QoS Control Schemes for Two-Stage Ethernet Passive Optical Access Networks

Abdallah Shami, *Member, IEEE*, Xiaofeng Bai, *Student Member, IEEE*, Nasir Ghani, *Senior Member, IEEE*, Chadi M. Assi, *Member, IEEE*, and Hussein T. Mouftah, *Fellow, IEEE*

***Abstract*—Ethernet passive optical networks (EPONs) have emerged as the one of the most promising candidates for next-generation access networks. These new architectures couple low-cost optics with advanced edge electronics to offer vastly improved scalability over competing digital subscriber line and cable modem offerings. This paper proposes several novel architectural enhancements for EPON, which will help increase the viability of optical access over a broader range of subscriber access scenarios. Specifically, this paper proposes a two-stage EPON architecture that allows more end-users to share an optical line terminal link, and enables longer access reach/distances (beyond the usual 25 km distance). In addition, a new dynamic bandwidth allocation (DBA) algorithm is proposed to effectively allocate bandwidths between end users. This DBA algorithm can support differentiated services in a network with heterogeneous traffic. We conduct detailed simulation experiments to study the performance and validate the effectiveness of the proposed architecture and algorithms.**

***Index Terms*—Access network, dynamic bandwidth allocation algorithm (DBA), Ethernet-based passive optical network (EPON), quality-of-service (QoS), simulation and modeling.**

I. Introduction

RECENTLY, there has been a dramatic increase in data traffic, driven primarily by the explosive growth of the Internet, as well as the proliferation of corporate *virtual private networks* (VPNs) [1]. As traffic demands have grown, many carriers have been prompted to add capacity quickly and in the most cost-effective way possible. As result, new core optical networks have been extensively deployed, and in particular, the use of *dense wavelength-division-multiplexing* (DWDM) technology has dramatically increased the capacity of these networks [2], [4]. At the same time, enterprise *local-area networks* (LANs) technologies have steadily scaled tributary speeds progressively from 10 and 100 Mb/s upwards toward multigigabit speeds, e.g., 1.0, 10 Gb/s Ethernet.

Overall, the above developments have led to a growing "access bottleneck," where metro/regional and backbone capacities are vastly out-scaling last-mile bandwidths. Although access technologies such as *digital subscriber line* (DSL) and *cable modem* (CM) offer affordable solutions for residential data users, they pose fundamental distance and bandwidth limitations. For example, many renditions largely limit end-users to speeds under 10 Mb/s and distances under 5 km. Hence, these technologies lack broader universality for business (or small business) settings. Clearly, new and improved access solution technologies are required. These offerings must be inexpensive, yet still be capable of scaling to delivering bundled data, voice and video over the same high-speed connection. Additionally, other prime concerns are *quality-of-service* (QoS) guarantee provisions, and the ability to purchase bandwidth on an as needed basis [6].

It is here that *Ethernet passive optical networks* (EPONs) [6] have emerged as the best candidate for next-generation access networks. Propelled by rapid price declines in fiber optics and Ethernet components [3], [5], these new EPON architectures combine the latest in optical and electronic advances and are poised to become the dominant means of delivering bundled services over a single platform [3], [5]. An EPON is basically a *point-to-multipoint* (1:N) optical access network with no active elements in the signal path. The network provides two-way operation (Fig. 1), in which traffic from an *optical line terminal* (OLT) is sent to/from multiple *optical network units* (ONUs). Namely, OLT-ONU traffic is called "downstream" (point-to-multipoint) and meanwhile, reverse ONU-OLT direction traffic is called "upstream" (multipoint-to-point) [3]. The latter requires contention resolution (arbitration) mechanisms to avoid upstream transmission collision between ONU senders.

The OLT typically resides in a *central office* (CO) location and connects the optical access network to the metro (backbone) network. Meanwhile, the ONU is usually located at or near end-user locations and must support a wide array of services—broadband video, voice, data, etc. In particular, various ONU deployment possibilities exist, as per different architectures such as *fiber-to-the-curb* (FTTC), *fiber-to-the-building* (FTTB), and *fiber-to-the-home* (FTTH) [6]. Overall, the operational costs of these setups are minimal since no active elements are placed in the outside fiber plant, e.g., no maintenance is needed in the field. Moreover, by sharing the network equipment among the maximum number of customers, operators can amortize the cost of installation and operation in a much more economical manner.

However, despite the above saliencies, the overall adoption of EPON technologies in access networks has met with various obstacles. As a result, notable research efforts are underway in order to evolve more capable technical solutions at highly

Manuscript received June 10, 2004; revised April 26, 2005. This work was supported in part by the National Sciences and Engineering Research Council of Canada (NSERC).

A. Shami and X. Bai are with the Department of Electrical and Computer Engineering, University of Western Ontario, London, ON N6A 5B9, Canada (e-mail: ashami@eng.uwo.ca; xbai6@uwo.ca).

N. Ghani is with the Department of Electrical and Computer Engineering, Tennessee Tech University, Cookeville, TN 38505 USA.

C. M. Assi is with Concordia Institute for Information Systems Engineering, Concordia University, Montreal, QC H3G 1M8, Canada.

