

Supporting Ethernet in optical-burst-switched networks

Sami Sheeshia and Chunming Qiao

*Laboratory for Advanced Networking Design, Evaluation and Research,
Department of Computer Science and Engineering,
State University of New York at Buffalo, Buffalo, New York 14260
sheeshia@eng.buffalo.edu; qiao@computer.org*

Jeffrey U. J. Liu

*Computer and Communications Research Laboratories,
Industrial Technology Research Institute, HsinChu, Taiwan 310
ujliu@itri.org.tw*

Received 1 May 2002; revised manuscript received 19 June 2002

The reemergence of metropolitan area networks (MANs) is being stimulated by the continued growth of the Internet. Here we introduce the likely role that optical burst switching (OBS) will play in the development of 10-Gbit Ethernet (10GbE) metropolitan networks. Although the synchronous optical network (SONET) is being proposed to provide wide-area connectivity for 10GbE MANs, its synchronous time-division multiplexing (TDM) nature renders it inefficient for data-centric connections. OBS, however, provides a better sharing of network resources and when coupled with generalized multiprotocol label switching (GMPLS) provides a robust and more efficient transport for Ethernet services. © 2002 Optical Society of America

OCIS codes: 060.4250, 060.4510.

1. Introduction

Network service providers face a variety of challenges as they seek to capitalize on new opportunities resulting from emerging technologies and increasing user requirements. At the metro-area-network (MAN) level there is a tremendous pressure for expanded capacity to support broadband local access and high-speed wide-area networks (WANs) at a low cost. All these factors suggest that a flexible, proven MAN architecture combined with multi-vendor compatible implementations is needed so that new services such as real-time video streaming and high-speed Internet access can be introduced with fast market adoption. Standards-based 10-Gbit Ethernet (10GbE) promises to play an important role, offering a hierarchy of speeds, end-to-end protocol consistency, and technical features that are needed by both providers and users in a cost-effective way.

The concept of Ethernet-based communication over long distances is unique to 10GbE and would not be feasible with the original Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Protocol (i.e., shared media with contention). The restriction to full-duplex switched operation allows 10GbE to operate over long link spans (up to 40 km), repeaters, and other transport layers such as synchronous optical network (SONET) and dense wavelength-division multiplexing (DWDM) while maintaining the 802.3 medium access control (MAC) frame format and length.

Historically, wide-area connectivity has been provided by SONET rings, which were designed and built to carry voice traffic. However, the SONET transport infrastructure is not optimized for bursty data traffic and cannot scale to support the rapid growth of Internet

traffic in a cost-effective manner, because additional capacity requires expensive add-drop multiplexers (ADMs), hubs, and duplicate fiber rings.

Numerous enhancements such as carrier-class Ethernet (CCE) have been proposed to optimize existing fiber ring infrastructures and offer Ethernet-based services. More specifically, Gigabit Ethernet has been popular in the local area network (LAN) and the MAN because of its simplicity and economies of scale. Multiple branches of corporations can be connected via point-to-point or point-to-multipoint Ethernet-based virtual private networks (VPNs) at a much lower cost than with SONET. However, Ethernet by itself is not considered to be a carrier-class protocol because it does not offer SONET-like resiliency, carrier-to-carrier services, and service level agreement (SLA) guarantees. Several standards such as Ethernet over WDM, resilient packet ring (RPR), and packet over SONET (PoS), have been developed to enhance Ethernet with carrier-grade attributes.

Ethernet over WDM such as 10GbE provides long-haul (up to 40 km) connectivity between two Ethernet switches. Wavelength-based point-to-point connections are provisioned manually or via multiprotocol lambda switching (MP λ S),^{1,2} at the full wavelength rate regardless of the traffic load. This limits the total number of circuits that can be provisioned and implies that no traffic aggregation or grooming can be done inside the core. RPR leverages existing SONET ring infrastructures to offer CCE services in the MAN. RPR defines the way multiple nodes connecting to a shared optical ring can achieve controlled bandwidth allocation, bounded delays and jitters, and 50-ms service restoration in the event of a ring failure. The switching capabilities of RPR's MAC layer provide for more efficient use of core fiber and ring operation. However, RPR has not gained any momentum, because carriers have opted to improve the SONET transport instead.

Packet over SONET provides WAN connectivity without the layering of poorly integrated protocols such as Internet Protocol (IP) over asynchronous transfer mode (ATM) over SONET. It aggregates and encapsulates IP datagrams in Point-to-Point Protocol (PPP) frames without any differentiation among the different packet flows. As a result PoS lacks multicast and quality-of-service (QoS) capabilities.

Next-generation SONET provides Ethernet traffic with a better bandwidth granularity than SONET and POS by multiplexing time-division multiplexing (TDM) and data traffic onto the same timeslot. New techniques for concatenating Ethernet frames at rates lower than SONET's basic STS-1 rate have been proposed.³ These techniques are known as virtual concatenation (VC) and result in a better match between the Ethernet data rates and the SONET line rates, thereby improving circuit use. Packet and signal encapsulation such as the generic framing procedure (GFP)⁴ have also been proposed to adapt multiservices to the SONET infrastructure. GFP provides a uniform mapping of Ethernet frames and client signals such as fiber channel into fixed-length GFP frames. Single-bit and multibit header error correction capabilities and packet-level multiplexing further improve the use of the TDM channels. Dynamic provisioning specifications such as the link capacity access scheme (LCAS)⁵ provide SONET with bandwidth-on-demand capabilities in STS-1 increments. Although next-generation SONET loosens the rigid multiplexing schemes of traditional SONET, it provides only incremental gains, since it does not address the issues associated with using circuit-switching technology for data traffic.

