Supporting differentiated classes of service in Ethernet passive optical networks

Glen Kramer and Biswanath Mukherjee

Department of Computer Science, University of California, Davis, California 95616
kramer@cs.ucdavis.edu; mukherjee@cs.ucdavis.edu

Sudhir Dixit and Yinghua Ye

Nokia Research Center, 5 Wayside Road, Burlington, Massachusetts 01803
sudhir.dixit@nokia.com; yinghua.ye@nokia.com

Ryan Hirth

Terawave Communications, 755 Baywood Drive, Petaluma, California 94954
rhirth@terawave.com

Received 9 June 2002; revised manuscript received 4 July 2002

Ethernet passive optical networks (EPONs) are being designed to deliver multiple services and applications, such as voice communications, standard and high-definition video (STV and HDTV), video conferencing (interactive video), real-time and near-real-time transactions, and data traffic. To support these applications with their diverse requirements, EPONs need to have class-of-service (CoS) mechanisms built in. Here we investigate how the Multipoint Control Protocol (MPCP)—an EPON transmission arbitration mechanism—can be combined with a strict (exhaustive) priority scheduling that is a default scheduling algorithm specified in the Institute of Electrical and Electronics Engineers (IEEE) 802.1D standard. Specifically, packet delays for different classes of traffic are analyzed. We find that the queuing delay for lower-priority classes increases when the network load decreases (a phenomenon we call light-load penalty). We also suggest and analyze two different optimization schemes that eliminate the light-load penalty.

OCIS codes: 060.2330, 060.4250.

1. Introduction

Ethernet passive optical networks (EPONs), which represent the convergence of low-cost Ethernet equipment and low-cost fiber infrastructure, appear to be the best candidate for the next-generation access network.

A PON is a point-to-multipoint (PtMP) optical network with no active elements in the signals' path from source to destination. The only interior elements used in a PON are passive optical components, such as optical fiber, splices, and splitters. All transmissions in a PON are performed between an optical line terminal (OLT) and optical network units (ONUs) (Fig. 1). The OLT resides in the telecom central office (CO) and connects the optical access network to the metropolitan-area network (MAN) or wide-area network (WAN). The ONU is located either at the end-user location [fiber-to-the-home (FTTH) and fiber-to-the-business (FTTB) configurations], or at the curb, resulting in fiber-to-the-curb (FTTC) architecture. In the downstream direction, EPON is a broadcasting media; Ethernet packets transmitted by the OLT pass through a 1:*N* passive splitter and reach each ONU. An ONU

discards packets not destined to it before passing the rest of them to a user (Fig. 1).

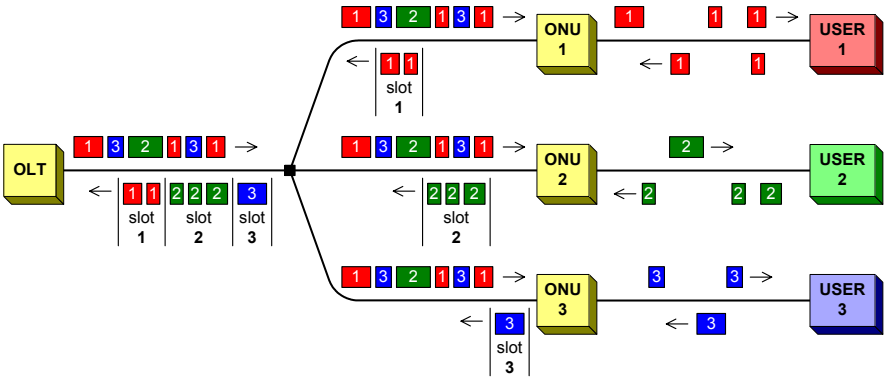

Fig. 1. Upstream and downstream transmissions in EPON.

In the upstream direction (from the ONUs to the OLT) the ONUs need to employ some arbitration mechanism to avoid data collisions and fairly share the fiber-channel capacity. This is achieved by allocation of a transmission window (timeslot) to each ONU. Each timeslot is capable of carrying several Ethernet packets. An ONU should buffer packets received from a subscriber until its timeslot arrives. When its timeslot arrives, the ONU should "burst" all stored packets at full channel speed. The timeslot size may be fixed (static) or variable (dynamic) depending, for example, on the amount of data stored in the ONU's queue.

1.A. Class-of-Service Considerations

Not being constrained by the limitations of copper outside plant, EPON is able to deliver tens to hundreds of megabytes per second to and from users. A giant step forward compared with cable modems and DSL technologies, EPON is expected to be a truly converged network, supporting voice communications, standard and high-definition video (STV and HDTV), video conferencing (interactive video), real-time and near-real-time transactions, and data traffic. To support this multitude of applications, EPON must exhibit an appropriate performance for each such application.

Performance of a packet-based network (and EPON in particular) can be conveniently characterized by several parameters: bandwidth, packet delay (latency), delay variation (jitter), and packet-loss ratio. Quality of service (QoS) refers to a network's ability to provide bounds on some or all these parameters on a per-connection (flow, session) basis. Not all networks, however, can maintain per-connection state or even identify connections. To support diverse application requirements, such networks separate all the traffic into a limited number of classes and provide differentiated service for each class. Such networks are said to maintain classes of service (CoS).

In this study we focus on CoS mechanisms in EPON. These mechanisms include *intra-ONU scheduling* and *inter-ONU scheduling* as shown in Fig. 2.

Being part of the Institute of Electrical and Electronics Engineers (IEEE) 802 family of standards, EPON must be compliant with bridging defined in IEEE 802.1D,[1] including compliance with CoS mechanisms in this standard. Specifically, IEEE 802.1D, clause 7.7.4, states that the default per-hop behavior (PHB) of bridges (intra-ONU scheduling, in our terminology) is a strict priority scheduling.

Currently, EPON architecture is being standardized by the IEEE 802.3ah task force. To allow efficient inter-ONU scheduling, EPON employs the Multipoint Control Proto-

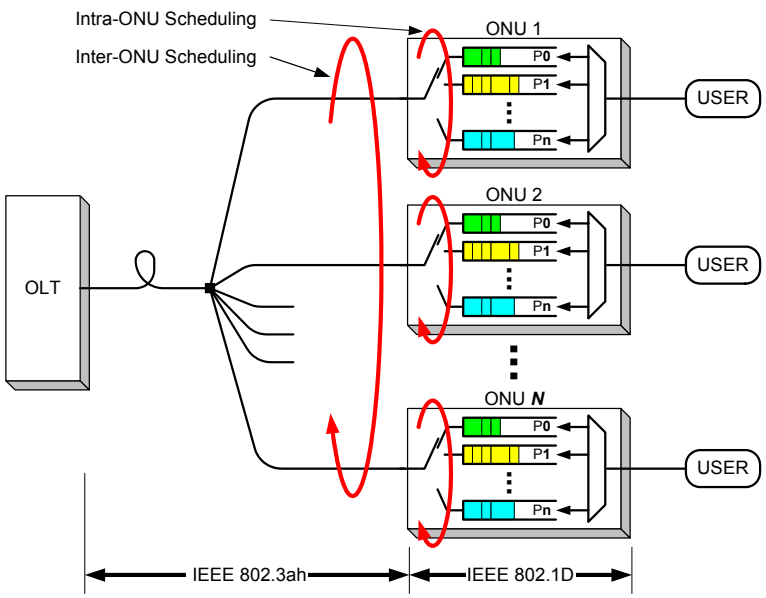

Fig. 2. Intra-ONU and inter-ONU scheduling.

col (MPCP). MPCP arbitrates (gates) transmission from multiple nodes to allow dynamic sharing of bandwidth while avoiding packet collisions. The standard does not specify a particular inter-ONU scheduling algorithm, allowing it to be vendor specific.

