Optical Coarse Packet-Switched IP-over-WDM Network (OPSINET): Technologies and Experiments

Maria C. Yuang, *Senior Member, IEEE,* Steven S. W. Lee, *Member, IEEE,* Po-Lung Tien, *Member, IEEE,* Yu-Min Lin, Julin Shih, *Student Member, IEEE,* Frank Tsai, and Alice Chen

***Abstract*— Optical Packet Switching (OPS) has been envisioned as a prominent future optical networking technology for data-centric IP over Wavelength Division Multiplexing (WDM) networks, or optical Internet. Such OPS technology however raises significant transport and Quality of Service (QoS) challenges due to technological limitations. To circumvent OPS limitations, we have proposed a new Optical Coarse Packet Switching (OCPS) paradigm, which uses in-band-controlled per-burst switching and advocates traffic control enforcement to achieve high packet-loss performance and differentiated services. Based on OCPS, we have constructed an experimental IP-over-WDM network, referred to as OPSINET. OPSINET consists of two major types of nodes- edge routers, and Optical Label Switched Routers (OLSRs), and is facilitated with an out-of-band Generalized Multi-protocol Label Switching (GMPLS) control network. In this paper, we first introduce the OCPS paradigm. We then present the architecture of OPSINET, describe the in-band header/payload modulation technique, and detail the operations of the edge routers, OLSRs, and GMPLS control.**

***Index Terms*— IP-over-Wavelength Division Multiplexing (WDM), Optical Packet Switching (OPS), Quality of Service (QoS), Optical Label Switched Routers (OLSRs), Generalized Multi-protocol Label Switching (GMPLS).**

I. Introduction

THE ever-growing demand for Internet bandwidth and recent advances in optical Wavelength Division Multiplexing (WDM) technologies [1] brings about fundamental changes in the design and implementation of the next generation IP-over-WDM networks or optical Internet. Current applications of WDM mostly follow the Optical Circuit Switching (OCS) paradigm by making relatively static utilization of individual WDM channels. Optical Packet Switching (OPS) technologies [2-4], on the other hand, enable fine-grained on-demand channel allocation and have been envisioned as an ultimate solution for data-centric optical Internet. Nevertheless, OPS currently faces some technological limitations, such as the lack of optical signal processing and optical buffer technologies, and large switching overhead. In light of this, while some work [3,4] directly confronts the OPS limitations, others attempt to tackle the problem by exploiting different switching paradigms, in which Optical Burst Switching (OBS) [5-12] has received most attention.

OBS [5] was originally designed to efficiently support all-optical bufferless [6,7] networks while circumventing OPS limitations. By adopting per-burst switching, OBS requires IP packets to be first assembled into bursts at ingress nodes. The most common packet assembly schemes are based on timer [13], packet-count threshold [7], and a combination of both [7,14]. Essentially, major focuses in OBS have been on one-way out-of-band wavelength allocation (e.g., Just-In-Time (JIT) [8], and Just-Enough-Time (JET) [6]), and the support of QoS for networks without buffers [6,7] or with limited Fiber-Delay-Line (FDL)-based buffers [9]. Particularly in the JET-based OBS scheme that is considered most effective, a control packet for each burst payload is first transmitted out-of-band, allowing each switch to perform just-in-time configuration before the burst arrives. Accordingly, a wavelength is reserved only for the duration of the burst. Without waiting for a positive acknowledgment from the destination node, the burst payload follows its control packet immediately after a predetermined offset time, which is path (hop-count) dependent and theoretically designated as the sum of intra-nodal processing delays. OBS gains the benefits of OCS and OPS. However, its offset-time-based design results in several complications [15]. These OBS design complications have been the primary motivators behind the design of the OCPS paradigm.

While OBS can be viewed as a more efficient variant of OCS; OCPS can be considered as a less stringent variant of OPS. Similar to OBS, OCPS supports per-burst switching, which are labeled-based, QoS-oriented, and either bufferless or with limited FDL-based buffers. However, unlike OBS using out-of-band control, OCPS adopts in-band control in which the header and payload are modulated and transported via the same wavelength. Such header/payload modulation technique, as will be shown, has been particularly designed for and beneficial to OCPS networks. Based on

Manuscript received August 19, 2005; revised January 15, 2006. This work was supported in part by the Phase-II Program for Promoting Academic Excellence of Universities, Taiwan, under Contract NSC94-2752-E009-004-PAE, in part by the NCTU/CCL Joint Research Center, and in part by the National Science Council (NSC), Taiwan, under Grant NSC94-2213-E-009-016.

M. C. Yuang and J. Shih are with the Department of Computer Science and Information Engineering, National Chiao Tung University, Taiwan (email: mcyuang@csie.nctu.edu.tw).

P. L. Tien is with the Department of Communication Engineering, National Chiao Tung University, Taiwan.

S. S. W. Lee, Y. M. Lin, and A. Chen are with the Optical Communications & Networking Technologies Department, Computer & Communications Research Labs, Industrial Technology Research Institute, Taiwan.

F. Tsai was with the Industrial Technology Research Institute. He is now with the University of California, San Diego (UCSD).

Digital Object Identifier 10.1109/JSAC-OCN.2006.22905.

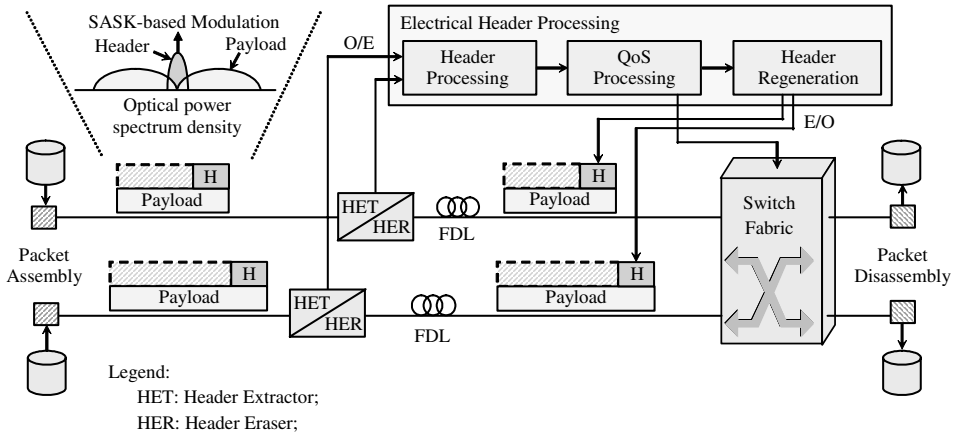


Fig. 1. Optical Coarse Packet Switching (OCPS).

OCPS, we have constructed an experimental optical IP-over-WDM network, referred to as OPSINET. OPSINET has been constructed under the collaborative project between National Chiao Tung University under the MOE Program of Excellence, and Computer/Communications Research Laboratories (CCL)/Industrial Technology Research Institute (ITRI) under the Intelligent Optical Networking project. The main objective is to examine and resolve fundamental OCPS transport and QoS challenges from both the system- and network-layer perspectives. OPSINET consists of three types of nodes- edge routers, optical lambda/fiber switches (OXCs), and Optical Label Switched Routers (OLSRs). To facilitate traffic engineering [16], OPSINET is augmented with an out-of-band Generalized Multiprotocol Label Switching (GMPLS) [17] control network.

The remainder of this paper is organized as follows. In Section 2, we introduce the OCPS paradigm and present the architecture of OPSINET. In Sections 3 and 4, we describe the architectures and operations of the ingress router and OLSR in OPSINET, respectively. The packet-loss performance is also shown in detail in Section 4. In Section 5, we delineate the GMPLS TE framework. Finally, concluding remarks are given in Section 6.