In short, the standards mentioned above either lack the end-to-end optical transparency or do not offer efficient use of core resources as traditional packet networks.⁶ In this paper we propose to take advantage of OBS technology to transmit Ethernet frames over WDM links directly to provide a scalable and data-optimized alternative to SONET-based connectivity.

The remainder of the paper is structured as follows. The implementation of 10GbE over SONET is discussed highlighting several disadvantages and quantifying its inefficiencies. Ethernet over OBS (EoB) is presented as a viable alternative that avoids the shortcomings

of SONET and MPLS. Ethernet-OBS specific integration issues such as burst size and aggregation are then discussed within the generalized multiprotocol label switching (GMPLS) framework.

2. Ethernet over SONET

SONET was designed primarily for voice applications by use of point-to-point links operating at a well-defined hierarchy of speeds (i.e., OC-1 at 51.840 Mbit/s up to OC-192 at 9953.281 Mbit/s). Its existing TDM-based transport infrastructure is based on circuit-oriented technology that is voice optimized. Dual ring and equipment (1 + 1 redundancy) topology enables SONET to implement a fast (sub-50-ms) protection mechanism that can restore connectivity by use of an alternate path in case of fiber cuts or equipment failure.

The 10GbE WAN physical layer is defined to be compatible with SONET. It is SONET friendly even though it is not fully compliant with all the SONET standards. Some of the implemented SONET features are the OC-192 link speed, the use of SONET framing, and some overhead processing. The most costly aspects of SONET, such as TDM support and performance and management functions, have been avoided.

There are well-known disadvantages and limitations to using SONET solutions, such as PoS, for transporting data traffic. SONET was designed for point-to-point, circuit-switched applications, and most of its limitations stem from these origins. The subsections below describe some of the disadvantages of using SONET rings for data transport.

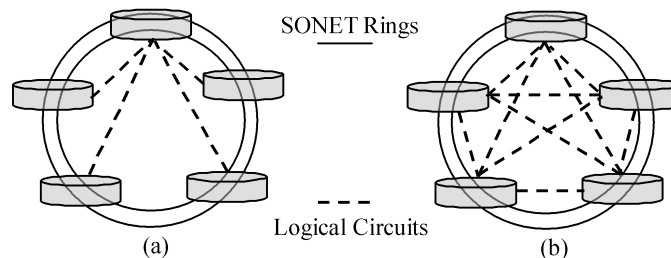


Fig. 1. SONET (a) access and (b) fully meshed networks.

2.A. Fixed Circuits

SONET provisions point-to-point circuits between ring nodes. Each circuit is allocated a fixed amount of bandwidth that is wasted when not used. For the SONET network that is used for access in Fig. 1(a) each node on the ring is allocated only one quarter of the ring's total bandwidth. That fixed allocation puts a limit on the maximum data burst traffic rates and increases the queuing delays between endpoints. Creating a logical mesh over a SONET ring as shown in Fig. 1(b) requires $N \times (N - 1)$ circuits, which are not only difficult and time consuming to provision but waste ring bandwidth. As traffic levels increase, the ease of deployment, maintenance, and upgradability becomes a critical operational requirement.

2.B. Multicast Traffic

On a SONET ring, layer-2 multicast requires each source to allocate a separate circuit for each destination. This requires a full mesh in which separate copies of the packet are sent

to each destination. The result is multiple copies of multicast packets traveling around the ring, thus wasting bandwidth.

2.C. Wasted Protection Bandwidth

Typically, 50% of ring bandwidth and equipment is reserved for protection. Although protection is obviously important, SONET does not achieve this goal in an efficient and cost-effective manner that gives the provider the choice of when and how much bandwidth to reserve for protection.

2.D. Added Overhead

When 10GbE frames are transported over circuits built from SONET links, then full SONET TDM capabilities, the SONET physical layer, and the various SONET management functions would all be included; the resulting SONET overhead for STS-192c frames, for example, is 3.7% (640/17280). However, at low traffic loads, the overall efficiency worsens as bursts of Ethernet frames are delivered in a fraction of the SONET time slot, resulting largely from a 10GbE high input-data rate.

Inefficiencies associated with carrying increasing quantities of data traffic over manually provisioned circuit-switched networks make it difficult to develop new services, and they increase the cost of building additional capacity.