1.B. Previous Research and Contributions of This Study

In Ref. 2 we studied EPON performance with a static bandwidth-assignment algorithm when all traffic belonged to a single class. In Ref. 3 we introduced a dynamic bandwidth-allocation scheme (called IPACT) and studied various flavors of this protocol. We found that the *limited service* scheme has the best performance, and we evaluated the performance of different classes of traffic under this scheme. Our experiments consisted of observing the behavior of one ONU (tagged ONU) with a fixed load while the ambient network load was varied.

Our focus in the present study is to investigate how a gated transmission mechanism (MPCP) and a dynamic bandwidth-allocation scheme (limited service) can be combined with priority scheduling. In this study we vary the load of all ONUs. This experiment allows us to observe how a combination of the limited service scheme (inter-ONU scheduling) and a priority queuing (intra-ONU scheduling) results in quite an unexpected network behavior. We find that the queuing delay for some traffic classes increases when the network load decreases (a phenomenon we call *light-load penalty*). Since the light-load penalty affects only some traffic classes, it violates the fairness property among the traffic classes (i.e., performance for some classes degrades, whereas it improves for other classes as the load is increased).

We also suggest some optimization schemes that improve the performance and eliminate the light-load penalty (either partially or completely).

In Section 2 below, we give an overview of EPON's principle of operation and the MPCP. In Section 3 we briefly describe the CoS mechanisms of IEEE 802.1D. A particular EPON model and traffic characteristics are given in Section 4. In Section 5 we analyze EPON's performance for different CoS and identify the effects of the combination of gated

transmission and priority scheduling on packet delay. Section 6 concludes this study.

2. Multipoint Control Protocol

The work on EPON architecture in the IEEE 802.3ah task force is still in progress. Although many issues are still to be resolved, the final solution is beginning to emerge.

To support a timeslot allocation by the OLT, the MPCP is being developed by this task force. The MPCP is not concerned with a particular bandwidth-allocation (or inter-ONU scheduling) scheme; rather, it is a supporting mechanism that can facilitate implementation of various bandwidth-allocation algorithms in EPON.

This protocol relies on two Ethernet messages: GATE and REPORT. (Additionally MPCP defines REGISTER_REQUEST, REGISTER, and REGISTER_ACK messages used for an ONU's registration.) A GATE message is sent from the OLT to an ONU, and it used to assign a transmission timeslot. A REPORT message is used by an ONU to convey its local conditions (such as buffer occupancy, and the like) to the OLT to help the OLT make intelligent allocation decisions. Both GATE and REPORT messages are MAC (media access control) control frames (type 88-08) and are processed by the MAC control sublayer.

Below, we illustrate the operation of the MPCP.

1. From its higher layer (MAC control client), the MPCP in the OLT gets a request to transmit a GATE message to a particular ONU with the following information: time when that ONU should start transmission and length of the transmission (Fig. 3).

2. The MPCP layer (in the OLT and each ONU) maintains a clock. Upon passing a GATE message from its higher layer to MAC, the MPCP time stamps the message with its local time.

3. When an ONU receives a GATE message matching its MAC address (GATE messages are unicast), the ONU will program its local registers with transmission start and transmission length times. The ONU will also update its local clock to that of the timestamp in the received GATE message.

4. When the local time reaches the start-transmission register value, the ONU will start transmitting. That transmission may include multiple Ethernet frames. The ONU will ensure that no frames are fragmented. If the next frame does not fit in the remainder of the timeslot, it will be deferred till the next timeslot, leaving some unused remainder in the current timeslot.

REPORT messages are sent by ONUs in the assigned transmission windows together with data frames. REPORT messages can be sent automatically or on the OLT's demand. A REPORT message is generated in the MAC control client layer and is time stamped in the MAC control (Fig. 4). Typically, REPORT would contain the desired size of the next timeslot based on the ONU's queue size. When requesting a timeslot, an ONU should account for additional overhead, namely, 64-bit frame preamble and 96-bit interframe gap (IFG) associated with every Ethernet packet.

When a timestamped REPORT message arrives at the OLT, it is passed to the MAC control client layer responsible for making the bandwidth-allocation decision. Additionally, the OLT will recalculate the round-trip time (RTT) to the source ONU as shown in Fig. 5. Some small deviation of the new RTT from the previously measured RTT may be caused by changes in fiber refractive index resulting from temperature drift. A large deviation should alarm the OLT about the ONU's potential mis-synchronization and should prevent the OLT from further granting any transmissions to that ONU until it is reinitialized (resynchronized).

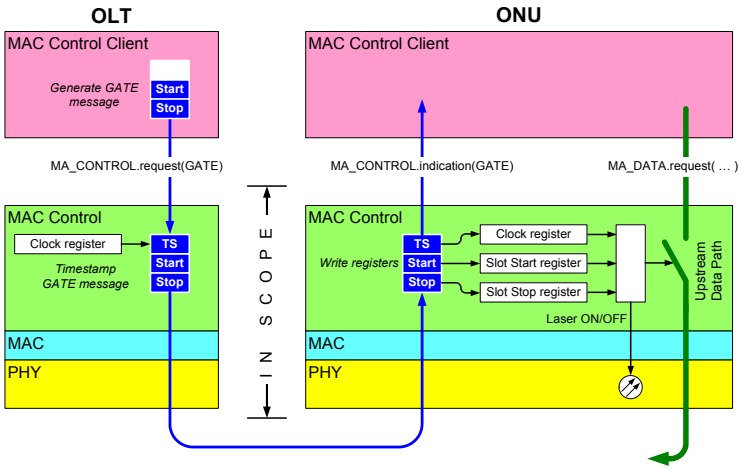

Fig. 3. MPCP–GATE operation.

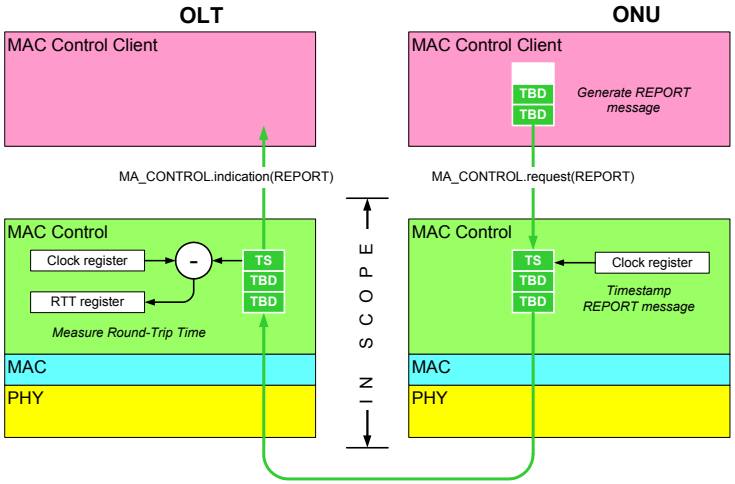

Fig. 4. MPCP–REPORT operation.

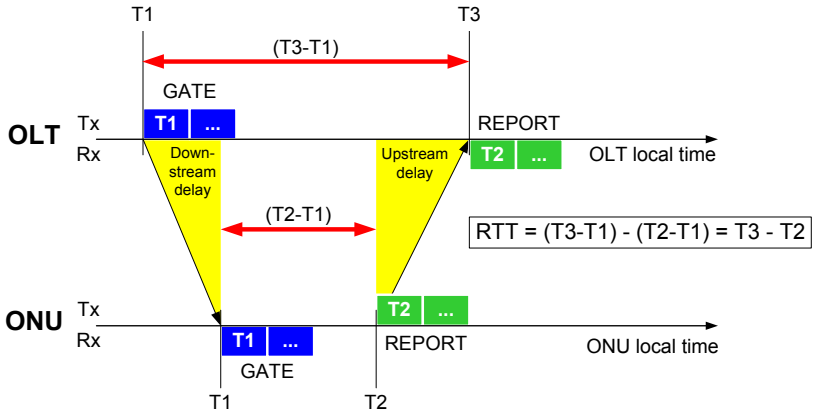

Fig. 5. Round-trip time measurement.