II. Optical Coarse Packet Switched IP-over-WDM Network (OPSINET)

A. Optical Coarse Packet Switching (OCPS)

IP packets in an OCPS network belonging to the same loss class and the same destination are assembled into bursts at ingress routers. As shown in Fig. 1, a header for a burst payload, which carries forwarding (i.e., label) and QoS (e.g., priority) information, is modulated with the payload based on our newly designed Superimposed Amplitude Shift Keying (SASK) technique [18], which will be described later. Besides, they are time-aligned during modulation via necessary padding added to the header. They are re-aligned in switching nodes should the burst be truncated. Notice that such design eliminates the payload length information carried in the header. The entire burst is then forwarded along a pre-established Optical Label Switched Path (OLSP).

At each switching node, the header and payload are first SASK-based demodulated. While the header is extracted and electronically processed, the burst payload with the header erased, remains transported optically in a fixed-length FDL achieving constant delay and data/protocol transparency. Provided with no buffer and that there is more than one burst payload at the switch destined for the same wavelength output, contention occurs and resolution is required. Each burst payload is then SASK-based re-modulated with the new header, and switched according to the label information in the header. Finally at egress nodes, the reverse burstification process is performed and IP packets are extracted from bursts.

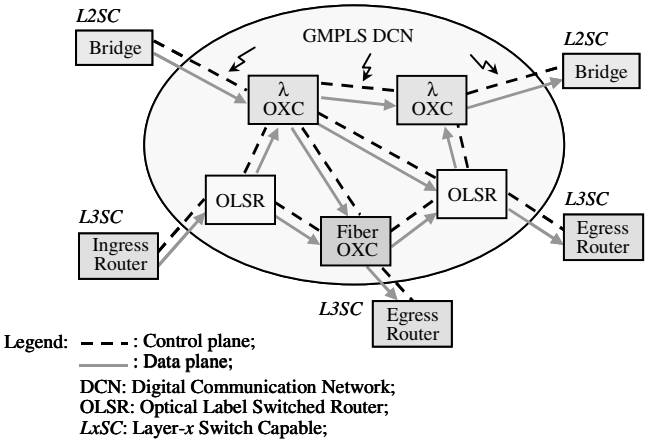


Fig. 2. OPSINET topology.

B. OPSINET Architecture

OPSINET consists of edge nodes (layer-3 routers and layer-2 bridges) that are interconnected via heterogeneous switching nodes, which are lambda/fiber OXCs and OLSRs, with multi-granularity switching capabilities, as shown in Fig. 2. A snapshot of OPSINET is displayed in Fig. 3. While lambda and fiber OXCs are layer-1 optical devices that switch on a single lambda and an entire fiber, respectively, the OLSRs are layer-3 optical nodes that route and switch packets on a *label* basis. The label-based routing and switching in OPSINET is managed by the control plane implemented by an out-of-band Fast-Ethernet-based GMPLS network. The GMPLS network connects a number of GMPLS controllers, each of which

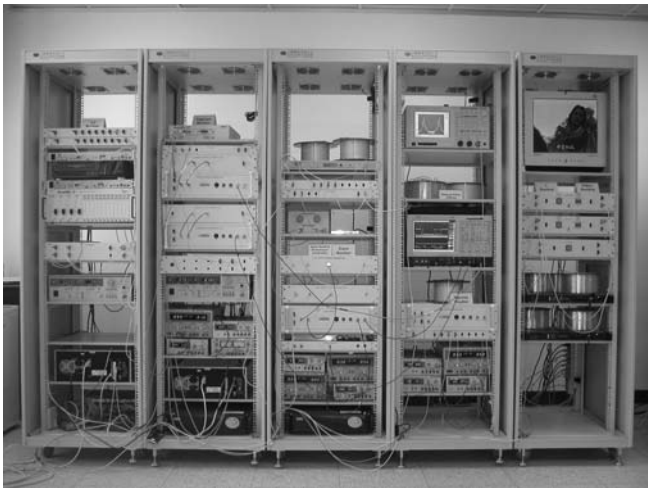


Fig. 3. OPSINET: a snapshot.

governs the routing/switching of an OPSINET node described above. Prior to the data transfer, an end-to-end optical Label Switched Path (LSP) is constructed by means of a routing and wavelength assignment algorithm proposed in a previous paper [16]. Other GMPLS operations are described in detailed in Section 5.

IP traffic is externally generated from SmartBits devices. OPSINET supports 2.5 Gb/s for payload, and 125 Mb/s for header due to easier recovery. Header and payload are encoded by the 8B/10B scheme and multiplexed via the SASK module. The header is 8 bytes long, and consists of six following fields: preamble, Start of Header (SOH), label, priority, Header Error Control (HEC), and End of Header (EOH), other than the padding. The burst payload is greater than or equal to 1500 bytes, which excludes the 68-byte overhead (e.g., preamble, Start/End of Payload), achieving a minimum of 95% efficiency. Specifically within the 64-byte preamble, 16 bytes are used for 2R reshaping, 32 bytes for 3R bit-resynchronization, and 16 bytes for word-resynchronization.

Significantly, the header and payload are time-aligned during modulation and remain aligned even after contention occurs. The rationale behind the design is described as follows. Notice that the payload length information is required for switching and reception processes, and thus has to be contained in the header. However, if contention occurs during switching, the payload is partially damaged. Such length information in the header is no longer valid. Therefore, with the time alignment design, the payload length information can be removed from the header, making the payload of any length recoverable at the receiver. Another side benefit of the design is that, since the header integrity must be maintained at all times, header timing can be used to serve as gating control during the burst-mode receptions of payloads. Such design can effectively alleviate the transient response problem resulting from the presence of back-to-back payloads with different powers.

C. Superimposed Amplitude Shift Keying (SASK) Technique

Since the information that is swapped at each switching node is the *label* inside the header, we use label rather than the header throughout this subsection. In general, SASK superimposes a low-speed ASK label on top of a high-speed DC-balanced line-coded ASK payload. At any intermediate switching node, the old ASK label is erased by modulating the combined payload and label signal with the inverse of the received ASK label. It has been shown [18] that such technique requires only low-speed external modulators and low-speed optical receivers to perform label swapping. As a result, sophisticated phase modulation devices [19] or optical components [20-21], such as MZI-SOA, can be eliminated.

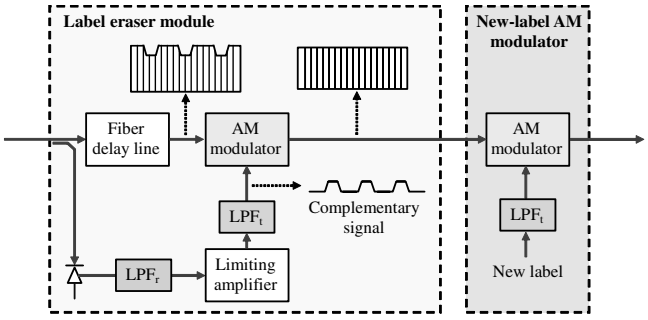


Fig. 4. The label swapping subsystem.

The basic building blocks of an optical transmitter are two-stage intensity modulators. A continuous-wave light source is first modulated by a high-speed Non-Return-to-Zero (NRZ) payload with a large modulation depth. It is subsequently modulated by a low-speed NRZ label with a small modulation depth. A DC-balanced line-encoder was adopted to suppress the low frequency energy of the payload signal. An 8B/10B line code has been adopted due to high practicability and bandwidth efficiency. It is worth noticing that the determination of a proper modulation depth for a label signal is crucial to the system performance. On one hand, a label with a low modulation index can not sustain multi-hop long-distance transmission due to payload interference and transmission noise. On the other hand, a label with a large modulation index may result in a decrease in payload signal power, and thus higher residual noise due to non-ideal label erasers.