3. Ethernet-over-SONET Efficiency

In the 10GBASE-W specifications, the WAN interface sublayer (WIS) provides SONET compatibility by throttling the MAC transmission to OC-192 speed and proper framing. However, because of the irregularities of Ethernet traffic such as burstiness and variable frame size, SONET framing introduces several inefficiencies. If the Ethernet frame size is modeled with an exponential distribution with parameter λ frame/Byte, then the efficiency of EoS is given in Eq. (1):

$$\eta_{EoS} = \eta_{SONET} \eta_{PE}, \quad (1)$$

where η_{PE} is the efficiency of packing Ethernet frames into the SONET payload (SPE). As seen in Fig. 2, the SONET payload carries a train of Ethernet frames, interarrival gaps (g), and an unused remainder (r).

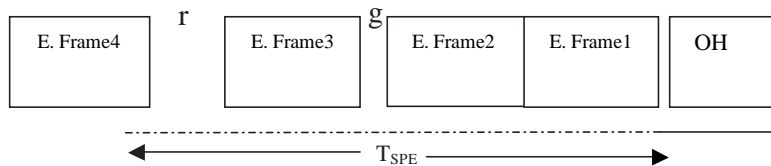


Fig. 2. Ethernet frames in SONET payload.

The unused remainder in each slot is due to low traffic load, or the size of an incoming Ethernet frame is larger than the remainder in the time slot. Hence the *packing efficiency* of Ethernet frames into a SONET T_{SPE} is as follows (see Appendix A for details):

$$\eta_{PE} = \sum_{k=1}^M \frac{\lambda^k}{(k+1)} \frac{\exp(-\lambda B) B^{k+1}}{(k-1)! T_{SPE}}, \quad (2)$$

where M is the maximum number of Ethernet frames that the SPE can hold and B is the size of the incoming Ethernet burst. Because of the bursty nature of Ethernet traffic, frame arrivals may become sporadic, separated with time intervals (g) possibly larger than T_{SPE} . As g varies, η_{PE} becomes proportional to the offered load.

4. Ethernet over Optical Burst Switching

The inefficiencies created by TDM suggest that a new approach is needed to carry Ethernet directly over optical fibers, and various schemes and technologies are being devised. However, these technologies still do not provide all-optical transparent paths between source and destination as the paths are interrupted at intermediate nodes (every 40 km or so) by optical–electronic–optical (O–E–O) conversions.

The recent arrival of all-optical cross-connects (OXC) and ultra-long-haul transmission technologies has made it possible to extend 10GbE paths to great distances (e.g., 4000 km) over multiple hops. This may be accomplished by provisioning of lightpaths (or wavelengths) between a source and a destination to provide 10-Gbit/s logical circuits. These constitute point-to-point and permanent reservation of resources, which can no longer be shared among different 10GbE paths. As a result, the efficiency of each lightpath is limited to the offered load, and the lack of statistical multiplexing leads to poor flexibility and scalability of such MPLS approaches.

OBS^{7,8} is a new technology that exploits the large bandwidths of DWDM transmission systems by avoiding the electronic processing of optical packets. The basic data block of OBS is a burst, which is a collection of data frames with the same destination address and QoS parameters. The optical burst is switched through a predetermined path in the optical core without any O–E–O conversions. The optical path is set up *a priori* by use of a control packet that is processed electronically at each core switch to determine the outgoing link and channel (wavelength). Once the burst arrives and flows through the switch, the channel is released, immediately following the burst transmission (or through an explicit control packet), and becomes available for other burst transmissions. This statistical multiplexing of the channel in the time domain among multiple flows improves the backbone efficiency and offers excellent scalability. Figure 3 shows a metropolitan edge switch with OBS and 10GbE interfaces.

Edge switches aggregate 10GbE frames into bursts and send out control packets to reserve bandwidth along predetermined end-to-end labeled optical-burst-switched (LOBS) paths.⁹ This implies that GMPLS labels are used to identify the destination edge switch and are mapped to an available wavelength at each intermediate OBS switch. However, unlike wavelength routing under MPLS, this scheme has many advantages, as described below.

4.A. Switched Paths

The LOBS paths do not constitute permanent reservations of resources, as is the case with SONET connections or wavelength paths under MPLS. They are set up dynamically to support the required QoS and torn down once the bursts are transmitted. In addition, the bandwidth is dynamically allocated and released on a burst-by-burst basis.

4.B. Peer-to-Peer Networking

GMPLS extends the WAN connectivity into the MAN thereby simplifying the interface among the multitier providers. Complex optical user-to-network (UNI) interfaces to demark the WAN–MAN boundaries between providers and customers are no longer required.

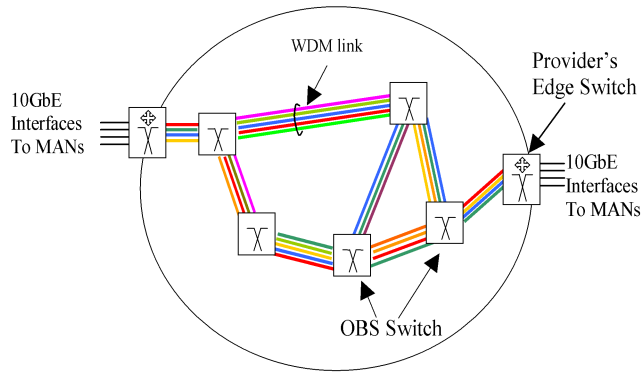


Fig. 3. OBS network with 10GbE interfaces.