H. T. Mouftah is with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON K1N 6N5, Canada.

Digital Object Identifier 10.1109/JSAC.2005.852185

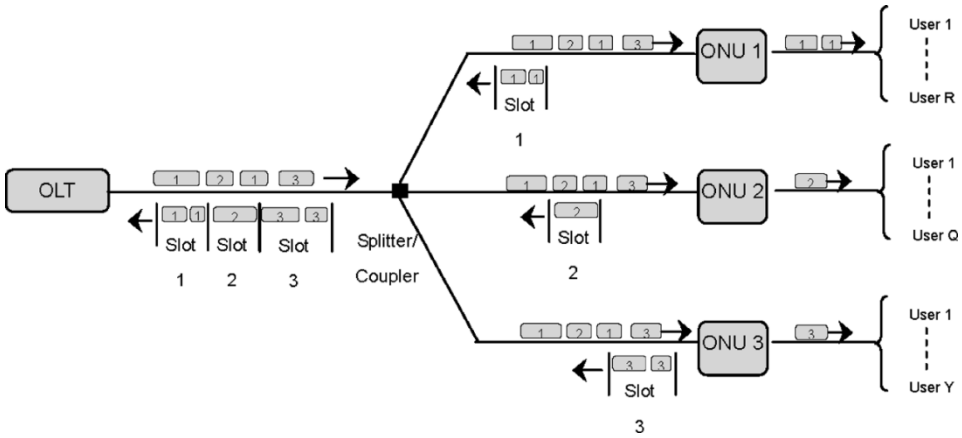


Fig. 1. EPON network architecture.

cost-effective price points. It is the purpose of this work to address some of these issues through devising and demonstrating several novel architectural enhancements for EPON, which will help increase the viability of optical access over a broader range of subscriber access scenarios. Specifically, this paper proposes a *two-stage* EPON architecture that allows more end-users to share an OLT link and enables longer access reach/distances (beyond the usual 25 km distance). Related *dynamic bandwidth allocation* (DBA) [11] algorithms are also presented to effectively allocate bandwidth between users in these two-stage EPON architectures. These DBA algorithms can support differentiated services in a network with heterogeneous traffic. We conduct detailed simulation experiments to study the performance and validate the effectiveness of the proposed architecture and algorithms.

The rest of this paper is organized as follows. Section II presents a background to motivate our work and proposes the two-stage EPON network model. In Section III, we present a novel bandwidth allocation algorithm with QoS support to reduce the average packet queueing delay and packet loss. The performances of the proposed bandwidth allocation algorithms are studied and analyzed in Section IV, and Section V concludes this paper.

II. Overview and Proposed Architecture

In order to properly introduce the planned architecture, it is instructive to first consider the existing body of work in Ethernet QoS and architectures.

A. Background and Previous Work

In general, an EPON network cannot be treated as a basic shared medium network, i.e., one using *carrier sense multiple access with collision detection* (CSMA/CD) type protocols. At the same time, neither can an EPON network be treated as a point-to-point network. Instead, it is a combination of both types. For example, consider upstream ONU-OLT communications. Here, due to large propagation delays across EPON infrastructures (which can easily exceed 20 km), the effectiveness of regular CSMA/CD protocols is greatly reduced. Instead, these protocols are more suited for local-area network (LAN) domains, where links are short and traffic predominantly comprises "best-effort" data. Since EPON access [6] must support much more stringent service requirements (QoS), related architectures must provide strict guarantees on such as packet delay and jitter performance. In order to accommodate diverse traffic types and to achieve bandwidth sharing and flow isolation, clearly efficient *bandwidth allocation* algorithms must be developed.

To date, many researchers [5]–[11] have surmised that *time-sharing* protocols represent the best method of optical channel sharing in optical access networks, i.e., *time-division multiple access* (TDMA). This approach allows the ONUs to share a single upstream wavelength in which the OLT allocates timeslots to each ONU to transmit its backlogged traffic. Overall, this yields a very cost-effective solution, and Fig. 1 illustrates time-shared data flow in an EPON. Time-sharing techniques can either be static or dynamic. In the former, each ONU is allocated a fixed timeslot to transmit data. Although this is a rather simple approach, its implementation is not contingent with EPON's requirements, e.g., QoS support, OLT link efficiency. Therefore, more effective and dynamic schemes will be needed in order to successfully implement service guarantees in the next generation access networks.

Along these lines, [9] presents a simple algorithm for dynamic bandwidth allocation based upon a time interleaving method, i.e., the *interleaved polling scheme with an adaptive cycle time* (IPACT) scheme. Here, an in-band signaling approach is used to allow the use of a single wavelength for both downstream data and grant transmission. Meanwhile, in [8], the authors studied EPON performance with a static bandwidth-assignment algorithm. Subsequently, a large body of work has defined various adaptive DBA schemes in order to boost OLT link efficiency and inter-ONU fairness, see [8], [11] and related references. Here, notable examples include the *limited*, *gated*, *linear credit*, and *elastic* allocation schemes. However, in all of the above-mentioned studies, the authors treat all traffic as belonged to a single class/type of service, i.e., no specific considerations made for differing service types at an ONU. This issue is addressed in [7] and [13], where the bandwidth allocation algorithms are augmented to support QoS in a differentiated services framework. To alleviate the light-load penalty, ONU nodes are partitioned into two groups,