At each intermediate switching node, label swapping is performed by an optical label swapping subsystem (Fig. 4) that is composed of a label eraser module and a new-label AM modulator. In the label eraser module, a portion of the input signal is detected through a passive optical tap and a photodiode. A Low Pass Filter (LPFr) at the receiver front end is used to remove the payload signal and out-of-band noise. A limiting amplifier and a Low Pass Filter (LPFt) are then used to provide a constant amplitude and to reshape the received label waveform, respectively. Notice that, the LPFt in the switching node should have a frequency response as close to that of the transmitting-end LPFt in order to inversely compensate the superimposed old label. Should the received label have a low error-rate performance, it can be considered as an analog copy of the original label signal. We use this re-shaped label, called complementary signal, to reverse modulate the optical signal via the AM modulator. Notice that a fiber delay line is placed before the AM modulator to minimize the deterministic phase error between the incoming and complementary signals. Consequently, most

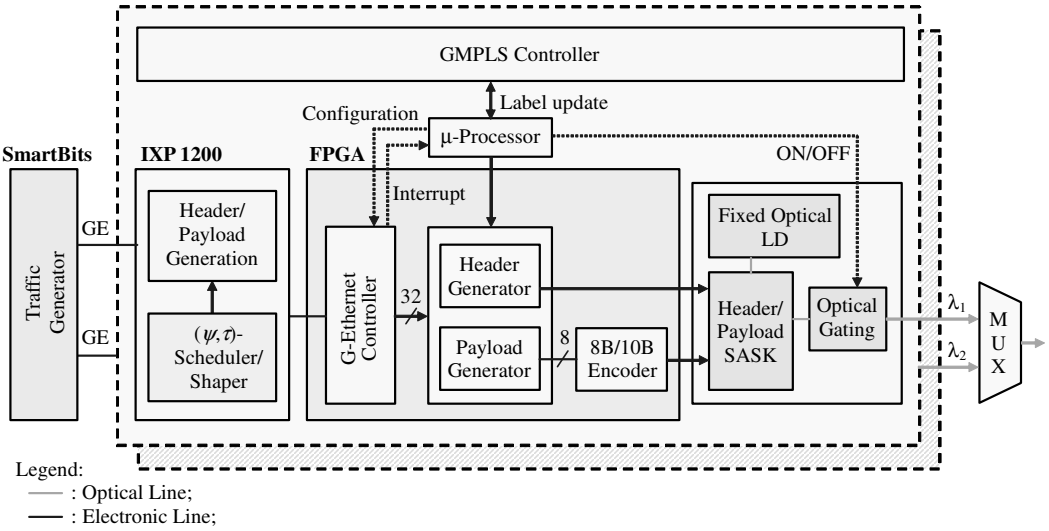


Fig. 5. Ingress router architecture.

of the incoming (old) label can be removed.

It is worth mentioning that the performance of the Label Eraser may be affected by the timing error during matching the path of the fiber delay line and that of the electrical signal (see Fig. 4). In the absence of noise, since all components on both paths are analog and not clock driven, the timing difference is a static value. Thus, timing control can be carried out by a manually tunable optical delay line, which can compensate the delay difference within tens of picoseconds. In the presence of noise, the random timing jitter problem arises from the Limiting amplifier as a result of the amplitude noise, subject to the selections of modulation index, optical amplifier spacing, and received power. A detailed performance analysis in this case has been presented in paper [18]. In general, using a modulation index of 0.22 and an optical received power of -14dBm, the label signal integrity can be fully maintained in 10-hop links [18].

Notice that the SASK technique has been specially designed for OCPS networks in which payloads have much larger data rate and packet length compared to headers. An important design parameter of SASK is the data rate for payload and label. To avoid payload's interference to the low-frequency label, not only the DC-null channel coding is used in payload signal, but also the label's data rate should be relatively low compared to the payload's data rate. In [18], for example, 100 Mb/s label together with 8 Gb/s payload, that is, 8 bytes label and a minimum of 640 bytes payload should be simultaneously sent out within a packet time. Such large payload length requirement for an OPS system becomes impractical. In contrast, in an OCPS system, the payload length that is always greater than 1500 bytes, rendering SASK superior for modulation and label swapping for OCPS networks.

III. Edge Router Architecture and Operations

The operations in ingress and egress routers differ in burstification and payload recovery. While the ingress router simply performs burstification, the egress router recovers the payload followed by the reverse burstification process. Since payload recovery is similar to header recovery provided with sufficient preamble, we only describe the architecture and operations of ingress routers.

A. Architecture

The ingress router consists of five major components (see Fig. 5): (ψ,τ)-Scheduler/Shaper, Gigabit-Ethernet (GE) Controller, Header/Payload Generator, 8B/10B Encoder, and SASK Optical Transmitter, in addition to the GMPLS controller and μ-processor interface. First, the (ψ,τ)-Scheduler/Shaper [15], which is implemented in an Intel IXP1200 network processor, performs QoS-enabled packet aggregation, with the aim of providing delay and loss class differentiations for OPSINET. Specifically, ψ and τ are the maximum burst size and maximum burst assembly time, respectively. A burst is generated and transmitted either when the burst size reaches ψ or τ expires. The scheme will be described in more detail in the next subsection. After having determined the packets to be aggregated, the Header/Payload Generation module in IXP1200 in turn performs Simple Data Link (SDL)-based [22] framing for packet delineation and recovery during the burstification process. Specifically, before placing each IP packet into the burst, the packet is encapsulated with a two-octet Packet Length Indication (PLI) and a two-octet header CRC, and is followed by a two-octet Frame Check Sequence (FCS).

It is worth noticing that there are three advantages of using the SDL-based framing protocol. First, with the PLI field, variable-size packets can be supported. Second, without such framing, any single bit error in the IP length field may result in false packet delineation from the remaining packets, namely error propagation, inside the burst. With the CRC field of SDL, the error propagation problem can be eliminated. Finally, notice that when a burst collision occurs, packets in the tail part of one burst are lost and the last received packet might be incomplete. Through the inconsistent PLI information and a missing FCS, such framing protocol allows the damaged incomplete packet to be identified, and thus the remaining packets to be recovered.

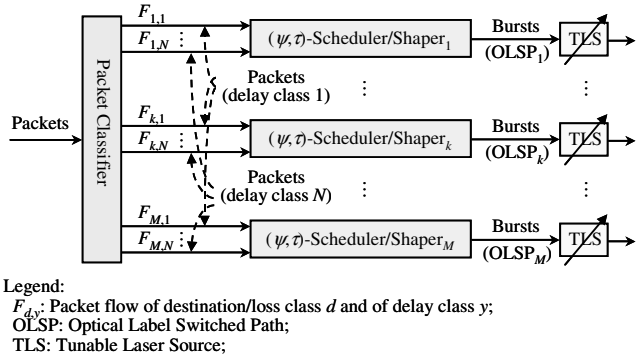


Fig. 6. (ψ,τ)-Scheduler/Shaper system architecture.

GE Controller provides the GE interface between IXP1200 and the FPGA, and ultimately passes the header and payload in parallel to the Header/Payload Generator within the FPGA. The payload is then encoded via the 8B/10B Encoder. At the output stage, the header and payload are SASK-based modulated and optically transmitted via an available wavelength. A snapshot of the FPGA part of an ingress router is displayed in Fig. 9(a).

B. (ψ,τ)-Scheduler/Shaper

Essentially, (ψ,τ)-Scheduler/Shaper is a dual-purpose scheme. It is a scheduler for packets, abbreviated as (ψ,τ)-Scheduler, which performs the scheduling of different delay class packets into back-to-back bursts. On the other hand, it is a shaper for bursts, referred to as (ψ,τ)-Shaper, which determines the sizes and departure times of bursts.

In any ingress node, incoming packets (see Fig. 6) are first classified on the basis of their destination, loss, and delay classes. Packets belonging to the same destination and loss class are assembled into a burst. Thus, a burst contains packets of various delay classes. In the figure, we assume there are M destination/loss classes and N delay classes in the system. For any one of M destination/loss classes, say class k, packets of flows belonging to N different delay classes are assembled into bursts through (ψ,τ)-Scheduler/Shaper$_k$ according to their pre-assigned delay-associated weights. Departing bursts from any (ψ,τ)-Scheduler/Shaper are optically transmitted, and forwarded via their corresponding, pre-established OLSP.