The above description represents a framework of the protocol being developed for the EPON. For more information on MPCP, consult Ref. 4.

3. Overview of IEEE 802.1D Support for Classes of Service

To support CoS, Ethernet networks must be able to classify traffic into CoS and provide differentiated treatment to each class. This was achieved by an introduction of two new standard extensions: P802.1p "Supplement to MAC bridges: traffic class expediting and dynamic multicast filtering" (later merged with P802.1D) and P802.1Q "Virtual bridged local area networks." P802.1Q defines a frame-format extension allowing Ethernet frames to carry a priority information field in their header. The standard distinguishes the following traffic classes:

1. *Network control.* Characterized by a "must get there" requirement to maintain and support the network infrastructure.
2. *Voice.* Characterized by less than 10-ms delay, and hence maximum jitter [one-way transmission through the local-area-network (LAN) infrastructure of a single campus].
3. *Video.* Characterized by less than 100-ms delay.
4. *Controlled load.* Important business applications subject to some form of "admission control," be that preplanning of the network requirement at one extreme to bandwidth reservation per flow at the time the flow is started at the other.
5. *Excellent effort.* Or "CEO's best effort," the best-effort-type services that an information services organization would deliver to its most important customers.
6. *Best effort.* LAN traffic as we know it today.
7. *Background.* Bulk transfers and other activities that are permitted on the network but that should not affect the use of the network by other users and applications.

If a bridge or a switch has less than seven queues, some of the traffic classes are grouped together. Table 1 illustrates the standard-recommended grouping of traffic classes.

Table 1. Mapping of Traffic Classes into Priority Queues (P802.1p)

Number of queues	Traffic types queue assignments						
1	Network Control	Voice	Video	Controlled Load	Excellent Effort	Best Effort	Background
2	Network Control	Voice	Video	Controlled Load	Excellent Effort	Best Effort	Background
3	Network Control	Voice	Video	Controlled Load	Excellent Effort	Best Effort	Background
4	Network Control	Voice	Video	Controlled Load	Excellent Effort	Best Effort	Background
5	Network Control	Voice	Video	Controlled Load	Excellent Effort	Best Effort	Background
6	Network Control	Voice	Video	Controlled Load	Excellent Effort	Best Effort	Background
7	Network Control	Voice	Video	Controlled Load	Excellent Effort	Best Effort	Background

P802.1D in clause 7.7.4 specifies the default bridge (switch) scheduling algorithm for multiple queues:

> **7.7.4 Selecting frames for transmission**
>
> The following algorithm shall be supported by all bridges as the default algorithm for selecting frames for transmission:
>
> (a) For each port, frames are selected for transmission on the basis of the traffic classes that the port supports. For a given supported value of traffic class, frames are selected from the corresponding queue for transmission only if all queues corresponding to numerically higher values of traffic class supported by the port are empty at the time of selection.
>
> (b) For a given queue, the order in which frames are selected for transmission shall maintain the ordering requirement specified in 7.7.3.
>
> Additional algorithms, selectable by management means, may be supported as an implementation option, so long as the requirements of 7.7.3 are met.

4. System Architecture: Integrating Priority Queuing in Ethernet Passive Optical Networks

In this study we consider an EPON access network consisting of an OLT and N ONUs (Fig. 6). The transmission speed of the EPON and the user access link may not necessarily be the same. In our model we consider R_U Mbit/s to be the data rate of the access link from a user to an ONU and R_N Mbit/s to be the rate of the upstream link from an ONU to the OLT (see Fig. 6). Line rates for each link are the same in upstream and downstream directions.

Every ONU is located at a certain distance from the OLT and has a certain propagation delay. A downstream propagation delay (from the OLT to the ONU) and an upstream propagation delay (from the ONU to the OLT) for each ONU are the same if a single fiber is used for bidirectional transmission; otherwise, the delays may be different. We denote L to be the largest distance between the OLT and an ONU.

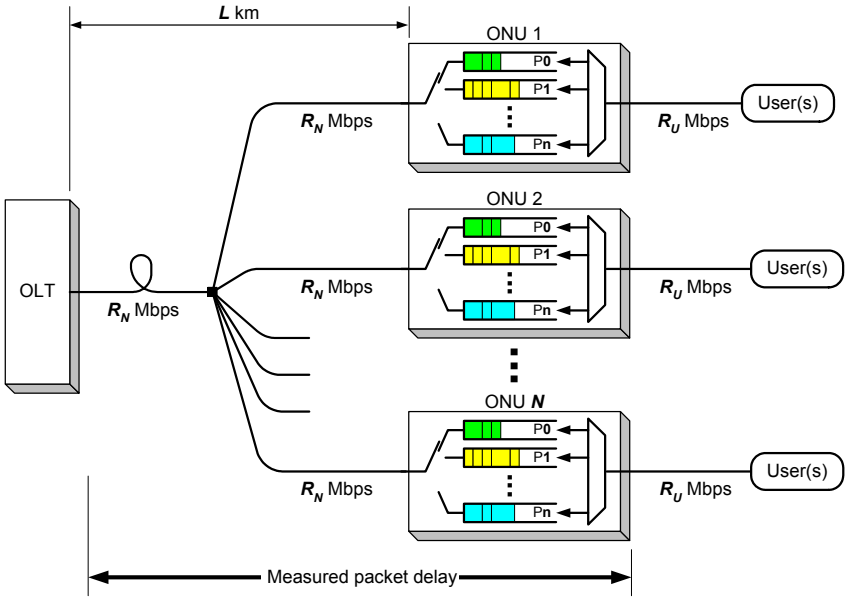

Fig. 6. Access network based on PON.

Each ONU is equipped with n queues serving n priority classes (denoted $P0, P1, \ldots, Pn$), with $P0$ being the highest priority and Pn being the lowest. When a packet is received from a user, the ONU classifies its type and places it in the corresponding queue. The queues in each ONU share common memory of size Q bytes. If an arriving packet with priority Pi finds the buffer full in the ONU, it can preempt one or more lower-priority packets Pk ($k > i$) from their queues, such that the Pi packet can itself be placed into the Pi queue. Between slots, an ONU stores all the packets received from the user in their respective queues. When a slot arrives, the ONU serves a higher-priority queue to exhaustion before serving a lower-priority queue.

In our model we used the *limited service* discipline described in Ref. 3. Under this discipline, the OLT assigns to an ONU a slot of size equal to what the ONU had requested (through a previous REPORT message), but not greater than some predefined maximum W_{MAX}. The limit W_{MAX} is needed to guarantee maximum interval between slots (cycle time) T_{MAX} and to avoid bandwidth hogging by a "hungry" ONU. This scheme was shown to share the bandwidth efficiently while maintaining fairness among ONUs.[3]

In our study we use simulation experiments rather than analytical methods. Our objective was to build a realistic model and evaluate the system behavior with a specific set of parameters; analytical modeling for such a system becomes extremely complex.

There is an extensive study showing that most network traffic flows [i.e., those generated by http, ftp, variable-bit-rate (VBR) video applications, and the like] can be characterized by self-similarity and long-range dependence (LRD) (see Ref. 5 for an extensive reference list). To obtain an accurate and realistic performance analysis, we generated synthetic self-similar traffic, using the method described in Ref. 6. We used a trimodal packet-size distribution similar to that observed in backbone networks[7] and in a cable network headend.[8] The three main modes correspond to most-frequent packet sizes: 64, 582/594, and 1518 Bytes (including Ethernet headers). The weights of the modes slightly differ in the backbone measurements and in the access network, so we used the distribution for the upstream traffic reported in Ref. 8.

An important characteristic of a self-similar process is its heavy-tailed behavior (with tail-decay coefficient $\alpha, 1 < \alpha < 2$). This leads to a property of infinite variance (i.e., all moments $m > \alpha$ do not exist). Therefore most analytical methods model only approximate behavior. Having built a simulation model, we could concentrate on the transient system behavior and evaluate the behavior when the offered load exceeds the network capacity.