4.C. Multicast Traffic

OBS switches support multicast at the WDM layer by using light-splitting techniques.¹⁰ These schemes require additional hardware that can further complicate the functionality of the switch control module. However, recent enhancements in the GMPLS-based multicast standards would provide simplification and scalability at no cost.

4.D. Protection and Restoration

The GMPLS fast reroute mechanism can be used to provide protection and restoration around a failed node or link by automatically rerouting traffic on an alternate LOBS. This can be accomplished by means of precomputing and preestablishing a number of protection LOBS paths, between the source and destination edge switches or around intermediate switches and links, at the same time that the primary LOBS paths are being established. These protection LOBS act as temporary tunnels activated only in case of path failures.

GMPLS-enabled OBS provides a pseudo-packet-based network solution that is more compatible with Ethernet's bursty nature and avoids the provisioning complications and inefficiencies of SONET and MPλS. However, integrating 10GbE services over GMPLS-enabled OBS requires attention to specific details such as the burst size and GMPLS extensions.

5. Ethernet-over-OBS Efficiency

In general, the efficiency of OBS is related to the fraction of time the reserved bandwidth along a LOBS path is used by the burst. It is then imperative to activate the bandwidth reservation as close as possible to the burst transmission time. As shown in Fig. 4, the OBS efficiency is dependent on t_s , the time at which the bandwidth reservation starts.¹¹

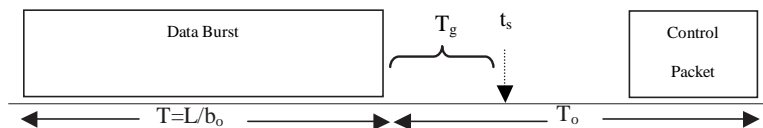


Fig. 4. Transmission efficiency and bandwidth reservation.

Neglecting any propagation delays, we give the efficiency in Eq. (3) where T is the burst time (the burst size L divided by the output data rate b_o) and T_o is the *offset time*. The *maximum efficiency* is achieved when $T \gg T_o$ (very large bursts) or $t_s \approx T_o$,

$$\eta_{\text{EoB}} = \frac{T}{T + (T_o - t_s)}. \quad (3)$$

End-to-end queuing delay, on the other hand, is dependent on whether fixed- or variable-size bursts are used. This is because fixed-size bursts allow the path to be set up as the burst is being assembled, whereas variable-size bursts do not. The *minimum offset time* between the control packet and the data burst is a function of the end-to-end queuing and processing delays, Δ , of the control packet,

$$T_o \gg (N - 1)\Delta, \quad (4)$$

where N is the number of OBS switches that the LOBS traverses. When the length of the burst is fixed, the control packet can be transmitted as soon as the first Ethernet frame arrives. The data burst encounters no edge delay as long as the *offset time* equals the time it takes to assemble the burst, as in Eq. (5), and the condition in Eq. (4) is met,

$$T_o = \frac{L}{b_{\text{in}}}, \quad (5)$$

where b_{in} is the input data rate. However, when the length of the burst varies, the control packet can be transmitted only after the entire burst has been assembled, that is, after a maximum interval T_i . This implies that the minimum edge delay becomes T_o .

6. Ethernet-over-OBS Burst Size

To improve efficiency, the OBS burst must be larger than $T_o - t_s$ but not so large as to cause significant queuing delays. However, high bandwidth and full-duplex operation requires a flow-control mechanism: the 10GbE source and destination exchange flow-control packets to throttle frame transmission and prevent data loss. These packets are typically 64 Bytes and must be transmitted with minimal queuing delays. As a result, flow-control packets should not be assembled into small bursts, since this affects efficiency, but transmitted with the control channels instead. The elasticity of the OBS burst, as opposed to the SONET fixed slot size, allows it to adapt to the bursty nature of Ethernet traffic and its varying frame size while accounting for any transmission overhead, which includes the interframe gaps (IFGs), preambles, and delimiters.

At 10 Gbit/s, the Ethernet adaptor must be able to buffer and process frames quickly; there are counters to be updated and timers to be synchronized with every frame received or transmitted. Ethernet requires a minimum IFG, or pause, between two successive frames; at 10GbE the IFG is ~ 9.6 ns, or the equivalent of 12 Bytes. Although the usefulness of the IFG has been debated in recent literature, it provides the receiving station with time to update its counters, check the CRC (cyclic redundancy check) of the previous frame, and better manage its buffers. There are two ways to implement the IFG in OBS: Insert the 12-Byte spacing at the 10GbE receiving interface during the O-E-O processing of the burst, or incorporate the IFG within the burst. The first approach places new requirements on the processing at the destination switch to seek the start of each frame within the burst instead of simply disassembling the burst. The second approach maintains the integrity of the Ethernet transmission at the expense of wasted bandwidth within each burst.