To provide delay class differentiation, for IP packet flows designated with delay-associated weights, (ψ,τ)-Scheduler performs packet scheduling and assembly into bursts based on their weights and a virtual window of size ψ. A flow of a higher delay priority class is given a greater weight, which corresponds to a more stringent delay bound requirement. The weight of a flow corresponds to the maximum number of packets (i.e., the credits) that can be accommodated in a window (or burst in this case) for the flow. For a flow with sufficient credits, its new packets are placed in the current window on a FIFO basis. Otherwise, its packets are placed in an upward appropriate window in accordance to the total accumulated credits. Such window-based scheduling allows simple FIFO service within the window and assures weight-proportional service at the window boundary.

To provide loss class differentiation, (ψ,τ)-Shaper facilitates traffic shaping with larger burst sizes assigned to higher loss priority classes. Analytical and experimental results [15] have shown that (ψ,τ)-Shaper yields substantial reduction, proportional to the burst size, in the coefficient of variation of the burst inter-departure time. Consequently, with (ψ,τ)-Shaper, the OCPS networks achieves more than five orders of magnitude reduction in burst loss probability under a traffic load of 0.8, ψ=100, and 50 wavelengths. The improvement of loss probability is even more compelling in the presence of a large number of wavelengths (W=100) due to higher statistical multiplexing gain.

It is worth pointing out that such traffic shaping is a preventive traffic control means to reduce the burst/packet loss probability. As will be shown in the next section, we also adopt reactive traffic control, namely prioritized contention resolution, at OLSRs in the presence of contention.

IV. Optical Label Switched Routers (OLSRs)

A. Architecture and Operations

The OLSR (see Fig. 7) consists of three major components for each input port (fiber), and one cyclic-frequency AWG switch for the entire node. The three components are: Header Extractor/Eraser, Burst Mode Receiver for header (BMR$_H$), and FPGA-based Core Switch Controller (CSC). First, the Header Extractor/Eraser extracts the header, and erases the header for the payload, by means of the SASK-based demodulation technique previously described. While the payload continues traveling optically along the internal FDL, the header is received and recovered (2R) in amplitude by BMR$_H$. The data recovery (3R) is then performed by burst-mode Clock Data Recovery (CDR) in a Xilinx Virtex-II 3000 FPGA via over-sampling the header with different phases.

With the recovered header, CSC performs label swapping, QoS control, and laser tuning control. First, notice that owing to the use of an AWG switch, once an OLSP is established, the path is determined locally via the binding from an old label to a new (label, wavelength) pair. All label and wavelength information have been in advance downloaded from GMPLS Controller through the μ-processor and saved in Content Addressable Memory (CAM). With CAM, label swapping is accomplished in three clock cycles.

Second, the QoS Control Processor (QCP) is responsible for prioritized contention resolution and header integrity assurance. It is worth noting that, due to AWG, any two bursts arriving from different input ports never contend. On the contrary, contention will occur for bursts arriving from the same input port but carried by different wavelengths, and destined for the same output port. Basically, to switch a burst to the destined output port, an idle wavelength is selected. If all wavelengths are busy, higher priority bursts receive absolute precedence over lower-priority bursts. That is, owing to bufferless, one of the lower-priority bursts being served is preempted and discarded. It is worth noting that if partially destructed lower-priority bursts are still transmitted, the loss probability can be much improved.

Such preemption resolution however raises a problem in which the header may be damaged resulting from contention.

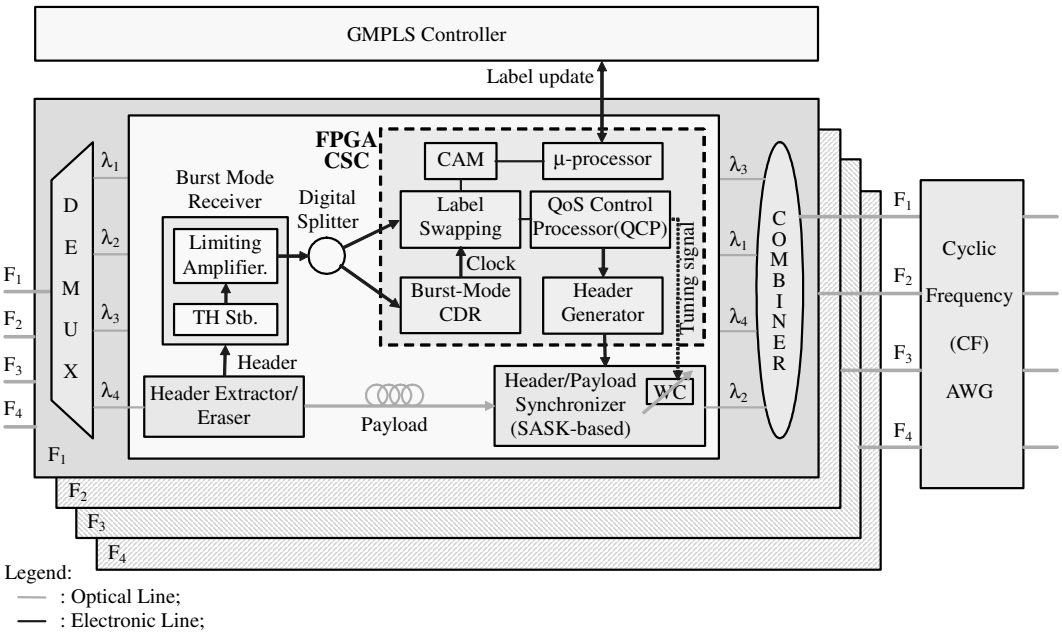


Fig. 7. Optical Label Switched Router (OLSR) architecture.

