primarily its complexity and high cost, are discussed. A new simplified architecture is introduced that consists of only an optical transport layer and IP layer. The transport layer is based on optical Ethernet, which has overcome the limitations of traditional Ethernet and become carrier class in terms of end-to-end QoS, high reliability with fast protection, scalability, and OAM. This paper has made a detailed description on each of these enhancements, This new transport layer not only can transmit IP traffic more efficiently, but also it supports professional services such as E-Line and E-LAN for business customers. With optical Ethernet, service providers have begun to transition their data networks to optical Ethernet in order to offer affordable and profitable professional services with very high bandwidth.

References

[1] M. Al-Chalabi, M. Swan, and S. Yin, "Service provider metrics: Another prism for analysts," RHK, South San Francisco, CA, 2003.

[2] "Telecom economics annual forecast update," RHK, South San Francisco, CA, 2003.

[3] "Multi-client study: Market opportunities for Ethernet services," RHK, South San Francisco, CA, 2002.

[4] "Ethernet layer 2 services definitions," Metro Ethernet Forum, Dec. 2002.

[5] D. Katz *et al*. (2003) Request for comments 3630: Traffic engineering (TE) extensions to OSPF version 2. [Online]. Available: http://www.faqs.org/ftp/rfc//pdf/rfc3630.txt.pdf

[6] D. Awduche *et al*. (2001) Request for comments 3209: RSVP-TE: Extensions to RSVP for LSP tunnels. [Online]. Available: http://www.ietf.org/rfc/rfc3209.txt

[7] B. Jamoussi *et al*. (2002) Request for comments 3212: Constraint-based LSP setup using LDP. [Online]

[8] F. Le Faucheur, Ed., (2003) IETF Internet draft: Protocol extensions for support of diff-serv-aware MPLS traffic engineering. [Online]. Available: http://www.ietf.org/proceedings/03nov/I-D/draft-ietf-tewg-diff-te-proto-05.txt

[9] E. Rosen *et al*. (2001) Request for comments 3032: MPLS label stack encoding. [Online]. Available: http://www.ietf.org/rfc/rfc3032.txt

[10] M. Mathis, J. Semke, J. Mahdavi, and T. Ott, "The macroscopic behavior of the TCP congestion avoidance algorithm," *Comput. Commun. Rev.*, vol. 27, no. 3, pp. 67–82, July 1997.

[11] E. Harrison, "Protection and restoration in MPLS networks," Data Connection Ltd., 2001.

[12] P. Pan *et al.*, Ed., Fast reroute extensions to RSVP-TE for LSP tunnels. [Online]. Available: http://www.ietf.org/proceedings/03jul/I-D/draft-ietf-mpls-rsvp-lsp-fastreroute-03.txt

[13] D.-H. Gan *et al.*. A method for MPLS LSP fast-reroute using RSVP detours. [Online]. Available: http://www.potaroo.net/ietf/old-ids/draft-gan-fast-reroute-00.txt
[14] IEEE 802.3ah OAM—July 2002 Presentation Materials [Online]. Available: http://grouper.ieee.org/groups/802/3/efm/public/jul02/oam/
[15] W. Simpson, Ed., (1994) Request for comments 1619: PPP over SONET/SDH. [Online]. Available: http://rfc1619.x42.com/
[16] W. Simpson, Ed., (1994) Request for comments 1662: PPP in HDLC-link framing. [Online]. Available: http://www.ietf.org/rfc/rfc1662.txt

Jiyang Wang received the B.S. degree in computer science from Tsinghua University, Beijing, China, in 1988and the M.S. degree in computer science from the Graduate School of China Academy of Science, Beijing, in 1991.

He has been working in the networking industry for 15 years with various global companies (Chipcom, Bay Networks, Cabletron, and 3Com) in the roles of System Engineer and Product Marketing Manager. He is currently Senior Product Marketing Manager of Atrica, Inc., Paris, France, a startup company engaging in carrier-class metro and access networks.