Ethernet's preamble (7 Bytes) and start-frame delimiter (1 Byte) must also be maintained within the optical burst. These fields serve to maintain clock synchronization between the sender and the receiver. Since the IFG, preamble, and frame check sequence

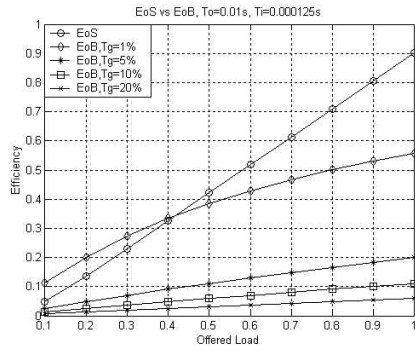


Fig. 5. $T_i = 1$ SONET slot.

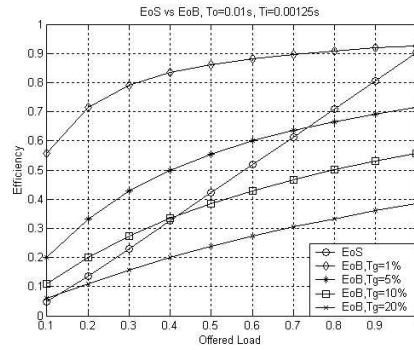


Fig. 6. $T_i = 10$ SONET slots.

(FCS) constitute a large overhead for the 1500-Byte Ethernet frame, jumbo, 9000-Byte, frames should be used to improve the burst's efficiency. For example, a 155,520-Byte (SONET slot equivalent) burst carrying standard 1500-Byte frames incurs approximately 1% overhead, whereas 9000-Byte frames require fewer IFGs or preambles, resulting in 0.1% overhead, a reduction of tenfold.

Figures 5 and 6 compare the performance of EoS versus EoB under different traffic loads and T_i (burst aggregation time) values of one and ten SONET slots, respectively. As shown, the efficiency of EoS is highly linear to the offered load regardless of the number of time slots allocated for a particular connection. This is because the SONET slots are fixed and do not adapt to varying traffic loads. When T_i is small and the offered traffic load is higher than 40%, EoS outperforms EoB by a significant margin.

On the other hand, the per-wavelength efficiency of EoB is a function of the burst size L and, to a lesser extent, on the guard time T_g , a minimum of $T_o - t_s$. As L increases, the burst time T becomes more dominant than $T_o - t_s$, meaning that the LOBS is being used for data transmission during most of its active time. Consequently, the efficiency of EoB for large bursts (50+ time slots in size) approaches unity under all traffic loads. However, for smaller bursts (less than 50 time slots), T_g plays a critical role as shown in Fig. 5; the smaller the T_g , the better the efficiency.

7. Ethernet-over-OBS Multiprotocol Label Switching Issues

10GbE is a data-link Ethernet technology optimized for data traffic; it remains connection-less in nature and lacks traffic-engineering capabilities. As a result, providing end-to-end differentiated services with Ethernet is not ideal.

Service creation over Ethernet such as transparent LAN services (TLS) across the WAN¹² requires extending GMPLS past the edge switch into the MAN. Virtual leased lines (VLLs), or Ethernet tunnels, can be provisioned dynamically among customer sites to create wide-area VLANs and VPNs as shown in Fig. 7. GMPLS-based signaling allows the creation of these tunnels across the OBS network and according to the customer-specified bandwidth and delay requirements.

This scheme provides peer-to-peer networking whereby the customer premise equipment (CPE) is part of the MPLS cloud. The CPEs must also act as Address Resolution Protocol (ARP) servers, resolving destination IPs to destination MACs, and ultimately, to destination labels for connected LANs. The CPE uses the Spanning Tree Protocol (STP) to discover all the VLAN-Port-MAC addresses of local stations. In Fig. 7 each VLAN-MAC-FEC combination (FEC is forwarding equivalence class) is assigned a unique label

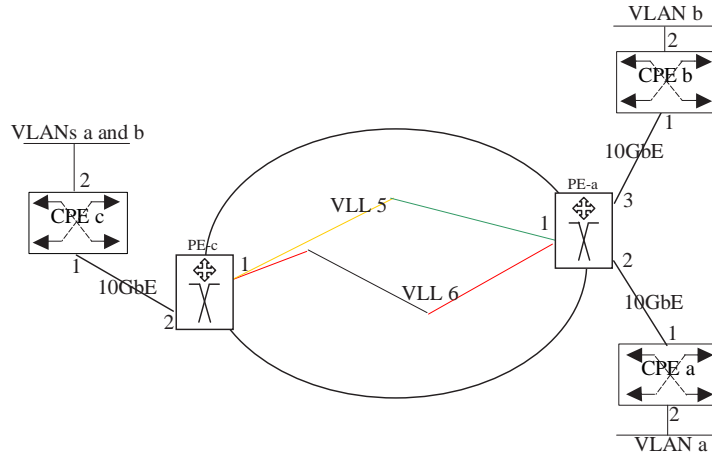


Fig. 7. Ethernet VLANs mapped with VLLs.

as shown in Table 1. The CPEs exchange their tables with the provider's edge (PE) switches using GMPLS signaling to create unified forwarding information tables.

Table 1. CPE: a Forwarding Information Base

VLAN	Port	MAC Address	Label	FEC
a	2	xx-xx-xx-xx-xx-aa	2	1
a	2	xx-xx-xx-xx-xx-aa	3	2

Ethernet frames arriving at the CPE from local LANs carry the original Ethernet fields and 802.1p/q headers as shown in Fig. 8.