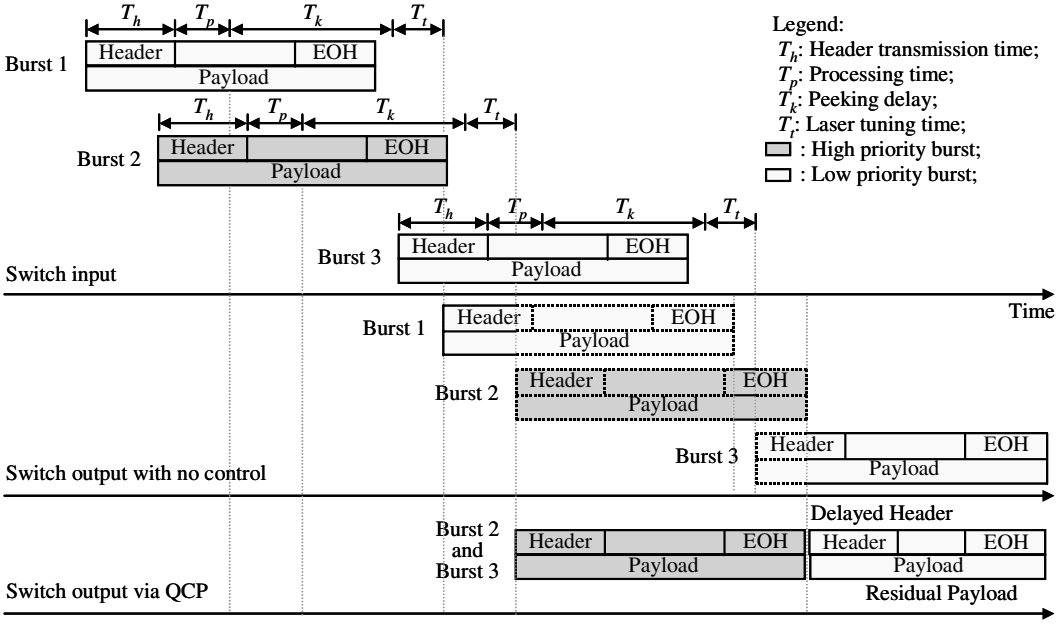
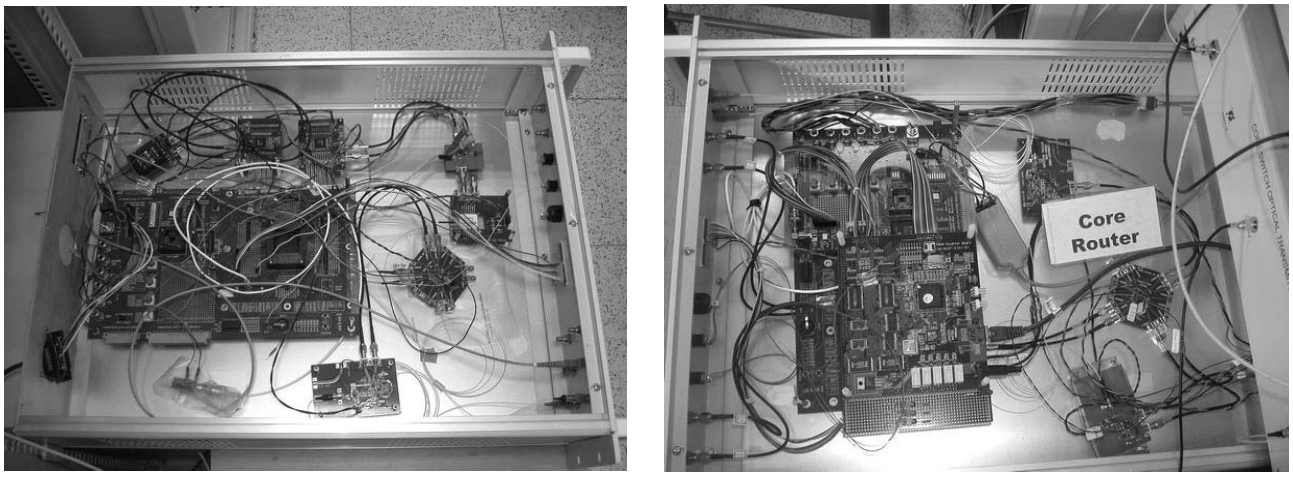


Fig. 8. QoS Control Processor (QCP): header integrity and contention resolution.

To assuring header integrity, QCP employs the following means. We first consider contention under no priority. For the ease of description, let T_h, T_p, and T_t denote the header transmission time, header processing time, and laser tuning delay, respectively. Notice that the header transmission time excludes that of the padding. If two payloads are distanced by at least T_h, the header can always be protected since the transmission of the first header is finished before that of the second header. However, the problem arises when two payloads are distanced by less than T_h. The problem is solved if such potential contention can be identified before the first header gets transmitted, i.e., if an extra delay, called the peeking delay (T_k), is imposed after the header is processed. Thus, header integrity can be maintained if $T_h+T_p+T_k+T_t > D + T_h + T_p$, where D is the distance between two bursts, and $0 \leq D \leq T_h$. The peeking delay can be assigned as: $T_k = \max(D - T_t) = T_h - T_t$.

With the peeking delay imposed, the operation of prioritized contention resolution with the support of preempted, partially collided bursts taken into consideration is described via three scenarios, as shown in Fig. 8. In the first scenario, a high-priority burst arrives after a low-priority burst by a distance of less than T_h. With no control, the headers and payloads of both bursts are damaged. With QCP, only the high-priority burst is transmitted in full. In the second scenario, a low-priority burst arrives after a high-priority burst by a distance



(a) (b)

Fig. 9. Ingress router and OLSR (snapshots): (a) FPGA of an ingress edge router; (b) Optical Label Switched Router (OLSR).

of greater than T_h. With control, the high-priority burst is first fully transmitted. To support partially collided bursts, QCP continues to transmit the remaining low-priority payload attached with a complete aligned header. In the last scenario, a high-priority burst arrives after a low-priority burst by greater than T_h. Not perceiving the arrival of the high-priority burst, QCP first transmits the low-priority burst. However, after identifying a potential collision, QCP terminates the low-priority burst transmission with the attachment of EOH, and transmits the high-priority burst in full.

Finally, with the new (label, wavelength) pair read from CAM, CSC generates the new header and sends a tuning signal to the gated tunable laser. The new header is re-synchronized with the payload having been traveled within the FDL. In OPSINET, the payload is imposed a total delay of 1.2 μs in FDL. It is comprised of 512 ns for header recovery, 100 ns for header processing, 362 ns for peeking delay, 76 ns for QoS processing, and 150 ns for laser tuning. A snapshot of an OLSR is displayed in Fig. 9(b).

B. Packet Loss Performance: Analytical and Simulation Results

We now draw a comparison of the loss probability with multiple priority class between QCP and a prioritized queueing system described in the sequel. Notice that the traffic entering QCP has been previously shaped via the (ψ,τ)-Scheduler/Shaper at the ingress routers. The prioritized queueing system we analyze contains Y priority classes with K wavelengths under Poisson arrivals and exponentially distributed service, namely an $\mathbf{M/M/K/K}$ **loss system with Y priorities**. In such system, a high-priority burst preempts a randomly selected lower-priority burst if all wavelengths are found busy upon arrival. Let λ_i and μ_i denote the arrival and service rates of class i, respectively. Class i has higher priority than class j if $i < j$. Let random variable $\tilde{n}_i(\geq 0)$ denote the total number of class-i bursts in the system. The system state is represented by Y-tuple $(\tilde{n}_1, \tilde{n}_2, \ldots, \tilde{n}_Y)$, where $\sum_{i=1}^{Y} \tilde{n}_i \leq K$. The loss probability for each class, say i, denoted as LP_i , can be derived from the limiting system distribution $\Pi = \{\pi_{n_1,\ldots,n_Y}, \sum_{i=1}^{Y} n_i \leq K\}$, where $\pi_{n_1,\ldots,n_Y}$ is the joint distribution of the Y-tuple.

The limiting distribution is solved based on two sets of balance equations- one corresponds to a system with at least one available server ($\sum_{i=1}^{Y} n_i < K$), and the other one corresponds to a busy system ($\sum_{i=1}^{Y} n_i = K$). Through derivation, they can be given as:
Case I: $\sum_{i=1}^{Y} n_i < K$,

$$\pi_{n_1,\ldots,n_Y}\left[\sum_{i=1}^{Y}(\lambda_i + n_i\mu_i)\right] =$$

$$\sum_{i=1}^{Y}\left[\lambda_i \pi_{n_1,\ldots,n_i-1,\ldots,n_Y} + (n_i+1)\mu_i \pi_{n_1,\ldots,n_i+1,\ldots,n_Y}\right]; \quad (1)$$

Case II: $\sum_{i=1}^{Y} n_i = K$,

$$\pi_{n_1,\ldots,n_Y}\sum_{i=1}^{Y}(\lambda_i^* + n_i\mu_i) = \sum_{i=1}^{Y}\lambda_i \pi_{n_1,\ldots,n_i-1,\ldots,n_Y}$$
$$+\sum_{i=1}^{Y-1}\left[\lambda_i \sum_{j=i+1}^{Y}\frac{n_j \pi_{n_1,\ldots,n_i-1,\ldots,n_j+1,\ldots,n_Y}}{\max(\sum_{l=i+1}^{Y} n_l, 1)}\right], \quad (2)$$

where $\lambda_i^* = \lambda_i$ if $\sum_{l=i+1}^{Y} n_l > 0$, otherwise $\lambda_i^* = 0$.