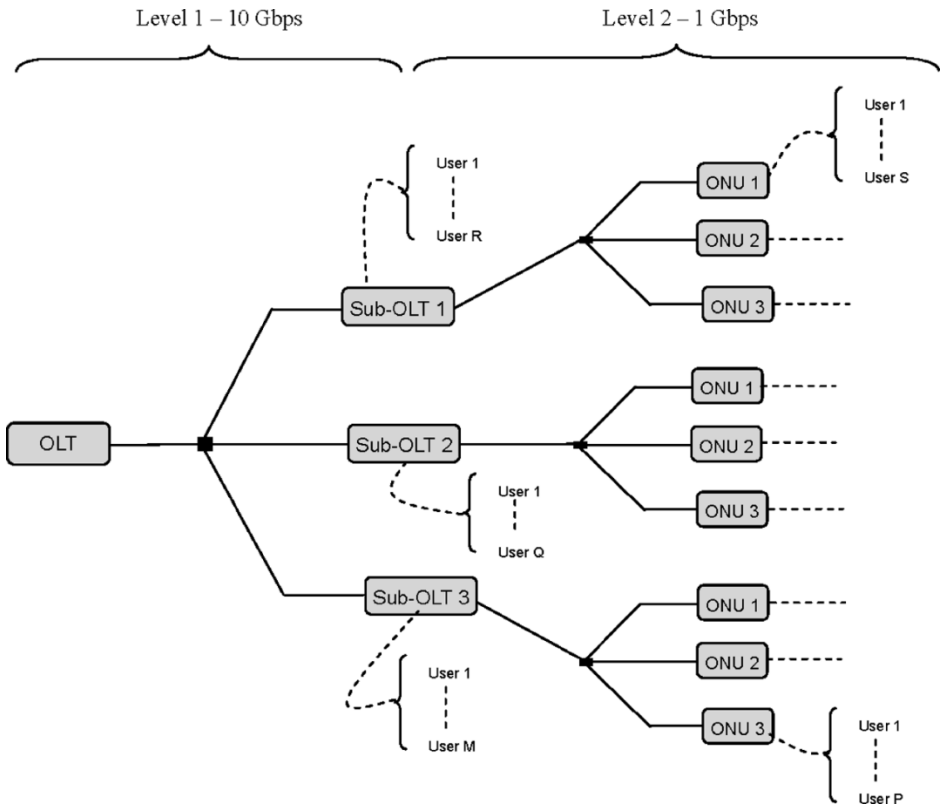


Fig. 2. P-EPON network architecture.

underloaded and *overloaded*. Here, the grants for underloaded group are issued before every ONU reports to the OLT, thereby the intertransmission time of an underloaded ONU is reduced, see DBA2 scheme in [7] for details. Another theoretical contribution on the fairness issue in EPON is given by the *FQSE* protocol introduced in [14]. FQSE extends sibling fair in one ONU to cousin fair at the global level. This protocol also provides a generalized approach to offer fairness guarantee for arbitrary levels of hierarchical structure. Jitter performance in EPON networks is studied in [15]. Here, a transmission cycle is divided into two subcycles, i.e., EF subcycle and AF subcycle. Though the scheduling frame size is not fixed, by protecting EF service in a separate subcycle, its jitter performance is considerably improved.

B. Proposed Architecture

Let us consider the actual practical deployment of EPON technologies. In order to increase the technology's universal appeal, ideally, network designers want solutions that are capable of achieving high penetration within the access loop. Namely, this implies the ability to serve a wide range of end-user types, both high-bandwidth users, and lower-to-moderate bandwidth users. However, field trial studies have shown PON to be costly, particularly, for the latter types of users. Hence, there is a crucial need to adapt the EPON framework in order to boost its appeal across a wider end-user based. In this paper, we present a cascaded two-stage EPON architecture which adds another intermediate level of ONU nodes to the network, termed sub-OLT, as shown in Fig. 2. Hereafter, we refer to this architecture as *penetrated-EPON* (P-EPON). The introduction of sub-OLT node is

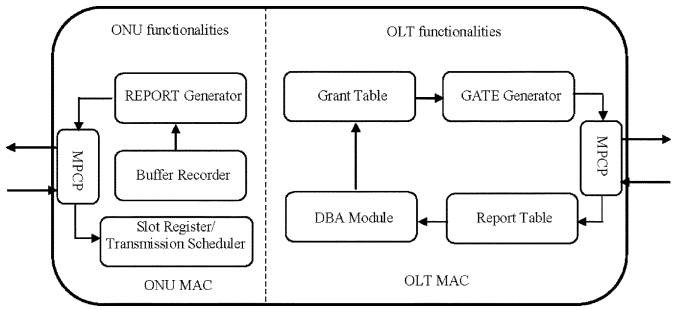


Fig. 3. Illustration of sub-OLT functionalities.

to offer the possibility of further dividing the capacity of one ONU amongst multiple lower level bandwidth users, via an even branched PON setup. Therefore, the sub-OLT node functions in a hybridized manner and should assemble the functionalities of a pair of back-to-back connected ONU and OLT nodes, as illustrated in Fig. 3.

Overall, there are several major motivations for deploying a two-stage P-EPON architecture. First, these architectures will allow more end-users (ONU nodes) to share the uplink OLT bandwidth, without incurring extra overhead for switch-over between users. Currently, due to limited power budget in practice only 4–64 ONU nodes can share an OLT link, which typically runs at 1.0 Gb/s capacity. This yields more than 10 Mb/s per user for FTTH deployment. Going forward, it may be very likely that many EPON users will still require much less bandwidth, e.g., 5 Mb/s range. In these cases, the use of a two-stage architecture will allow network operators to further segment OLT capacity amongst more slower speed end-users, yet still

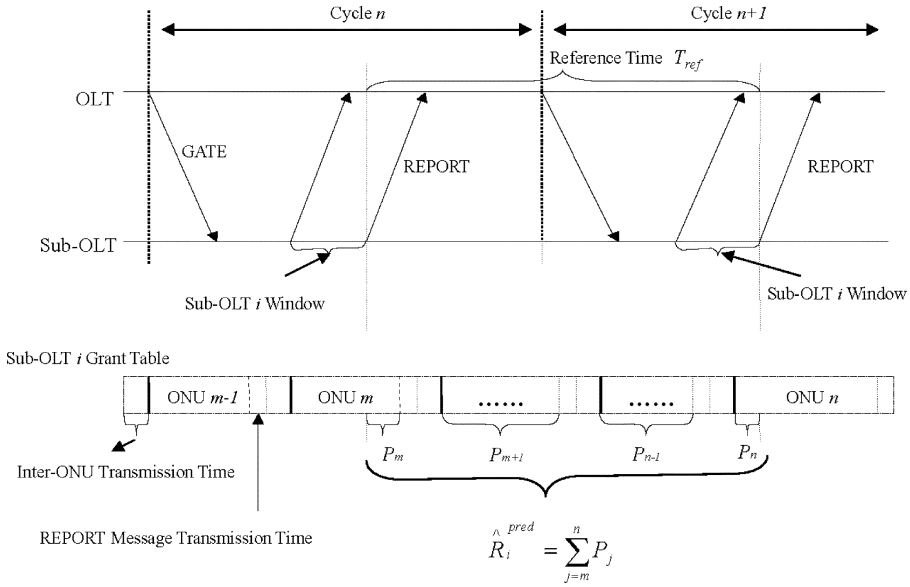


Fig. 4. Illustrative example.

expand their PON footprint/user-base. More importantly, a diverse range of end-user speeds can be supported by a single technology—namely, EPON.