D	S	E-type	802.1p/q	E-type	Ethernet
A	A	(0x8100)		(0x0800)	Payload

Fig. 8. VLAN-tagged frame arriving at CPE.

Once the frame is deemed valid, the frame is mapped to a user-defined FEC with the 802.1p/q and destination address fields. The FEC lookup yields the outgoing port and a virtual circuit (VC) label, which is then added to the frame and forwarded on the outgoing port toward the PE switch.

The PE switch maps the Ethernet frame into an OBS burst and provides the required signaling by extending the GMPLS label assignment. This can be achieved in one of two addressing schemes, flat and hierarchical.

In the flat assignment scheme, the PE constructs its label forwarding information base as shown in Table 2. Similar to the CPE label assignment, each label designates a unique VLAN-MAC-FEC and only one is assigned to the frame.

The advantage of this scheme is that it is simple to implement but provides no means of VC aggregation through the OBS cloud; it merely provides point-to-point connection

Table 2. PE: a Flat Label Forwarding Information Base

Ingress Port	Ingress Label	Egress Port	Egress Label
2	2	1	3
2	3	1	4
3	7	1	9

between source and destination. In addition, each source–destination pair requires multiple LOBS, one per FEC, thereby consuming many labels and increasing the amount of signaling.

In the hierarchical scheme shown in Fig. 7, each VLAN is assigned one LOBS or VLL label as shown in Table 3.

Table 3. PE: Label Forwarding Information Base

Ingress Port	Ingress Label	Egress Port	VLL Label
2	2	1	5
2	3	1	5
3	2	1	6

The PE looks up the incoming label, determines its VLAN membership, and adds a LOBS label as shown in Fig. 9. Frames with the same VLAN labels get the same LOBS label and are aggregated into the same burst. The MPLS Ethernet type (0x8847 for unicast and 0x8848 for multicast) field is added to facilitate the processing of the frame at the destination PE, whereas the destination address (DA) field is that of the destination PE 10GbE interface.

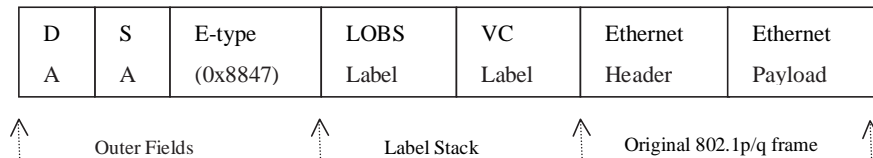


Fig. 9. Frame format within the burst.

The VLL label is used to aggregate flows belonging to the same VLAN. This allows for easy rerouting of VLAN traffic in the event of failure. Furthermore, once a packet is classified as multicast, the PE duplicates and forwards the frame to all VCs that belong to the VLAN represented by the VLL or LOBS label. The disadvantage of hierarchical labeling is the additional processing overhead.

Frame aggregation into bursts is based on destination and QoS parameters. This can be accomplished in two different ways depending on the PE addressing scheme used. Once the CPE determines the priority of an Ethernet frame using the 802.1p/q priority field, the frame can either be marked with the appropriate class of service (CoS) or it can be mapped to a specific QoS VC by use of the VC label.¹³ The VC label consists of a 20-bit label field, 3-bit CoS field, 1-bit bottom of stack, and an 8-bit time-to-live (TTL) field. When a single LOBS (hierarchical scheme) carries multiple VCs (classes of service), the CoS bits are

used to determine aggregation. However, when multiple LOBS (flat scheme) of different classes of service are used, the CoS field is used to identify the VC label. QoS and class differentiation through the OBS cloud are realized by adjustment of the *offset time* in the control packet.^{14,15} The PE maps the CoS or VC labels to the appropriate *offset times* when constructing the control packet.

The ability to treat pairs of Ethernet VCs as virtual interfaces that can be added to a VLAN allows transparent bridging to operate. A broadcast frame or a frame with an unknown destination address is flooded on all VCs that are part of the VLAN. Once the MAC addresses have been learned, frames are sent only on the proper VC.

EoB path protection and restoration can be provided at multiple layers: WDM, GMPLS, or the Ethernet layer. At the optical layer the protection and restoration schemes match those of SONET but are inflexible and remain subjects of current research. At the Ethernet layer, the STP is used to provide redundant paths. However, STP's long convergence time prohibits it from providing such functionality, especially because LOBS paths are constantly set up and torn down. The GMPLS layer is best suited to provide flexible and SONET-like protection and restoration procedures. There are two such procedures, head end and local reroute.

Both reroute algorithms precompute and preestablish a number of protection LOBS between source and destination. When a failure takes place, the OBS switch signals the head-end switch to use a backup LOBS¹⁶ or makes a local decision to redirect the burst over a detour path such as the case with local reroute. The usefulness of these reroute mechanisms is determined by the speed with which a detour path is established, the time it takes to reroute traffic, and the ease of configuring detour paths within the network.