The left hand sides of Equations (1) and (2) differ in that the non-busy system allows any arrival of any class, whereas a busy system only permits a preemption of a lower-priority burst (if it exists) by a higher-priority burst. Moreover, at the right hand side of Equation (2), the second term indicates the preemption of class j by class i, making the size of class-j reduced by one and the size of class i incremented by 1. The probability of being preempted is proportional to the size of the class. Finally, a burst is lost if either the burst arrives at a busy system and there is no lower-priority burst that can be preempted, or the burst is later preempted by another newly arriving burst with higher priority. Accordingly, we obtain

$$LP_i = \sum_{n_1+\ldots+n_i=K} \pi_{n_1,\ldots,n_i,0,\ldots,0}$$

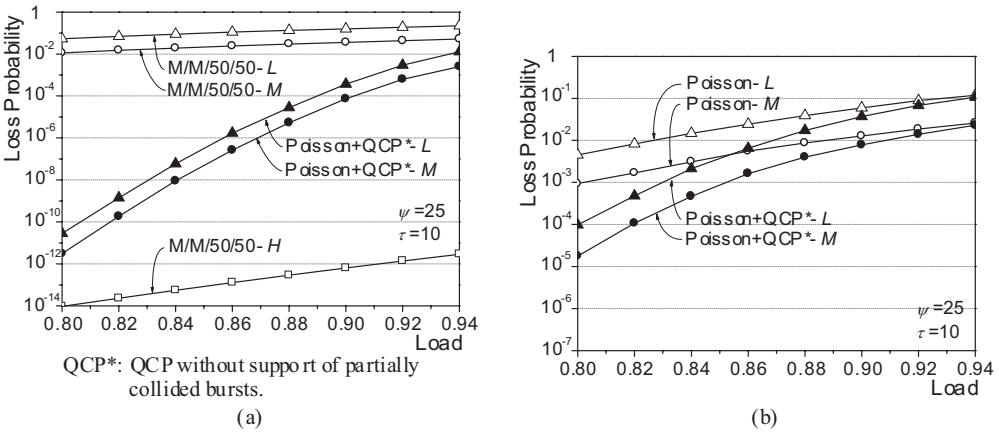


Fig. 10. Performance of QCP (without support of partially collided bursts): (a) single node; (b) 24-node network.

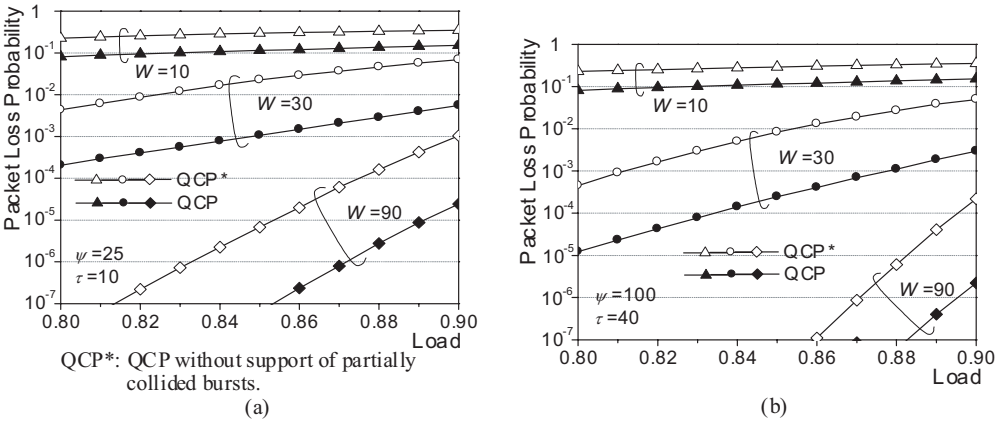


Fig. 11. Performance of QCP (with and without support of partially collided bursts): (a) smaller burst size (ψ=25); (b) greater burst size (ψ=100).

$$+\frac{1}{\lambda_i}\sum_{j=1}^{i-1}\left[\lambda_j \sum_{n_1+\ldots+n_Y=K} \frac{n_i \pi_{n_1,\ldots,n_Y}}{\max(\sum_{l=j+1}^{Y} n_l, 1)}\right]. \quad (3)$$

We draw comparisons of loss probability between the M/M/50/50 and QCP systems supporting three priorities. In the simulation, we computed the loss probability of the ARPANET network with 24 nodes and 48 links, in which 14 nodes are randomly selected as edge routers. There are 50 wavelengths (1 Gb/s per wavelength) on each link, and the wavelength is randomly assigned. OLSPs are determined subject to load balance of the network. Under any given load, say A, the total amount of traffic injected to the network is 50A Gb/s. Analytical and simulation results are plotted in Fig. 10. Under both cases as shown in Fig. 10, compared to the M/M/50/50 system, the QCP system yields superior performance for all three classes, due to traffic shaping. Notice that, due to super low loss probability for the H class (lower than 10^{-14} under load 0.94), the plotting is omitted in the figure.

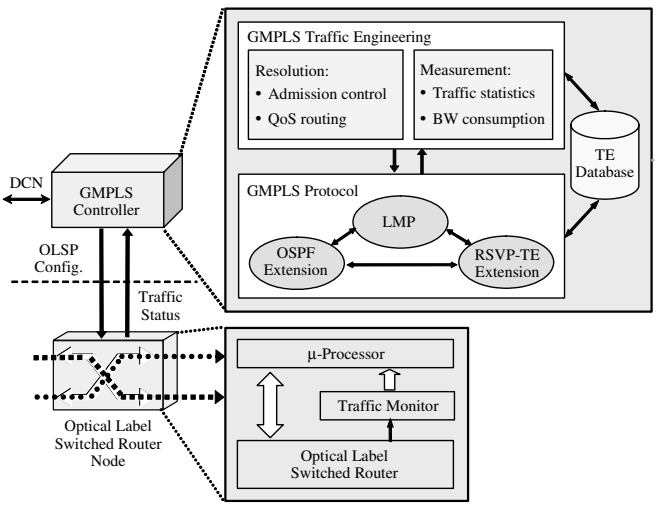


Fig. 12. OLSR node and its GMPLS controller.

Finally, we examine the performance of QCP with respect to packet loss probability with and without supporting partially collided bursts, under two different ψ values, and three different numbers of wavelengths (W=10, 30, and 90). In the simulation, there are two priorities (H and L) of traffic, both of which are IPP distributed with b=4 and shaped via (ψ,τ)-Shaper. At OLSRs, a higher priority burst that finds no wavelength available upon arrival will preempt the lower-priority burst with the least remaining service time. Simulation results of packet loss probability for L-class traffic are displayed in Fig. 11. As was expected, the loss performance is noticeably improved with QCP. Specifically, the loss probability declines by more than two orders of magnitude under loads of 0.9 and below, ψ=100 and W=90.

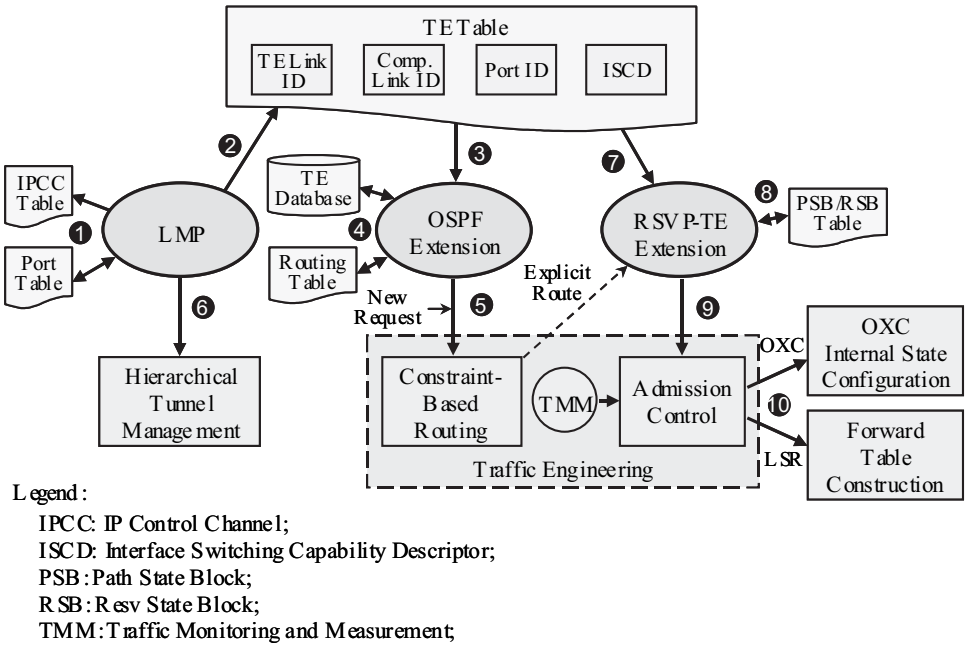


Fig. 13. GMPLS traffic engineering and protocols.