Second, a two-stage EPON architecture will enable longer access reach/distances, since the use of intermediate sub-OLT nodes (Fig. 2) effectively adds another level of electrical regeneration. This will allow operators to extend EPON service offerings beyond the usual 25 km distance, and further service footprint growth. Third, the addition of an extra stage will help reduce OLT hardware complexity significantly. Namely, intermediate sub-OLT nodes can subsume key bandwidth allocation (QoS support) functionalities for downstream ONU nodes, reducing signaling loads, and information state requirements at the head-end OLT. As such, network designers can even consider hybrid-rate EPON setups, in which the OLT can run at higher speeds (e.g., 10 Gb/s) and host distributed sub-OLT nodes running at slower 1.0 Gb/s speeds. Clearly, this setup will help facilitate a very staged, and cost-effective migration from 1.0 Gb/s to future 10 Gb/s EPON networks.

As within EPON settings, DBA algorithms will also be a crucial necessity in P-EPON architectures. Clearly, a comprehensive DBA algorithm has to be developed in order to achieve high (collision-less) throughput and efficient bandwidth sharing among the ONUs (and intermediate sub-OLT nodes). Additionally, this architecture must also support service guarantees for multiple, different *class-of-service* (CoS) types. In particular, since the P-EPON architecture introduces a second level of ONUs (i.e., added buffering), the minimization of end-users packet delay will be a key requirement.

III. Dynamic Bandwidth Allocation With QoS Support for IP-EPONS

In the upstream direction an EPON network acts like a shared medium setting in which all ONU devices can potentially contend while transmitting data. Ideally, at any given time only one ONU should be allowed to occupy the medium. It is also important to state that there is no direct communication between

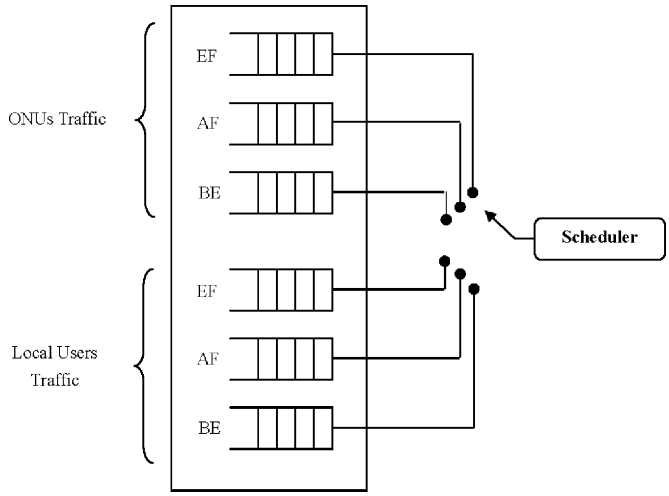


Fig. 5. Sub-OLT queue structure.

ONUs in an EPON-based network. As such, the OLT is the only network element that is able to communicate with the other units in an EPON setup. Note that in the basic EPON standard, the new *multipoint control protocol* (MPCP) is used to implement bandwidth arbitration by the OLT [12].

At the beginning of each transmission cycle, the OLT sends a GATE message to each ONU through the downstream connection. This GATE message contains the following information: time when the ONU should start transmission and the length of its transmission window. Upon receiving its GATE message, the ONU performs synchronization and updates its local parameters. When its local clock matches the transmission start time sent by the OLT, the ONU starts sending information packets to the OLT. At the end of its transmission window, the ONU sends a REPORT message to the OLT to report its current buffer occupancy and request bandwidth for the next transmission cycle.

Upon receiving REPORT message from the ONU, the OLT updates its report table and passes the message to the DBA module responsible for bandwidth allocation decision. At the