Table 2 summarizes the default system parameters used in our simulation experiments.

Table 2. Default System Parameters

Parameter	Description	Value
N	Number of ONUs	16
n	Number of priority classes (also number of queues in an ONU)	3
R_U	Line rate of user-to-ONU link	100 Mbit/s
R_N	EPON line rate	1000 Mbit/s
Q	Buffer size in ONU	8 Mbit
L	Maximum distance between OLT and an ONU	20 km
W_{MAX}	Maximum slot size	15,000 Bytes
B	Guard time between adjacent slots (At the time of writing, the IEEE 802.3ah task force was investigating a possibility of reducing the guard time to 1 μs. In our simulation experiments we used a conservative value of 5 μs)	5 μs
T_{MAX}	Maximum cycle time $T_{\mathrm{MAX}} = N\left(B + \frac{W_{\mathrm{MAX}}}{R_N}\right)$	2 ms

Traffic Modeling

As a multiservice access network, the proposed architecture should support a multitude of services, i.e., best-effort data, VBR video stream, constant-bit-rate (CBR) stream [for legacy equipment such as plain old telephone service (POTS) lines, private branch exchange (PBX) boxes], and so on.

In our simulation experiments we divide our data into three priority classes: $P0$, $P1$, and $P2$. The three classes may be used for delivering voice, video, and data traffic. Using three classes also allows easy mapping of DiffServ's Expedited Forwarding (EF), Assured Forwarding (AF), and Best Effort (BE) classes into 802.1D classes.

Class P0 is used to emulate a circuit-over-packet connection. $P0$ traffic has CBR. In our model we chose to emulate a T1 connection. The T1 data arriving from the user are packetized in the ONU by means of placing 24 Bytes of data in a packet. Including Ethernet and UDP/IP (User Data Protocol/Internet Protocol) headers results in a 70-Byte frame generated every 125 μs. Hence the $P0$ data consume 4.48 Mbit/s of bandwidth.

Class P1 in our experiment consists of VBR video streams that exhibit properties of self-similarity and LRD (as was shown in Ref. 9 for real MPEG-coded video streams). Since the $P1$ traffic is highly bursty, it is possible that some packets in long bursts will be lost. This will happen if the entire buffer is occupied by $P1$ and $P0$ packets. Packet sizes in $P1$ streams range from 64 to 1518 Bytes.

Class P2 has the lowest priority. This priority level is used for non-real-time data transfer. There are no delivery or delay guarantees for this service. This is also self-similar and long-range-dependant traffic with variable-size packets ranging from 64 to 1518 Bytes.

When we vary the load in our simulation experiments, we always keep the $P0$ load constant (4.48 Mbit/s) and split the remaining load between $P1$ and $P2$ equally. For example, an ONU's offered load of 40 Mbit/s means that $P1$ and $P2$ classes generated $(40-4.48)/2 = 17.76$ Mbit/s each. In all the performance diagrams, the ONU's offered load values are normalized to the ONU's ingress link capacity (R_U = 100 Mbit/s).

5. Packet Delay Analysis

We start our performance analysis by investigating the limited service discipline. This discipline grants the requested number of bytes but no more than W_{MAX}. As our performance measures, we consider average and maximum packet delay (Fig. 7). Each point on the plots corresponds to a sample of 500 million packets. The horizontal axis represents the load of an individual ONU. The network load Φ can be derived from the ONU's load ϕ as

$$\Phi = \frac{R_U}{R_N} \sum_{i=1}^{N} \phi_i. \tag{1}$$

In our simulation experiments, all ONUs have a uniform load. Thus 100% of network load corresponds to ONU's load of 0.625. Understandably, the delay plots show clear knees at a load of $\sim$0.625. At this point, the network begins to exhibit signs of saturation: Buffers are full, and a large amount of packets is dropped.

One can immediately observe that combining default priority-queuing PHB with a simple polling mechanism in an EPON results in a very interesting phenomenon: As the load decreases from moderate ($\sim$0.25) to very light ($\sim$0.05), the average delay for the lowest-priority class ($P2$) increases significantly. The average packet delay at a load of 0.05 (or 5 Mbit/s) is 17.8 ms, more than 1200% higher than the 1.4-ms delay at a load of 0.25 (25 Mbit/s). Similar behavior is observed for the maximum packet delay for $P1$ and $P2$ classes. We refer to this phenomenon as *light-load penalty*.

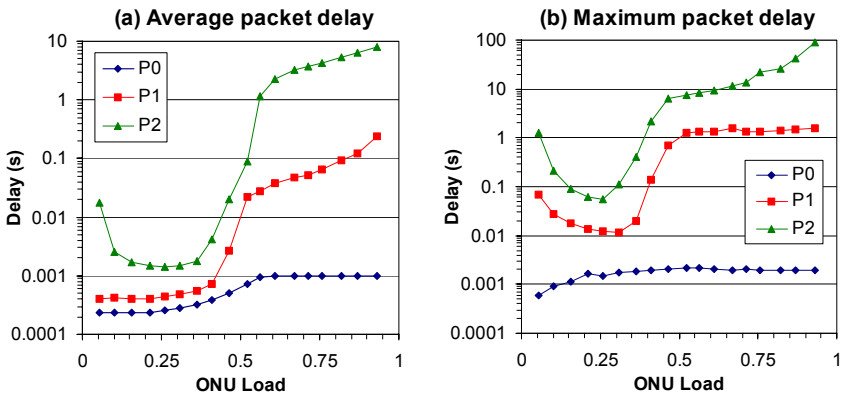

Fig. 7. Packet delay for limited service.

Light-Load Penalty

A trace-level analysis of the polling scheme combined with priority queuing reveals the cause of light-load penalty. At the end of every timeslot, an ONU generates a REPORT message containing the number of bytes that remain in the queue (*residual queue occupancy*). The residual queue occupancy is almost always less than the maximum slot size W_{MAX}, because the light-load penalty occurs only at light loads. This means that whatever slot size an ONU requested in the REPORT message, the OLT will grant the requested slot size through the next GATE message to that ONU. However, during the time lag between ONU's sending a REPORT and the arrival of its assigned timeslot (i.e., between sending a REPORT and transmission of the reported data), more packets arrive to the queue. Newly arrived packets may have higher priority than some packets already stored in the queue, and they will be transmitted in the next transmission slot before the lower-priority packets. Since these new packets were not reported to the OLT, the given slot cannot accommodate all the stored packets. This causes some lower-priority packets to be left in the queue. This situation may repeat many times, causing some lower-priority packets to be delayed for multiple cycle times. A lower-priority packet will finally be transmitted when more lower-priority packets (bytes) accumulate (and are reported to the OLT) behind a given packet than higher-priority packets cut in front of it. But, since $P0$ traffic is periodic (CBR) and $P2$ traffic is bursty (i.e., a new $P2$ packet may not arrive for a long time), on average, at a load of 0.05, $P2$ packets are delayed by $\sim$80 cycles. As the load increases, the queue behind a lower-priority packet grows faster and the light-load penalty decreases. At load 0.25, the average delay for $P2$ packets is only $\sim$4.6 cycles.

Since Ethernet packets cannot be fragmented (according to IEEE 802.3), packet preemption results in an *unused slot remainder* (unless an added higher-priority packet is the same size as a preempted lower-priority packet, which is rare). We investigate the properties of such remainders in Subsection 6.B.

It should be clear that the light-load penalty can never happen with the first-come–first-served (FCFS) queuing discipline (with no priorities); later-arriving packets are appended to the end of the queue, and they cannot displace earlier packets from their place in the queue (i.e., arriving packets do not change the delineation of packets already stored in the queue and reported to the OLT).