To achieve a timely detour path setup, a protection path cannot be established when a failure is detected. Thus using a precomputed and preestablished detour path is essential for burst traffic. Shortest reroute time implies that the detour decision must be made as close to the failure point as possible, since it may take a significant amount of time to notify the head node in the LOBS path. This is why the local reroute mechanism is preferred to the head-end algorithm. However, since it is impossible to predict where failure may occur along the LOBS path, every switch and link along the path must be protected. Local reroute requires establishing $(N - 1)$ detour paths as shown in Fig. 10, where N is the number of OBS switches that the LOBS traverses. These detour paths can be set up dynamically at the same time the primary path is being set up; the ingress switch replicates the control packet and forwards it on the backup route. This implies that every GMPLS port-label association must have a backup port-label pair that is node and link disjoint.

It is worth noting that whereas GMPLS provides generalized packet label-switched paths (LSPs) in the electronic and circuit LSPs in the optical domains, several differences between these LSPs and LOBS remain.

First, packet LSPs generally exist in the electronic domain where electronic buffering and processing at intermediate nodes are feasible. In addition, edge PE devices run in-band signaling and IGP protocols across permanent Ethernet tunnels [LSP or general routing encapsulation (GRE)], to exchange VLAN-MAC-label information. Once learned, VC labels are added to Ethernet frames and transmitted onto LSPs within the proper VLAN tunnels. QoS is supported by giving precedence to high-priority traffic and reserving resources (queues) for a particular LSP within the tunnel.

LOBS paths are all-optical transparent connections in which intermediate nodes do not buffer or process data bursts; also, they do not consume core resources (wavelengths) as their optical domain counterparts (called wavelength paths or lightpaths) do. With LOBS, the edge PE devices perform any data aggregation and prioritization functions and use

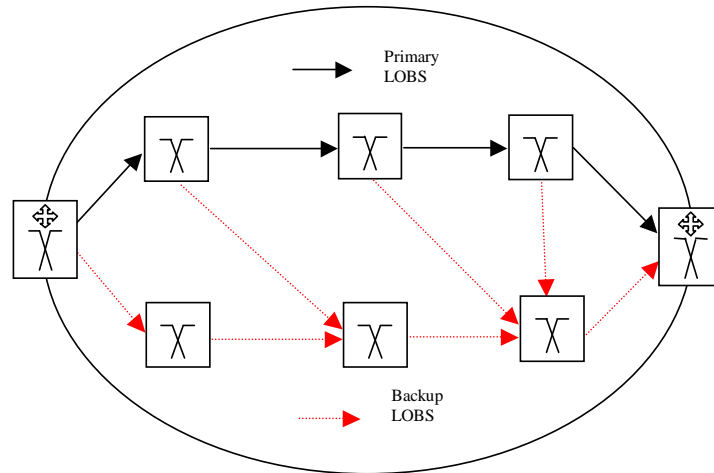


Fig. 10. Primary and backup LOBS.

out-of-band signaling on dedicated control channels to exchange VLAN-MAC-label information and activate or deactivate LOBS paths. In addition, QoS is supported by addition of an extra offset time to the bursts of high priorities. Because of these and other differences, necessary extensions to the existing GMPLS framework are needed and will likely be the subject of future study.

8. Conclusion

Extending Ethernet services over OBS provides better scalability and bandwidth efficiency than with SONET, which is inherently limited by its TDM nature and by MPLS, which provides no statistical multiplexing.

As DWDM technology evolves, more and more wavelengths will be supported on each fiber. This increase in bandwidth does not require new fiber, ADMs, or hubs; it does however require an efficient transport protocol such as OBS and a dynamic provisioning and management plane such as GMPLS. The simplicity and reliability of Ethernet make it an ideal technology for the LAN-MAN connectivity, whereas OBS provides the efficient sharing of backbone resources. GMPLS provides a standardized control mechanism to integrate, manage, and extend Ethernet over OBS networks and bridge distant MANs and customer sites. It also allows the provisioning of new services such as wide-area VLANs and circuit-like connections with protection and restoration, but without the permanent reservation of expensive backbone resources.

These features are attractive for service providers because they leverage their installed infrastructure among many customers. In addition, the GMPLS peer-to-peer integration allows providers and customers to develop new long-term services such as outsourcing; providers improve their revenue stream while customers reduce their network operational costs.

Appendix A

Let B be the burst size and Y be the random variable representing the train of Ethernet frames L_1, L_2, \dots, L_K where L_n represent the lengths of these frames. If L_n are independent and identically distributed random variables with an exponential density function $f_L(l)$, then

$$Y = \begin{cases} 0 & \text{if no traffic} \\ L_1 & \text{if } L_1 \leq B < L_1 + L_2 \\ L_1 + L_2 & \text{if } L_1 + L_2 \leq B < L_1 + L_2 + L_3 \quad . \\ \sum_{k=1}^K L_k & \text{if } \sum_{k=1}^K L_k \leq B < \sum_{k=1}^{K+1} L_k \end{cases}$$