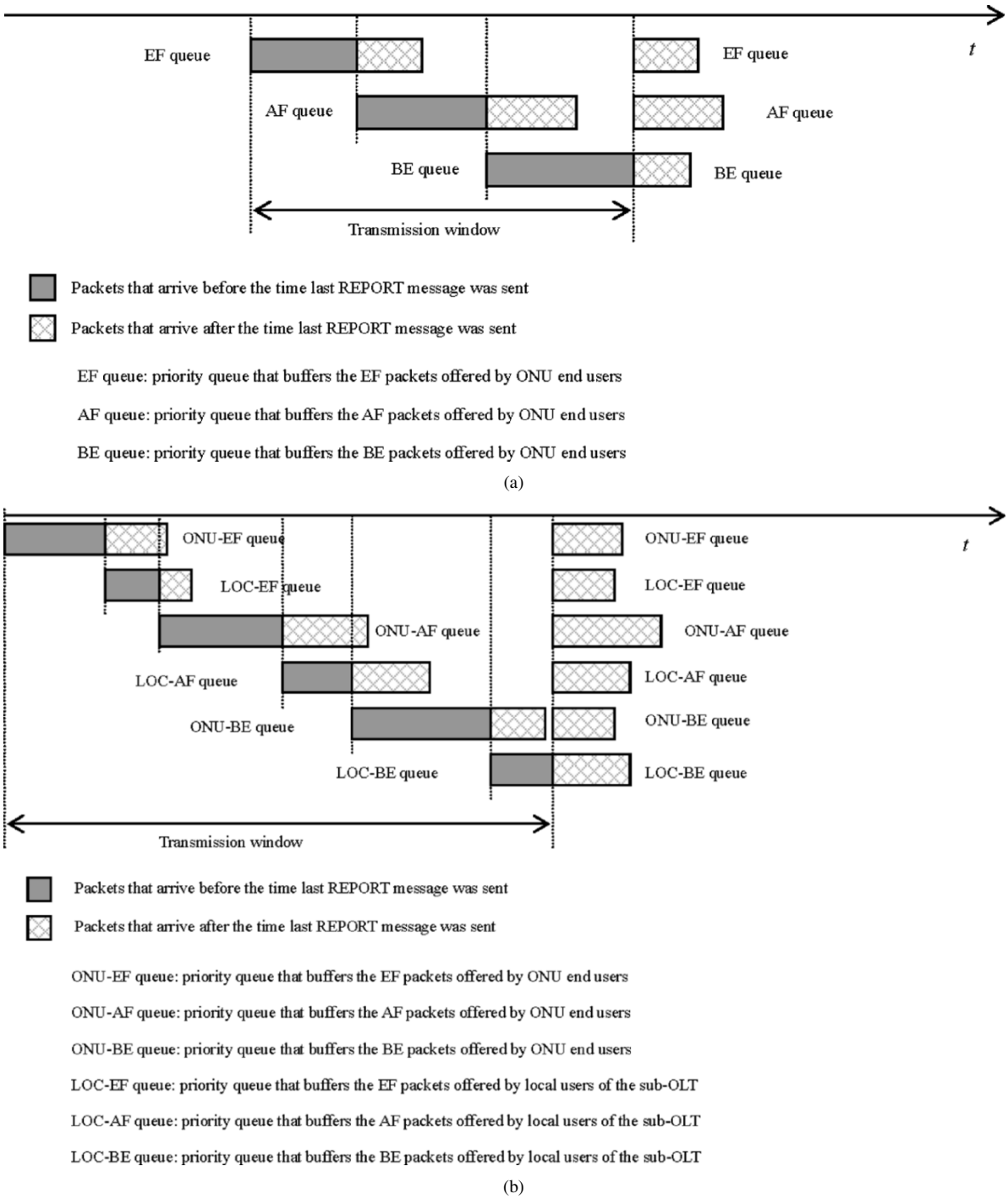
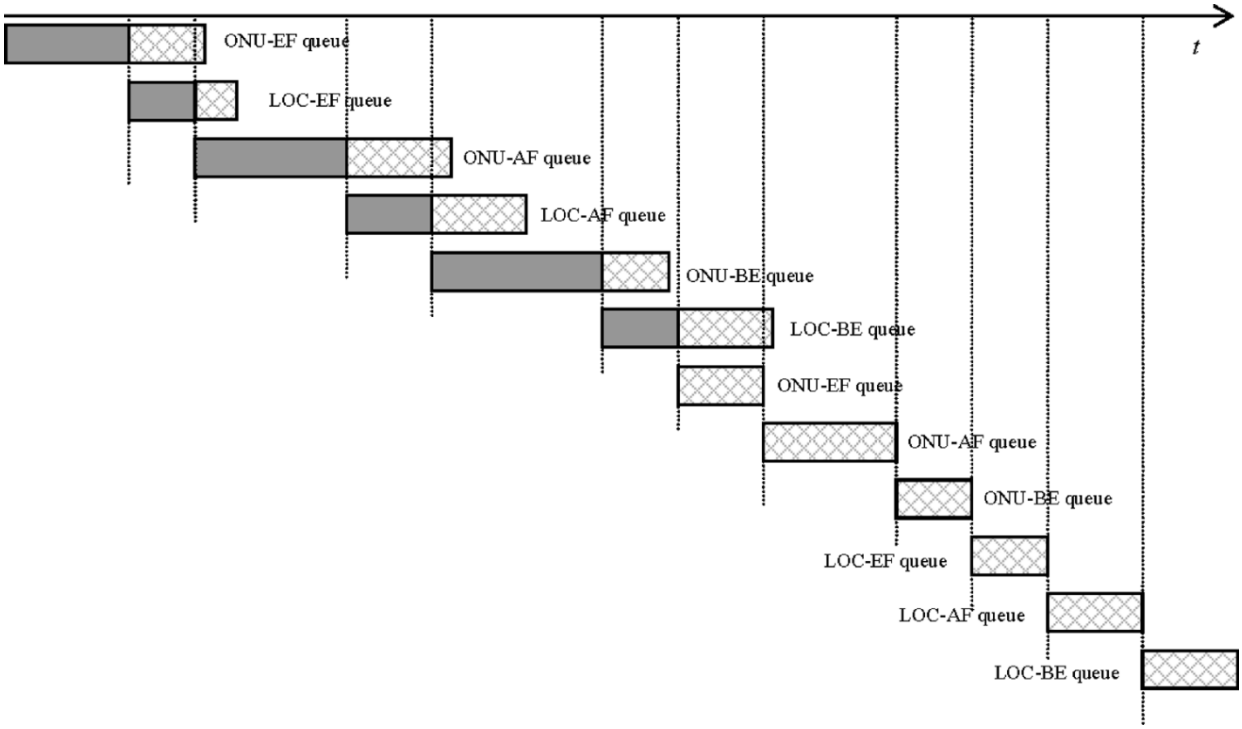


Fig. 6. (a) Intra-ONU scheduling (E-DSA1). (b) Intrasub-OLT scheduling (E-DSA1).

end of the transmission cycle and upon receiving REPORT messages from all ONUs in the network, the DBA module in the OLT calculates the new window size for every ONU in the network. Thereafter, a series of GATE messages are generated and broadcast continuously to the ONUs by the OLT to initiate the next transmission cycle. The MPCP only defines the mechanism for the OLT to arbitrate the transmissions of its attached ONUs and does not specify any details about the bandwidth allocation amongst ONUs. Thus, the development of efficient EPON scheduling algorithms (i.e., dynamic bandwidth allocation schemes) that can accommodate multiple, diverse traffic types, e.g., voice, video, and data is critical for the deployment of such networks.