Some higher-layer protocols [i.e., Transmission Control Protocol (TCP) Vegas[10], or cprobe and bprobe[11]] rely on the packet delay as a measure of network congestion. The light-load penalty may have a detrimental effect on the operation of such protocols: A

random fluctuation that reduces the load could increase the delay, which in turn could be interpreted by a data source as increased congestion and will force it to reduce its load (sending rate), thus increasing the delay even more. This chain reaction can lead to unstable behavior of the higher-layer protocol or may prevent its proper operation altogether.

Another reason EPON designers should eliminate or mitigate the light-load penalty is that it may encourage low-priority applications to artificially generate heavier-than-required load in order to get better performance from the network. Although this may improve the performance for the lower-priority class, higher-priority classes will be adversely affected.

6. Optimization Schemes

6.A. Two-Stage Buffers

One way to eliminate the light-load penalty is to implement a two-stage queue in an ONU (Fig. 8). In a two-stage system, stage I consists of multiple-priority queues and stage II consists of one FCFS queue. When a timeslot arrives, data packets from stage II are transmitted to the OLT, thereby vacating the queue; simultaneously, data packets from stage I are advanced into vacant spaces in the stage-II queue. At the end of the current timeslot, the ONU reports to the OLT the occupancy of the stage-II queue (to get a corresponding slot size in the next cycle). The total size of the stage-II buffer can be made exactly W_{MAX} Bytes so that the ONU never requests a slot greater than W_{MAX} Bytes. This configuration will ensure that the given slot is always 100% utilized, i.e., that the unused remainder is always zero.

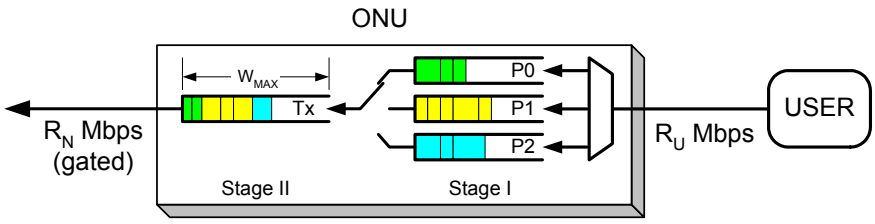

Fig. 8. Two-stage queue at an ONU.

Figure 9 presents the average and maximum packet delay in a two-stage queuing system. Immediately, we can see that the light-load penalty has been eliminated. The drawback of this scheme, however, is the increased delay for the highest-priority class ($P0$). The average delay for $P0$ has increased three times (under heavy-load conditions, the average delay in the two-stage scheme settles at 3 ms, whereas in the limited service, it was 1 ms). This can be intuitively explained by the fact that a packet that arrived at a random time will wait, on average, half a cycle in the stage-I queue and exactly one cycle in the stage-II queue. If just one priority queue is implemented (without stage II), the average $P0$ delay is only half a cycle time. The maximum delay for $P0$ has increased two times. The reason for this is also clear: A high-priority packet can spend at most one cycle time in each stage—two cycle times total. In the limited service scheme, the maximum delay for $P0$ packets was limited to one cycle time (refer to Fig. 7).

The increased delay in a two-stage queuing scheme could sometimes become a problem. All high-priority packets such as system alarms, failure indication, and the like may have to endure a longer delay. For example, consider the delay budget for voice traffic. International Telecommunication Union Telecommunication Standardization Sector (ITU-T) Recommendation G.114 “One-way transmission time” specifies 1.5-ms one-way propaga-

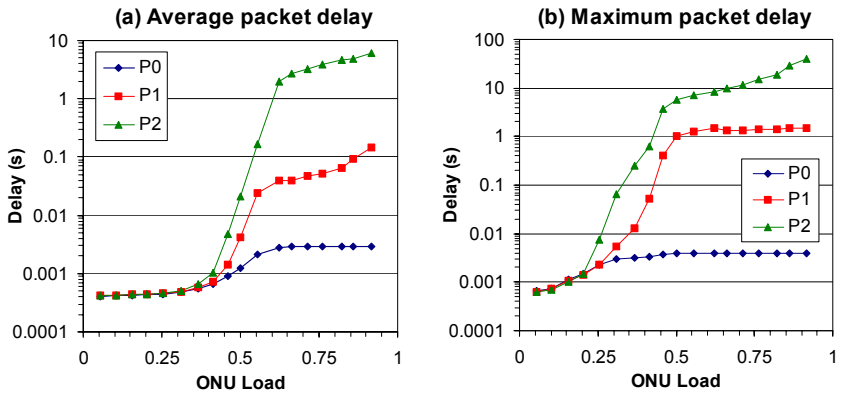

Fig. 9. Packet delay for two-stage queuing scheme.

tion delay in an access network (digital local exchange). To keep the average delay within this bound, the maximum cycle time T_{MAX} will have to be reduced to 1 ms. With our default configuration parameters (see Table 2), this increases the guard-time overhead under heavy-load conditions from 4% to 8% (overhead is NB/T_{MAX}). However, we note that, whereas the guard-time overhead increases, the total overhead increases not as much; this is because the guard-time overhead is partially compensated for by a complete elimination of the unused slot remainder (slot-packing overhead). In Ref. 2 we derived the following formula for the expected size of the unused slot remainder (denoted by R):

$$E(R) = \frac{1}{E(X)} \sum_{r=1}^{M-1} r\,[1 - F_X(r)], \tag{2}$$

where M is maximum Ethernet packet size, X is a random variable representing packet size (X is assumed to be independent identically distributed), and $F_X(x)$ is a cumulative distribution function.

Using the above formula, we can estimate the total overhead. For the packet size distribution from Ref. 8 and the limited service discipline with T_{MAX}= 2 ms, we have

$$\text{overhead} = \frac{N[B + E[R] \times 8\ (\text{ns/Byte})]}{T_{\mathrm{MAX}}} = \frac{16(5\ \mu\text{s} + 555\ \text{Byte} \times 8\ \text{ns/Byte})}{2\ \text{ms}} \approx 7.55\%.$$

For the two-stage buffer scheme the unused slot remainder is always zero, and hence for T_{MAX} = 1 ms we get

$$\text{overhead} = \frac{NB}{T_{\mathrm{MAX}}} = \frac{16 \times 5\ \mu\text{s}}{1\ \text{ms}} \approx 8\%.$$

It is clear that, in the two-stage system, the overhead can be reduced compared with the limited service scheme if the guard time B can be made smaller than the estimated unused slot remainder $E[R]$.

6.B. Constant-Bit-Rate Credit

Another interesting solution to the light-load penalty (without increasing the delay of the highest-priority class beyond one cycle time as in a two-stage scheme) is to predict the amount of high-priority packets that are expected to arrive at the ONU and to adjust the granted timeslot size accordingly. Of course, to predict the traffic with any reasonable accuracy, we need to have some knowledge about the traffic behavior. In our case, we have

this knowledge about the *P*0 traffic; namely, we know that this is a CBR flow with a given data rate. Therefore, when deciding on the size of the next timeslot for an ONU, the OLT can estimate the time of the next transmission and increase the timeslot size by the amount of CBR data it anticipates. We call this scheme *CBR credit*, since the additional timeslot size increment (credit) is based on the known CBR arrival rate.

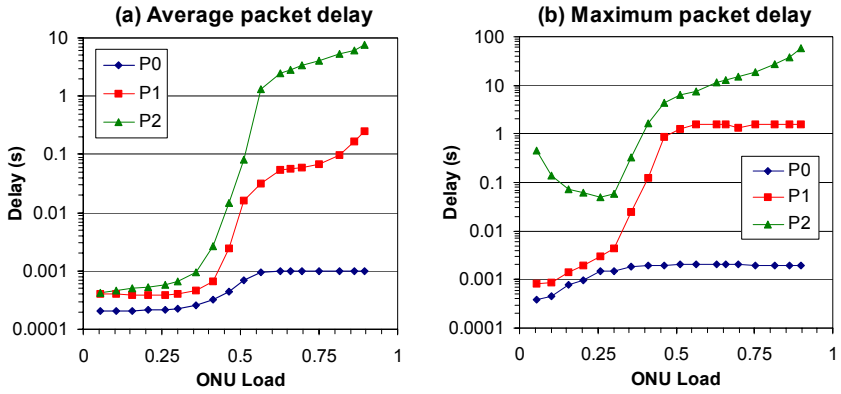

Fig. 10. Packet delay for CBR credit scheme.