Define $Z(K) = \sum_{k=1}^K L_k$; then $f_{Z(K)}(l)$ is the convolution of $f_L(l)$ K times,

$$f_{Z(K)}(l) = f_{L1}(l) * f_{L2}(l) * \cdots * f_{LK}(l).$$

The convolution of K exponential distributions leads to a K -type Erlang distribution,

$$f_{Z(K)}(l) = \frac{\lambda \exp(-\lambda l) (\lambda l)^{k-1}}{(k-1)!},$$

where $1/\lambda$ is the average Ethernet frame size. Since $y = z$ if $z \leq B$ and $B < z + l_{K+1} \Rightarrow$

$$\begin{aligned} E[Y] &= \sum_{k=1}^{\infty} y_k P_r(y_k) = \sum_{k=1}^{\infty} y_k \int_0^{\infty} f_Y(y_k) dy \\ &= \sum_{k=1}^{\infty} z_k \int_0^B f_{Z(K)}(z) dz P_r(l > B - z) \\ &= \sum_{k=1}^{\infty} \int_0^B z_k f_{Z(K)}(z) dz [1 - F_L(B - z)] \\ &= \sum_{k=1}^{\infty} \frac{\lambda^k \exp(-\lambda B)}{(k-1)!} \frac{B^{k+1}}{(k+1)}. \end{aligned}$$

Hence the efficiency of packing Ethernet frames in slot T_s is

$$\eta_{PE} = \frac{E[Y]}{T_s}.$$

References and Links

1. N. Ghani, S. Dixit, and T. Wang, "On IP-over-WDM integration," IEEE Commun. Mag. (March 2000), pp. 72–84.
2. P. Green, "Progress in Optical Networking," IEEE Commun. Mag. (January 2001), pp. 54–61.
3. ITU-T Recommendation G.707, "Network node interface for the synchronous digital hierarchy" (International Telecommunication Union, March 2001), <http://www.itu.int/ITU-T/>.
4. ITU-T Recommendation G.7041/Y.1303, "Generic framing procedure" (International Telecommunication Union, December 2001), <http://www.itu.int/ITU-T/>.
5. ITU-T Recommendation G.7042/Y.1305, "Link capacity adjustment scheme for virtual concatenated signals" (International Telecommunication Union, November 2001), <http://www.itu.int/ITU-T/>.
6. A. Jourdan, D. Chiaroni, and E. Dotaro, "The perspective of optical packet switching in IP-dominant backbone and metropolitan networks," IEEE Commun. Mag. (March 2001), pp. 136–141.
7. C. Qiao and M. Yoo, "Optical burst switching (OBS)—a new paradigm for an optical Internet," J. High Speed Netw. **8**, 69–84 (1999).
8. C. Qiao and M. Yoo, "Choices, features and issues in optical burst switching (OBS)," Opt. Netw. Mag. (April 2000), pp. 36–44.

9. C. Qiao, "Labeled optical burst switching for IP-over-WDM integration," *IEEE Commun. Mag.* (September 2000), pp. 104–114,
10. C. Qiao, X. Zhang, and J. Wei, "Constrained multicast routing in WDM networks with sparse light splitting," *J. Lightwave Technol.* **18**, 1917–1927, (2000).
11. M. Yoo and C. Qiao, "Just-enough-time (JET): a high speed protocol for bursty traffic in optical networks," in *Digest of IEEE/LEOS Summer Topical Meetings on Technologies for a Global Information Infrastructure* (Institute of Electrical and Electronics Engineers, New York, 1997), pp. 26–27.
12. P. Menezes, L. Anderson, A. Smith, P. Lin, J. Heinanen, G. Heron, R. Haberman, T. Soon, N. Slabakov, and L. Martini, "Virtual private LAN services over MPLS," Internet Draft, draft-lasserre-vkompella-ppvnp-vpls-01.txt (Internet Engineering Task Force, March 2002), pp. 1–21, <http://www.ietf.org/internet-drafts/draft-lasserre-vkompella-ppvnp-vpls-01.txt>.
13. G. Heron, S. Vogelsang, C. Liljenstolpe, V. Radoaca, D. Tappan, K. Kompella, and A. Malis, "Encapsulation methods for transport of ethernet frames over IP and MPLS networks," Internet Draft, draft-martini-ethernet-encap-mpls-00.txt (Internet Engineering Task Force, April 2002), pp. 1–9, <http://www.ietf.org/internet-drafts/draft-martini-ethernet-encap-mpls-00.txt>.
14. M. Yoo, C. Qiao, and S. Dixit, "QoS performance of optical burst switching in IP-over-WDM networks," *IEEE J. Sel. Areas. Commun. Special Issue on Protocols for Next Generation Optical Networks* **18**, 2062–2071 (2000).
15. M. Yoo, C. Qiao, and S. Dixit, "Optical burst switching for service differentiation in the next-generation optical internet," *IEEE Commun. Mag.* (February 2001), pp. 98–104.
16. D. Haskin and R. Krishnan, "A method for setting an alternative label switched path to handle fast reroute," work in progress, Internet Draft, draft-haskin-mpls-fast-reroute-05.txt (Internet Engineering Task Force, November 2002), pp. 1–9, <http://www.netzmafia.de/rfc/internet-drafts/draft-haskin-mpls-fast-reroute-05.txt>.