A. Transmission Cycle Time

We denote transmission cycle time by T_{TCT}, which is the sum of all ONUs' transmission time. In one transmission cycle time every ONU must be able to transmit or/and report to the OLT. Now clearly, if T_{TCT} is too large, the average packet queueing delay inside ONU nodes might increase randomly. In other words, packets that arrive at one ONU between two



(c)

Fig. 6. *Continued:* (c) Intrasub-OLT scheduling (E-DSA2).

consecutive transmissions of this ONU might experience unacceptable delays. Therefore, T_{TCT} should not be too large. We denote the maximum transmission cycle time by $T_{\mathrm{Max-TCT}}$. This T_{TCT}, however, cannot be too small. Under heavy load, if T_{TCT} is restricted too much, the average packet delay might still increase unnecessarily, because more bandwidth is wasted on frequent MPCP messaging and inter-ONU guard times, i.e., the transmissions of two consecutive ONUs are usually separated by a guard time t_g. Moreover, the delay for larger packet sizes may also increase as fragmented transmissions are not permitted at the end of each transmission window [12].

In simple systems, static bandwidth allocation algorithms can be used, in which each ONU is granted a fixed transmission timeslot per transmission cycle. Namely, the ONU timeslot can be calculated as

$$t_i = (T_{\mathrm{TCT}} - N \times t_g) \times w_i \quad \text{with} \quad \left(\sum_{i=1}^{N} w_i = 1\right). \quad (1)$$

Where t_i is the timeslot allocated to ONU i, w_i is the weight of ONU i according to its service level agreement (SLA), and N is the number of ONU nodes. Here, t_i is a fixed transmission timeslot of ONU i under any network traffic load and hence T_{TCT} is consequently fixed, too. However, assigning static capacity to ONU nodes will preclude idle capacity reuse and yields generally lower utilizations for bursty traffic.

To mitigate the above concerns, the authors in [7] proposed an inter-ONU DBA scheme in which ONU nodes are portioned into two groups, *underloaded* and *overloaded*. Here, *underloaded* ONU nodes are those requesting bandwidth below their minimum guarantee defined by the maximum transmission cycle time $T_{\mathrm{Max-TCT}}$, i.e., $B_i^{\min}$ and, hence, their unused capacity is shared in a weighted manner amongst the *overloaded* ONU nodes. This algorithm can be described as follows:

$$T_i = \begin{cases} r_i, & r_i \le t_i^{\min} \\ t_i^{\min} + t_i^{\mathrm{excess}}, & r_i > t_i^{\min} \end{cases} \quad (2)$$

$$t_i^{\mathrm{excess}} = \frac{r_i}{\sum_{k \in K} r_k} t_{\mathrm{total}}^{\mathrm{excess}} \quad (3)$$

$$t_{\mathrm{total}}^{\mathrm{excess}} = \sum_{l \in M} \left(t_l^{\min} - r_l\right) \quad \left(t_{\mathrm{total}}^{\mathrm{excess}} > 0\right). \quad (4)$$

Where T_i and $t_i^{\min}$, respectively, are the timeslot allocated to ONU node i and the minimum guaranteed timeslot (when $B_i^{\min}$ is offered) of this node, r_i is the requested transmission time of ONU node i, $t_{\mathrm{total}}^{\mathrm{excess}}$ is the total excess transmission time saved from *underloaded* ONU nodes (i.e., $r_i < t_i^{\min}$), t_i^{excess} is the corresponding share of the total excess transmission time allocated to *overloaded* ONU node i (i.e., $r_i > t_i^{\min}$), and K and M are the set of *overloaded* and *underloaded* ONU nodes, respectively.

Nevertheless, the above algorithm may still yield some wasted (unused) bandwidth capacity. Namely, this results from the fixed transmission cycle time of $T_{\mathrm{Max-TCT}}$, in which the total excess transmission time might not always be fully occupied by the *overloaded* ONU nodes. To address this deficiency, we modify this algorithm with a more advanced dynamic cycle time upper bounded at $T_{\mathrm{Max-TCT}}$ as follows:

$$t_{\mathrm{total}}^{\mathrm{demand}} = \sum_{k \in K} \left(r_k - t_k^{\min}\right) \quad \left(t_{\mathrm{total}}^{\mathrm{demand}} > 0\right) \quad (5)$$

$$T_i = \begin{cases} r_i, & r_i \le t_i^{\min} \text{ or } t_{\mathrm{total}}^{\mathrm{excess}} \ge t_{\mathrm{total}}^{\mathrm{demand}} \\ t_i^{\min} + t_i^{\mathrm{excess}}, & \text{Otherwise} \end{cases}. \quad (6)$$

```
// This pseudo-code is for the bandwidth allocation of the OLT and sub-OLT nodes in E-DSA1 (no prediction applied):
for (i=0; i<n; ++i)      // n=the number of Sub-OLT nodes in level 1 section or ONU nodes in level 2 section
{
    if (total_excess>=total_demand) {
        grant_table[i].window_size=report_table[i]+trans_time(REPORT);
                                                    // satisfy every bandwidth request
    }
    else if (report_table[i]<=guaranteed_slot) {
        grant_table[i].window_size=report_table[i]+trans_time(REPORT);
                                                    //under loaded nodes get what they requested
    }
    else {
     grant_table[i].window_size=guaranteed_slot+trans_time(REPORT)
                                                + report_table[i]/total_report(Overloaded)*total_excess;
    }                                               //over loaded nodes share the total saved bandwidth fairly.
}
```