Figure 10 shows the average and the maximum packet delay for the CBR credit scheme. We can see that the light-load penalty is eliminated for average delay values. The penalty remains (in somewhat lesser degree) for the maximum delay values for the *P*2 class only. This behavior is expected, since the penalty for class *P*2 is caused not only by *P*0 traffic (CBR) but by *P*1 traffic as well. The CBR credit scheme makes no attempt to predict the arrivals of *P*1 traffic (which is a highly bursty, LRD traffic). Luckily, the probability of a *P*1 packet displacing a *P*2 packet at light load is not high, and the light-load penalty does not affect the average delay of *P*2 packets.

To compute the size of the credit, the OLT first determines the credit interval τ shown in Fig. 11. The size of the next timeslot should be increased (credited) to accommodate CBR packets that are to arrive during the credit interval. The credit interval can be calculated as follows. Given that t_R is timestamp in a received REPORT message, t_S is start time of a granted slot, s is slot size, and R_N is EPON line rate, we have

$$\tau = t_S + s/R_N - t_R. \tag{3}$$

Taking T_{CBR} to be the period of CBR packet arrivals (in s/packet), we can expect n_{CBR} CBR packets to arrive during the interval τ; i.e.,

$$n_{\mathrm{CBR}} = \tau / T_{\mathrm{CBR}} = (t_S + s/R_N - t_R)/T_{\mathrm{CBR}}. \tag{4}$$

But the slot size s itself depends on the number of additional CBR packets it should accommodate. Thus,

$$n_{\mathrm{CBR}} = \frac{1}{T_{\mathrm{CBR}}}\left(t_S + \frac{v + n_{\mathrm{CBR}} S_{\mathrm{CBR}}}{R_N} - t_R\right) = \frac{t_S + v/R_N - t_R}{T_{\mathrm{CBR}} - S_{\mathrm{CBR}}/R_N}, \tag{5}$$

where S_{CBR} is the size of CBR packets (70 Bytes in our experiment; see Section 4, Traffic Modeling) and v is a number of bytes (requested slot size) reported by the ONU. Thus the OLT assigns the slot size on the basis of the following formula:

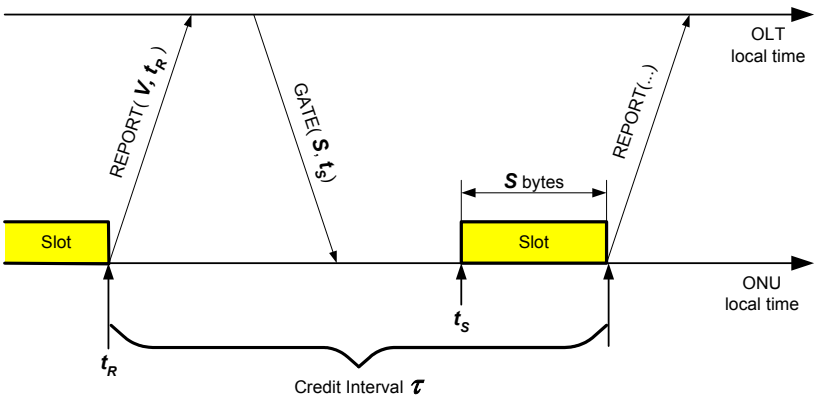

Fig. 11. Calculation of the credit interval.

$$w_i = \min\begin{cases} v_i + \lceil n_{\mathrm{CBR},i} \rceil \, (S_{\mathrm{CBR}} + \mathrm{IFG}) \\ W_{\mathrm{MAX}} \end{cases}$$
$$= \min\begin{cases} v_i + \left\lceil \frac{t_{S,i} + v_i/R_N - t_{R,i}}{T_{\mathrm{CBR}} - S_{\mathrm{CBR}}/R_N} \right\rceil (S_{\mathrm{CBR}} + \mathrm{IFG}) \\ W_{\mathrm{MAX}} \end{cases}, \tag{6}$$

where w_i is the slot size assigned to ONU i, v_i is the requested slot size from ONU i, $t_{R,i}$ is the timestamp of the REPORT message received from ONU i, $t_{S,i}$ is the start time of the slot assigned to ONU i, IFG is the minimum interframe gap (includes 64-bit preamble as well as 96 bits of IFG), and W_{MAX} is the maximum limit on slot size.

The $t_{S,i}$ value is updated after each slot assignment as

$$t_{S,i+1} = t_{S,i} + (w_i/R_N) + B; \tag{7}$$

i.e., the OLT expects the data (first bit) from ONU $i+1$ to arrive exactly after the guard time B after the data (final bit) from ONU i. Recall that B is the guard time (see Table 2).

The value $\lceil n_{\mathrm{CBR}} \rceil \, (S_{\mathrm{CBR}} + \mathrm{IFG})$ in Eq. (6) represents the CBR credit given to ONU i. The ceiling function is used to accommodate an integer number of packets. Obviously, in some instances, our prediction may be too generous, and the ceiling function will give an ONU more credit than it actually needs (i.e., only $\lfloor n_{\mathrm{CBR}} \rfloor$ packets may arrive at the ONU in interval τ). Below, we show that it is more efficient to give extra credit than not enough. If the OLT credited $\lceil n_{\mathrm{CBR}} \rceil \, (S_{\mathrm{CBR}} + \mathrm{IFG})$ Bytes, but only $\lfloor n_{\mathrm{CBR}} \rfloor$ CBR packets arrived, the slot will have an unused remainder of size $S_{\mathrm{CBR}} + \mathrm{IFG}$ Bytes exactly (i.e., 70 + 20 = 90 Bytes in our case). If, however, in an alternative case, the OLT conservatively credited $\lfloor n_{\mathrm{CBR}} \rfloor \, (S_{\mathrm{CBR}} + \mathrm{IFG})$ Bytes and $\lceil n_{\mathrm{CBR}} \rceil$ CBR packets arrived, they will all be sent ahead of other lower-priority packets, displacing one or more lower-priority packets from the slot. If this happens, the worst-case unused remainder is one less than the largest packet size (with associated IFG and preamble); i.e., $1518 + 20 - 1 = 1537$ Bytes of wasted slot space.

To illustrate the advantages of the CBR scheme, we built a distribution (histogram) for the unused slot remainder and compared it with this distribution for the limited service (Fig. 12). We obtained these plots by simulating the transmission of 10^9 packets through the EPON at an average ONU load of 0.05 (5 Mbit/s).

We measured and found the average unused slot remainder at a load of 0.05 to be 540 Bytes. This is the average slot space that will remain unused with every credit mispredic-

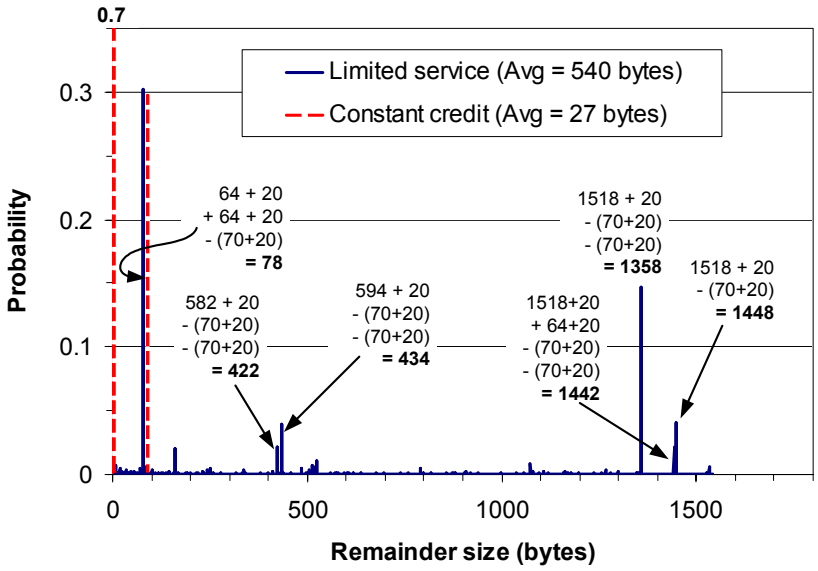

Fig. 12. Unused slot remainder distribution.

tion. Obviously, it is much cheaper (in terms of use) for the OLT to credit more than to credit less.