V. GMPLS Traffic Engineering (TE)

OPSINET is facilitated with GMPLS TE and control via a Fast-Ethernet network with the same topology as that of the data plane of OPSINET (shown in Fig. 2). Each switch/router in the data plane is connected to a GMPLS controller in the control plane, which performs either OXC configuration for optical lambda/fiber switches, or traffic engineering and control for OLSRs. Fig. 12 illustrates the major functions and interface of an OLSR and its GMPLS controller. To support TE, an OLSR node is equipped with a traffic monitor, periodically passing traffic status information to the GMPLS controller via the μ-processor interface. In response, the GMPLS controller makes frequent updates to the TE database for determining and establishing OLSPs upon new connection requests arrive.

The GMPLS controller is composed of two parts: TE and control protocols. In TE, the measurement function collects traffic status, derives statistics, and computes statistical bandwidth consumption. Such statistical bandwidth information is not only updated in the local TE database but also made available to other OLSRs through the GMPLS OSPF-extension flooding protocol. The resolution function, which is the brain of the GMPLS controller, performs admission control and QoS routing for establishing new OLSPs. The establishment requires full participation of three GMPLS protocols working together (detailed next) to provide unified control across optical lambda/fiber switches and OLSRs. As a result, the OLSP is associated with Inbound and Outbound (I/O) (port, wavelength) pairs at each lambda OXC, I/O ports at each fiber OXC, and I/O (label, wavelength) pairs at each OLSR. Notice that the biding of wavelength in the last OLSR case is due to the use of AWG switch in OPSINET. Finally, an OLSP configuration command carrying new (label, wavelength) pairs is triggered and passed to the OLSR, which ultimately makes the CAM updates for efficient data forwarding during burst transmissions.

The operations and interworking of three GMPLS control protocols (i.e., LMP, OSPF-extension, and RSVP-TE extension) are further elucidated in a sequence of procedures shown in Fig. 13. First, LMP consists of two main tasks-control channel management, and link property correlation. Control channel management constructs (by "config") and frequently maintains (by "hello") control link connectivity between physically adjacent nodes. An IP Control Channel (IPCC) table is created and managed. Link property correlation is used to exchange the local and remote interface property mapping for the data channel, maintaining the Port table for the later port-biding use during OLSP management. Second, a crucial result of the link correlation task is the formation of the TE table, which is served as the common database manipulated by the three protocols during the operations.

Third, the TE table includes TE link ID, component link ID, port ID, and interface switching capability descriptor (ISCD). The TE link, which is formed by the aggregation of component links, is used for efficient management purpose. Component links are multiplexing-capable data-bearing links. Ports, on the other hand, are also data-bearing links but multiplexing incapable. In the case between photonic switches, ports are fibers and component links are the wavelengths within each fiber. In OPSINET, for hierarchical tunnel management (described in procedure six), the component link, or lambda ID is designated as the channel ID defined in International Telecommunications Union Telecommunication (ITU-T). The last ISCD field indicates the type of a switch/router. In our case, it can be layer-3 switch capable (OLSR), or lambda/fiber switch capable (lambda/fiber OXC). Forth, with the TE table produced, OSPF extension is responsible for the distributed construction and the maintenance of the TE database through the original flooding protocol to sustain a global view at each switch/router in the network. The result of the task is the

(a) (b)

Fig. 14. An OPSINET experiment (video data transfer): (a) data transfer state when the OLSP is up; (b) idle state when the OLSP is down.

building of TE database and the forward table. As shown in the fifth step, the major work behind OSPF extension is the conduction of route computation, such as Constraint-Based Routing (CBR), in response to new requests for OLSPs.

In OLSP management, due to the possibility of possessing dissimilar switching capabilities in two neighboring nodes, it requires a means of creation and maintenance of a hierarchy of OLSPs. Accordingly as shown in the sixth step, hierarchical tunnel management has been proposed as another task performed in LMP. Seventh, based on the RSVP signaling protocol, RSVP-TE extension has been proposed to encompass OLSP management for optical networks with OXCs. Major enhancements implemented in OPSINET include: the use of generalized labels, establishment of hierarchical OLSPs (with the aid of LMP) through PATH and RESV messages, and the use of a suggested label that is passed through an explicit route computed by OSPF in advance. RSVP-TE constructs the PSB/RSB table as part of OLSP set-up task, as shown in the eighth step. Ninth, the major work behind OLSP establishment is the admission control intelligence for making the connection acceptance or rejection decision. A potential solution is the measurement-based admission control, which is still an ongoing work in OPSINET. Finally, in the tenth step, should an OLSP be successfully established, the OXC internal state is configured in case of a lambda/fiber switch. Otherwise under the OLSR case, the forward table with label mapping information is constructed.

Upon having established the OLSPs, bursts are transmitted from ingress nodes to egress nodes. In Fig. 14, we display the snapshots taken at an egress node for a video data transfer experiment over OPSINET. Part (a) of the figure shows video playout while the OLSP is up, and Part (b) display the absence of data received after the OLSP has been terminated.

VI. Conclusion

In this paper, we have presented the architecture of OPSINET, an IP-over-WDM experimental network operating at a data rate of 2.5 Gbps per wavelength, based on an optical coarse packet switching (OCPS) paradigm. The OCPS paradigm advocates the enforcement of traffic control to realize bandwidth-on-demand on sub-wavelength basis. In the basic transport, OPSINET performs efficient per-burst switching by means of the time-aligned design and SASK-based modulation of the header and burst payload. At ingress routers, (ψ,τ)-Scheduler/Shaper performs scalable traffic scheduling and shaping providing delay and loss differentiation for the network. At OLSRs, prioritized contention resolution is exerted with the support of partially collided bursts, while header integrity is maintained at all times. Through this experiment, we perceive that data-centric optical Internet can become a reality based on the OCPS technology.

Acknowledgment

The authors would like to thank the team members, including Y. L. Chou, Y. Chen, D. Z. Hsu, Y. T. Wang, S. K. Yang, Y. C. Liaw, J. J. Duanmu, and Y. Y. Chang, for their significant contribution to the OPSINET project. The authors would also like to thank Dr. Sheng Ching Jeng, Director, and Dr. Chung H. Lu, vice President of CCL/ITRI, for their abiding support of the work. Furthermore, the authors would especially like to thank Dr. Tingye Lee and Prof. G. K. Chang for their valuable technical advice and encouragement throughout the construction of the testbed. Finally, the authors gratefully acknowledge the constructive suggestions of the anonymous reviewers.

References

[1] B. Mukherjee, "WDM optical communication networks: progress and challenges," *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, pp. 1810-1824, Oct. 2000.

[2] T. El-Bawab and and J.-D. Shin, "Optical packet switching in core networks: between vision and reality," *IEEE Commun. Mag.*, vol. 40, no. 9, pp. 60-65, Sept. 2002.