Fig. 7. Pseudocode for E-DSA1 DBA scheduler.

```
//This pseudo-code is for the bandwidth allocation of OLT in E-DSA2 (prediction applied):
for (i=0; i<n; ++i)        //n=the number of Sub-OLT nodes
{
        if (total_excess>=total_demand) {
                        // if imminent bandwidth requests can be fully satisfied, predictions will be considered.
            grant_table[i].window_size=report_table[i].immi+trans_time(REPORT);
                        // satisfy every imminent bandwidth request first, then try to accommodate predictions.

            if (total_excess-total_demand>=total_predict) {
                    //satisfy every predicted bandwidth request if the leftover of
                    // the total saved bandwidth from under loaded nodes permits it,
                grant_table[i].window_size+=report_table[i].pred;
            }
        else {
             //otherwise share the leftover of the total saved bandwidth fairly
            grant_table[i].window_size+=report_table[i].pred/total_predict*(total_excess-total_demand);
            }
     }
    else if (report_table[i].immi<=guaranteed_slot) {
                    //if imminent bandwidth requests cannot be fully satisfied, do not consider
                    // predictions.   Therefore, under loaded nodes get what they requested,
        grant_table[i].window_size=report_table[i].immi+trans_time(REPORT);
        }
    else {
                //and over loaded nodes share the total saved bandwidth fairly.
    grant_table[i].window_size=guaranteed_slot+trans_time(REPORT)
                            + report_table[i].immi/total_immi_report(Overloaded)*total_excess;
      }
}
```

Fig. 8. Pseudocode for E-DSA2 DBA scheduler.

Where $t_{\text{total}}^{\text{demand}}$ denotes the total extra transmission time requested by *overloaded* ONU nodes. This modified DBA scheme essentially excludes potential overgranting to ensure higher bandwidth utilization.

B. Sub-OLT Scheduling

Now, we further consider DBA extensions for P-EPON settings. Here, the proposed two-stage model can effectively be segmented into two sections, i.e., first, section one from the OLT to the attached sub-OLT nodes (level 1) and second, section two from sub-OLT nodes to the attached ONU nodes (level 2). Different bandwidth allocation algorithms, therefore, can be applied into these two logically independent sections. The first option for these algorithms may be to apply the above modified DBA scheme into both sections. Thereafter, we refer to this algorithm as *EPON dynamic scheduling algorithm 1* (E-DSA1). Moreover, there is also a special characteristic of P-EPON that we can take advantage of to improve its performance. Specifically, since each sub-OLT maintains a grant table to keep the

grant information for every ONU node attached to it, the future incoming traffic of sub-OLT nodes is predictable in a certain extent, compared with the future incoming traffic of the ONU nodes. This fact permits a sub-OLT to prerequest some bandwidth in its REPORT message for absent packets that will arrive before the next transmission of this sub-OLT, i.e., those that were reported by the ONU nodes in their last REPORT messages. Hence, this scheme can exempt some packets that were not reported to the OLT by the sub-OLT in E-DSA1 from waiting at least one cycle time before being transmitted, i.e., unlike E-DSA1.

In order to implement this algorithm, each sub-OLT must estimate its future incoming traffic from the attached ONUs within some reference time (described in the following paragraph) based upon the information in its grant table. This estimate must be added to the bandwidth request in the REPORT message of the sub-OLT. Upon receiving REPORT messages from sub-OLT nodes, the OLT will first try to satisfy the imminent bandwidth requests (i.e., those bandwidth requests for the packets that are present in the sub-OLT before transmitting the REPORT message). Subsequently, it will try to satisfy the predicted bandwidth requests using the total available bandwidth defined by $T_{\mathrm{Max-TCT}}$. The overall bandwidth prediction approach in a sub-OLT is shown in Fig. 4.

To explain the bandwidth prediction approach more clearly, in Fig. 4, we assume that the link capacity of the level-1 and level-2 sections is the same. The predicted bandwidth request (transmission time) for this sub-OLT is $\hat{R}_i^{\mathrm{pred}}$ and T_{ref} is the reference time. As shown in Fig. 4, the value of $\hat{R}_i^{\mathrm{pred}}$ can be calculated by summing up the scheduled packet transmission time P_j (transmission window minus REPORT message transmission time) in each selected transmission window in the grant table of the sub-OLT. These selected transmission windows should at least partially fall within T_{ref}.

If level-1 and level-2 sections run at different link capacities, the predicted bandwidth request of the sub-OLT, i.e., R_i^{pred} can be simply converted as follows:

$$R_i^{\mathrm{pred}} = \frac{C_2}{C_1}\hat{R}_i^{\mathrm{pred}} = \frac{C_2}{C_1}\sum_{j=m}^{n} P_j \tag{7}$$

where C_1 and C_2 are the link capacities of the level-1 and level-2 sections, respectively, and m and n are the first and last ONU nodes that will transmit in the reference time window T_{ref}. Thereafter, we refer to this improved algorithm that takes into consideration both imminent and predicted bandwidth requests as *EPON dynamic scheduling algorithm 2* (E-DSA2).

Previous studies [5] showed that the traffic in access and local area networks is bursty and has a certain degree of predictability, i.e., if we observe a long burst of data, this burst is likely to continue for some time into the future. This makes it reasonable to assume the transmission cycle time T_{TCT} of the level-1 section will not change dramatically between two consecutive transmission cycles. The reference time T_{ref} mentioned before, therefore, can be estimated as equal to the current cycle time, which is delivered by the GATE message from the OLT to the sub-OLT nodes. Note that EPON DBA schemes require all REPORT messages to be received within a frame period (one transmission cycle), i.e., frame size greater than the largest round-trip