In the CBR credit scheme the unused remainder is reduced to only 27 Bytes. We also can observe an interesting periodicity in the limited service remainder distribution. It can be explained by the pronounced modality of the underlying packet-size distribution. The trimodal packet-size distribution has been demonstrated in backbone[7] and in access networks.[8] The three modes correspond to the most frequent packet sizes: 64 Bytes (46%), 582/594 Bytes (10%), and 1518 Bytes (12%). [Packet sizes include Ethernet header (18 Bytes). Also, all payload sizes less than 46 Bytes are padded to 46 Bytes to comply with the 64-Byte minimum IEEE 802.3 frame size.] The interaction of these three modes among themselves and with CBR traffic results in the periodic pattern of the remainder distribution. The plot in Fig. 12 illustrates which combinations of packets result in particular peaks of the remainder distribution. For example, the remainder of 1448 Bytes is left when one CBR packet (70 Bytes) displaces one 1518-Byte packet. A remainder of 1358 Bytes is left when two CBR packets displace one 1518-Byte packet, and so on.

The plot for CBR traffic shows two major distributional peaks: one at 0 Bytes (70%) and the other at 90 Bytes (29.8%). The peak at zero represents all slots where the OLT exactly predicted the number of CBR packet arrivals. The peak at 90 Bytes represents cases in which the ceiling function overestimated the CBR arrivals and granted the slot size for one extra CBR packet (i.e., when $\lfloor n_{\mathrm{CBR}} \rfloor$ packets arrived instead of $\lceil n_{\mathrm{CBR}} \rceil$ packets).

A limitation of the CBR credit scheme is that external knowledge of the arrival process is necessary. Even though, for some time-critical applications, we may have such knowledge, it is by no means a universal case. This scheme can be applied for circuit-emulation services with a CBR arrival process. A fairly straightforward modification of this scheme would allow it to be used with voice-over-packet traffic, even in silence-suppression (SS) mode (i.e., when no packets are transmitted during silence intervals). Now, the OLT would start crediting an ONU when a talk spurt is detected and will stop crediting when silence is detected (the OLT can detect talk spurts and silence by presence or absence of voice packets). The packet rate within a talk spurt is constant, and so the crediting scheme would work.

The fact that the OLT will mispredict the beginning and the end of a talk spurt should not introduce any significant overhead, since the average 3-s misprediction window [assuming that average talk spurt is 1.65 s and that average silence interval is 1.35 s (Ref. 12)] is much larger than the cycle time (2 ms maximum). Analysis of the prediction accuracy and efficiency of the CBR credit scheme with the *P*0 class consisting of voice-over-packet traffic remains a topic of future research.

7. Bandwidth Utilization

Bandwidth utilization in EPON is determined by the cycle time, the guard time, and the size of unused slot remainder. Figure 13 presents the absolute and the normalized values of the average slot remainder.

At higher loads (0.6 and above), the remainder values for the limited service scheme and for the CBR credit scheme are the same, and they correspond to the value obtained from Eq. (2). However, Eq. (2), which assumes slot size independent of packet size and large enough to fit many packets, cannot be applied at light load. At light load this independence assumption breaks down, since ONUs request small slots, just enough to fit a few packets. Thus the slot-size distribution has a strong correlation with packet-size distribution. The hump in the limited service plot at a load of 0.2 is another manifestation of the light-load penalty. A large lower-priority packet, which is continuously being preempted by small higher-priority packets, will result in a larger remainder for many cycles.

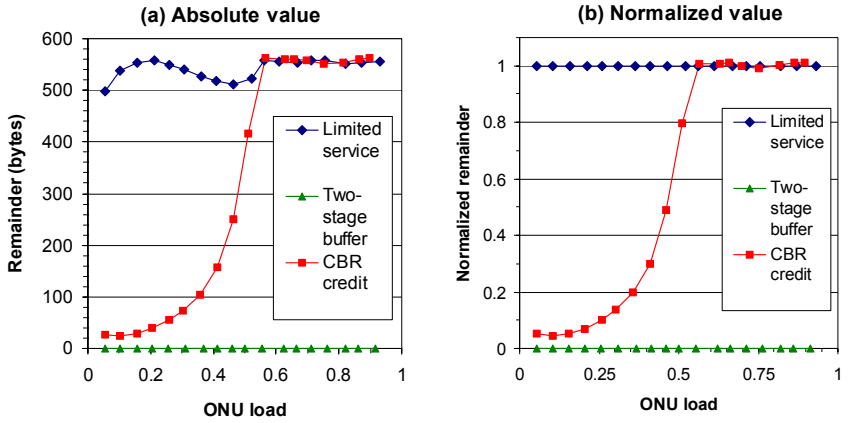

Fig. 13. Average slot remainder.

Figure 13(b) shows that, starting at a load of 0.6, the average unused remainders in limited service and CBR credit schemes are the same [and have an absolute value of approximately 555 Bytes as predicted by Eq. (2)]. This is expected, since, at high loads, the ONUs request windows larger than W_{MAX} Bytes. According to Eq. (6), the OLT will grant them a W_{MAX}-Byte slot, ignoring the credit value. Therefore the CBR credit scheme's performance at high loads is the same as in the limited service scheme.

It is interesting to observe that, in the CBR credit scheme, even though a larger slot size is granted to the ONUs, the cycle time is reduced. Figure 14 shows the average cycle time for all three schemes described above. At a load of 0.35, the average cycle time for the CBR credit scheme is 301 μs, a 30% reduction from 422 μs in the limited service scheme. This advantage is gained through the reduction in unused slot remainder, which results in increased network utilization. The two-stage buffer system is found to have even larger cycle-time reduction.

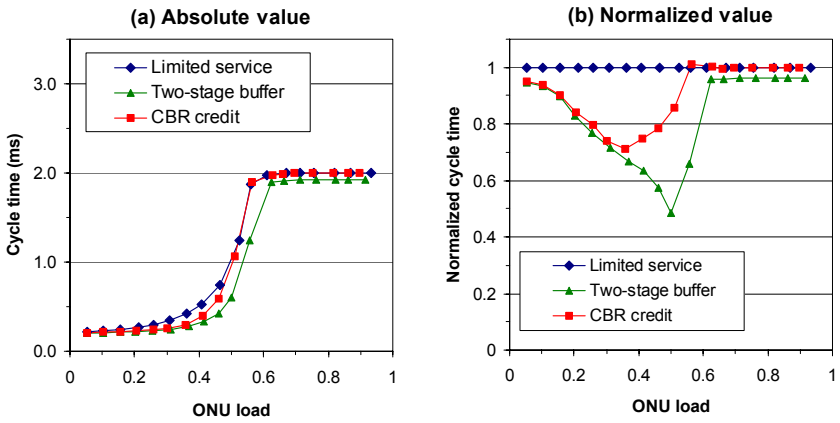

Fig. 14. Average cycle time.