[3] F. Callegati, G. Corazza, and C. Raffaelli, "Exploitation of DWDM for optical packet switching with quality of service guarantees," *IEEE J. Select. Areas Commun.*, vol. 20, no. 1, pp. 190-201, Jan. 2002.

[4] H. Dorren *et al.*, "Optical packet switching and buffering by using all-optical signal processing methods," *J. Lightwave Technol.*, vol. 21, no. 1, pp. 2-12, Jan. 2003.

[5] T. Battestilli and H. Perros, "An introduction to optical burst switching," *IEEE Commun. Mag.*, vol. 41, no. 8, pp. S10-S15, Aug. 2003.

[6] M. Yoo, C. Qiao, and S. Dixit, "Optical burst switching for service differentiation in the next generation optical Internet," *IEEE Commun. Mag.*, vol. 39, no. 2, pp. 98-104, Feb. 2001.

[7] V. Vokkarane and J. Jue, "Prioritized burst segmentation and composite burst-assembly techniques for QoS support in optical burst-switched networks," *IEEE J. Select. Areas Commun.*, vol. 21, no. 7, pp. 1198-1209, Sept. 2003.

[8] J. Wei and R. McFarland, "Just-in-time signaling for WDM optical burst switching networks," *J. Lightwave Technol.*, vol. 18, no. 12, pp. 2019-2037, Dec. 2000.

[9] M. Yoo, C. Qiao, and S. Dixit, "QoS performance of optical burst switching in IP-over-WDM networks," *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, pp. 2062-2071, Oct. 2000.

[10] J. White, M. Zukerman, and H. Vu, "A framework for optical burst switching network design," *IEEE Commun. Lett.*, vol. 6, no. 6, pp. 268-270, June 2002.

[11] N. Barakat and E. H. Sargent, "Dual-header optical burst switching: a new architecture for WDM burst-switched networks," in *Proc. IEEE INFOCOM 2005*, pp. 685-693.

[12] Z. Rosberg, Hai Le Vu, M. Zukerman, and J. White, "Performance analyses of optical burst-sswitching networks," *IEEE J. Select. Areas Commun.*, vol. 21, no. 7, pp. 1187-1197, Sept. 2003.

[13] A. Ge, F. Callegati, and L. Tamil, "On optical burst switching and self-similar traffic," *IEEE Commun. Lett.*, vol. 4, no. 3, pp. 98-100, March 2000.

[14] M. Yuang, J. Shih, and P. Tien, "QoS burstification for optical burst switched WDM networks," in *Proc. IEEE OFC 2002*, pp. 781-783.
[15] M. Yuang, P. Tien, J. Shih, and A. Chen, "QoS scheduler/shaper for optical coarse packet switching IP-over-WDM networks," *IEEE J. Select. Areas Commun.*, vol. 22, no. 9, pp. 1766-1780, Nov. 2004.
[16] S. Lee, M. Yuang, P. Tien, and S. Lin, "A Lagrangean relaxation based approach for routing and wavelength assignment in multi-granularity optical WDM networks," *IEEE J. Select. Areas Commun.*, vol. 22, no. 9, pp. 1741-1751, Nov. 2004.
[17] E. Mannie *et al.*, "Generalized multi-protocol label switching (GMPLS) architecture, draft-ietf-ccamp-gmpls-architecture-07.txt, Nov. 2003.
[18] Y. Lin, M. Yuang, S. Lee, and W. Way, "Using superimposed ASK label in a 10 Gbps multi-hop all-optical label swapping system," *J. Lightwave Technol.*, vol. 22, no. 2, pp. 351-361, Feb. 2004.
[19] Z. Zhu, Z. Pan, and S. J. B. Yoo, "A compact all-optical subcarrier label-swapping system using an integrated EML for 10-Gb/s optical label-switching networks," *IEEE Photon. Technol. Lett.*, vol. 17, pp. 426-428, Feb. 2005.
[20] N. Deng, Y. Yang, C. K. Chan, W. Hung, and L. K. Chen, "Intensity-modulated labeling and all-optical label swapping on angle-modulated optical packets," *IEEE Photon. Technol. Lett.*, vol. 16, pp. 1218-1220, Apr. 2004.
[21] N. Chi, J. Zhang, P. Holm-Nielsen, C. Peucheret, and P. Jeppsesn, "Transmission and transparent wavelength conversion of an optically label signal using ASK/DPSK orthogonal modulation," *IEEE Photon. Technol. Lett.*, vol. 15, pp. 760-762, May 2003.
[22] B. Doshi *et al.*, "A simple data link (SDL) protocol for next generation packet network," *IEEE J. Select. Areas Commun.*, vol. 18, no. 10, pp. 1825-1837, Oct. 2000.

Maria C. Yuang received the B.S. degree in applied mathematics from the National Chiao Tung University, Taiwan, in 1978; the M.S. degree in computer science from the University of Maryland, College Park, Maryland, in 1981; and the Ph.D. degree in electrical engineering and computer science from the Polytechnic University, Brooklyn, New York, in 1989. From 1981 to 1990, she was with AT&T Bell Laboratories and Bell Communications Research (Bellcore), where she was a member of technical staff working on high speed networking and protocol engineering. She was also an Adjunct Professor at the Department of Electrical Engineering, Polytechnic University, during 1989-1990. In 1990, she joined National Chiao Tung University, Taiwan, where she is currently a Professor of the Department of Computer Science and Information Engineering. Her current research interests include optical and broadband networking, wireless local/access networking, multimedia communications, and performance modeling and analysis.

Steven S. W. Lee received the Ph.D. degree in electrical engineering from National Chung Cheng University, Taiwan, in 1999. He joined Computer & Communications Laboratories of Industrial Technology Research Institute (CCL/ITRI), Taiwan, in fall of 1999, where he is mainly involved in the projects of optical communications. Since 2004, he has been with NCTU-CCL Joint Research Center, National Chiao Tung University, Taiwan, where he is currently a research associate professor. His research interests include optical networks, network planning, and mathematical programming.

Po-Lung Tien received the B.S. degree in applied mathematics, the M.S. degree in computer and information science, and the Ph.D. degree in computer and information engineering, from the National Chiao Tung University, Taiwan, in 1992, 1995, and 2000, respectively. In 2005, he joined National Chiao Tung University, Taiwan, where he is currently an assistant professor of the department of communication engineering. His current research interests include optical networking, wireless networking, multimedia communications, performance modeling and analysis, and applications of soft computing.

Yu-Min Lin received the B.S. degree in electrical engineering from National Tsing-Hua University, Taiwan, R.O.C., in 1996 and the Ph.D. degree in communication engineering from National Chiao-Tung University, Hsinchu, Taiwan, R.O.C., in 2003. He joined the Department of Optical Communications and Networks, Industrial Technology Research Institute (ITRI), Taiwan, R.O.C., in 2004. His research interests include broad-band optical networking and optical packet switching.

Julin Shih received the B.S. degree in management and information system from the National Central University, Chung-li, Taiwan, in 1999 and he received the M.S. degree in 2001 and currently a Ph.D. candidate, in computer science and information engineering from the National Chiao Tung University, Hsin-chu, Taiwan. His currently research interests include high speed networking, optical networking, and performance modeling and analysis.

Frank Tsai is currently pursuing his Ph.D. degree in University of California, San Diego (UCSD). He received his BS and MS degree in Electrical Engineering from National Chiao-Tung University, Hsinchu, Taiwan, in 1999 and 2001 respectively. From 2001 to 2005, he worked as an engineer in Computer and Communication Research Laboratory, Industrial Technology Research Institute, Hsinchu, Taiwan.

Alice Chen received the B.S. degree in electronics engineering from the National Chiao Tung University, Taiwan in 1984 and the M.S. degrees in computer science and information engineering from the National Chiao Tung University, Taiwan in 1992. Currently, she is a senior engineer at Computer & Communications Research Laboratories of Industrial Technology Research Institute, Taiwan, where she works on network control and management of optical networks.