Both the cycle time and the slot remainder affect the bandwidth utilization in EPON. The following formula allows us to compute bandwidth utilization U:

$$U = 1 - \frac{N(B + R/R_N)}{t_C}, \tag{8}$$

where N is number of ONUs, B is guard time, R is remainder, and t_C is cycle time. Based on Eq. (8), Fig. 15 presents the absolute and the normalized bandwidth-utilization values. We can see that both two-stage buffer scheme and the CBR credit scheme result in considerable improvement in bandwidth utilization at light loads. The two-stage buffer scheme also shows slightly better utilization at high loads, because of its complete elimination of the unused remainder.

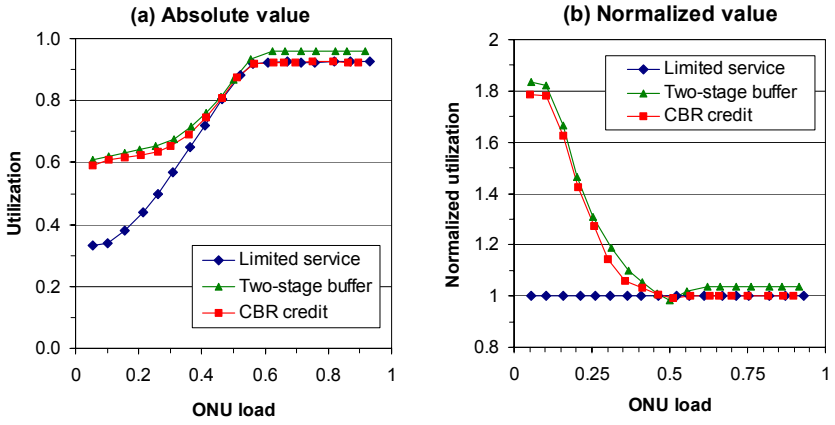

Fig. 15. Bandwidth utilization.

8. Conclusion

For successfully integrating Ethernet PONs (EPONS) into an access-network environment, a host of technical issues need to be solved. In this paper we have focused on one important issue—support for differentiated classes of service. We limited our investigation to packet delay and bandwidth-utilization characteristics.

We found that a combination of a default scheduling algorithm (priority scheduling) and MPCP (limited service) results in an interesting phenomenon in which some classes of traffic are treated unfairly when the network load is light. In fact, under light loads, ONUs with FCFS queue perform better than ONUs with priority scheduling. We call this phenomenon light-load penalty.

To alleviate this penalty, we proposed and examined the characteristics of two optimization schemes with different trade-offs. A two-stage queuing scheme eliminates the light-load penalty completely at the expense of increased packet delay for all the classes of traffic.

Another scheme (CBR credit) attempts to predict high-priority packet arrivals. This scheme eliminates light-load penalty for most packets (even though some low-priority packets are delayed excessively in a buffer, the number of such packets is small and does not affect the average packet delay). The limitation of this scheme is that some external knowledge of the traffic-arrival process is needed.

Even though we focused only on packet-delay optimization and bandwidth utilization, they are by no means the only parameters affecting the network's performance. Other measures, such as guaranteed and best-effort bandwidth, jitter, and packet loss are also important. We refer the reader to Ref. 3 for a description of various ONU-scheduling disciplines in EPON. The packet-loss characteristic, although outside the scope of this paper, is nevertheless an important performance parameter. The default packet discarding policy (drop-tail) shows that it is unfair to large packets; i.e., under heavy load, larger packets are more likely to be dropped (we call this the large-packet penalty). This behavior is understandable and even expected [this is one of the reasons for implementing the random-early-discard (RED) policy]. What is interesting in EPON settings is that the large-packet penalty affects higher-priority $P1$ traffic more than it affects lower-priority $P2$ traffic.

Being part of IEEE 802 family of standards, EPON must be compliant with bridging defined in IEEE 802.1D, including compliance with CoS mechanisms in this standard. In this study we focused only on strict priority scheduling, because of its status as a default scheduling algorithm in IEEE 802.1D-compliant bridges and switches. Priority queuing is easy to implement. It provides low delay to high-priority traffic, but it may have some performance shortcomings such as better-than-needed performance for high-priority queues and starvation of low-priority queues.

A large amount of research has been done in developing scheduling algorithms with improved fairness in resource sharing (a family of *fair queuing* protocols based on a concept of generalized processor sharing[13]). Integration of such schedulers in EPON is not a trivial task, because of EPON's distributed nature and unique properties such as limited control-plane bandwidth, large propagation delay, and significant switching overhead. In a forthcoming study we plan to investigate the feasibility of implementing a fair-queuing scheduling in EPON.

Acknowledgments

G. Kramer and B. Mukherjee were supported by a research gift from Nokia.

The authors thank Bob Gaglianello of Lucent Technologies; Gerry Pesavento, J. C. Kuo, and Ed Boyd of Teknovus; and Ariel Maislos and Onn Haran of Passave Networks for their insightful comments and suggestions. We are grateful to Dolors Sala of Broadcom Corporation for providing head-end traffic statistics.

References and Links

1. ANSI/IEEE Standard 802.1D, 1998 ed., "IEEE standard for information technology—Telecommunications and information exchange between systems—Local and metropolitan area networks—Common specifications. Part 3: media access con-

trol (MAC) bridges (Institute of Electrical and Electronics Engineers, 1998), http://standards.ieee.org/getieee802/download/802.1D-1998.pdf.
2. G. Kramer, B. Mukherjee, and G. Pesavento, "Ethernet PON (ePON): design and analysis of an optical access network," Photon. Netw. Commun. **3**(3), 307–319 (2001).
3. G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT: a dynamic protocol for an Ethernet PON (EPON)," IEEE Commun. **40**(2), 74–80 (2002).
4. G. Kramer, B. Mukherjee, and A. Maislos, "Ethernet passive optical networks," in *Multiprotocol over DWDM: Building the Next Generation Optical Internets*, S. Dixit, ed. (to be published).
5. W. Willinger, M. S. Taqqu, and A. Erramilli, "A bibliographical guide to self-similar traffic and performance modeling for modern high-speed networks," in *Stochastic Networks*, F. P. Kelly, S. Zachary, and I. Ziedins, eds. (Oxford University, Oxford, UK, 1996), pp. 339–366.
6. M. S. Taqqu, W. Willinger, and R. Sherman, "Proof of a fundamental result in self-similar traffic modeling," ACM/SIGCOMM Comput. Commun. Rev. **27,** 5–23 (1997).
7. K. Claffy, G. Miller, and K. Thompson, "The nature of the beast: recent traffic measurements from an Internet backbone," in *Proceedings of the Internet Society (INET '98)* (Internet Society, 1998), http://www.isoc.org/isoc/conferences/inet/98/proceedings/.
8. D. Sala and A. Gummalla, "PON functional requirements: services and performance," presented at the IEEE 802.3ah meeting in Portland, Ore., July 2001. Available at http://www.ieee802.org/3/efm/public/jul01/presentations/sala_1_0701.pdf.
9. M. W. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," *Proceedings of the Conference on Communications Architectures, Protocols and Applications* (Association for Computing Machinery, 1994), pp. 269–280, http://doi.acm.org/10.1145/190314.190339.
10. L. Brakmo, S. O'Malley, and L. Peterson, "TCP Vegas: new techniques for congestion detection and avoidance," in *Proceedings of the Conference on Communications Architectures, Protocols and Applications* (Association for Computing Machinery, 1994), pp. 24–35, http://doi.acm.org/10.1145/190314.190317.
11. R. L. Carter and M. E. Crovella, "Measuring bottleneck link speed in packet-switched networks," TR-96-006 (Department of Computer Science, Boston University, Boston, Mass., 15 March 15 1996).
12. J. Diagle and J. Langford, "Models for analysis of packet voice communications systems," IEEE J. Sel. Areas Commun. **4,** 847-855 (1986).
13. A. K. Parekh and R. G. Gallager, "A generalized processor sharing approach to flow control in integrated services networks—the single node case," IEEE/ACM Trans. Netw. **12,** 344-357 (1993).