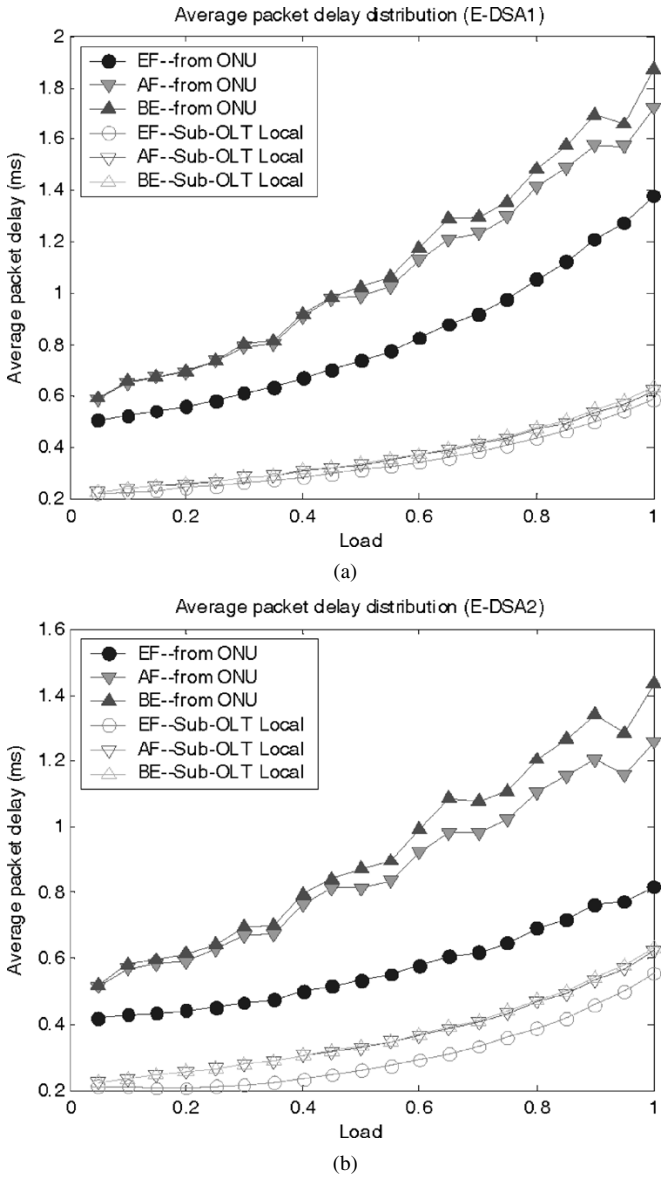


Fig. 9. (a) E-DSA1: Average packet delay. (b) E-DSA2: Average packet delay.

delay. However, since end-of-frame computation time can lead to increased inter-frame idling, underloaded ONU GATE messages can be sent immediately to increase efficiency, see [7] for more details. We also consider this point in E-DSA2.

C. Priority Queueing

In order to support different classes of service with varying packet delay and delay jitter requirements, we introduce three prioritized service classes in our P-EPON model. Namely, the *expedited forwarding* (EF) class provides the highest priority for strict delay sensitive services such as voice. Meanwhile, the *assured forwarding* (AF) class provides a lower, i.e., medium, priority level for services of nondelay sensitive nature, but still requiring guaranteed bandwidth, e.g., video applications. Finally, the *best effort* (BE) class provides the lowest priority for delay tolerable services such as web browsing, e-mail, and file transfer. In every ONU, a separate queue is maintained for each traffic class's packets [7], i.e., *class-based queueing* (CBQ). Received packets are first segregated and classified and then placed

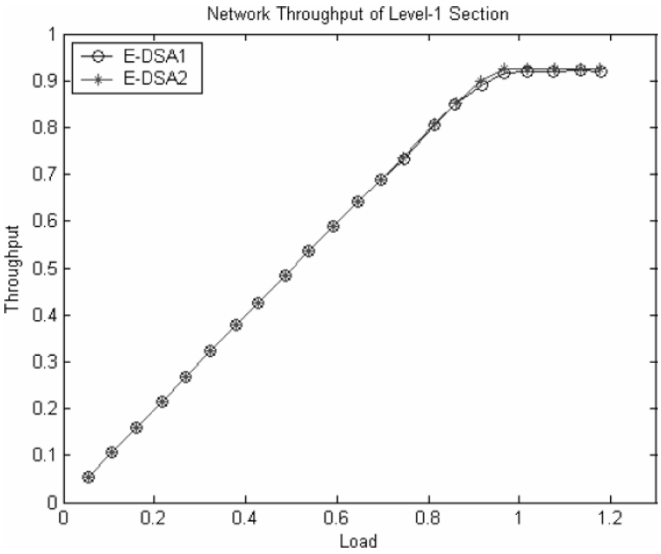


Fig. 10. Level-1 section throughput.

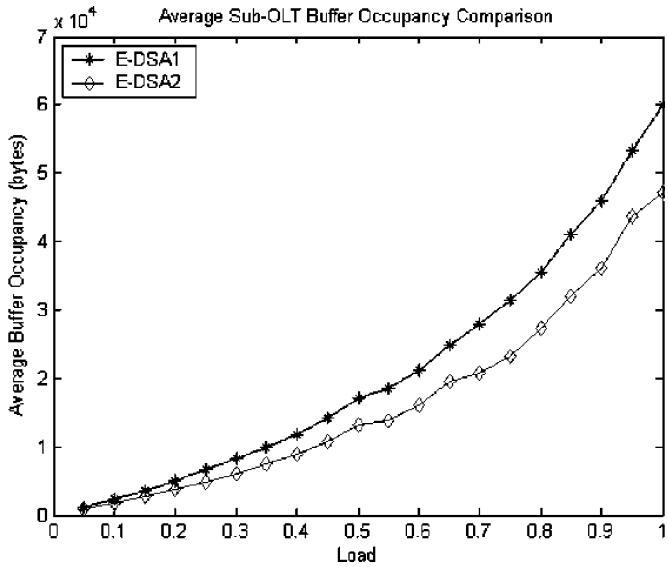


Fig. 11. Average buffer occupancy of sub-OLT node.

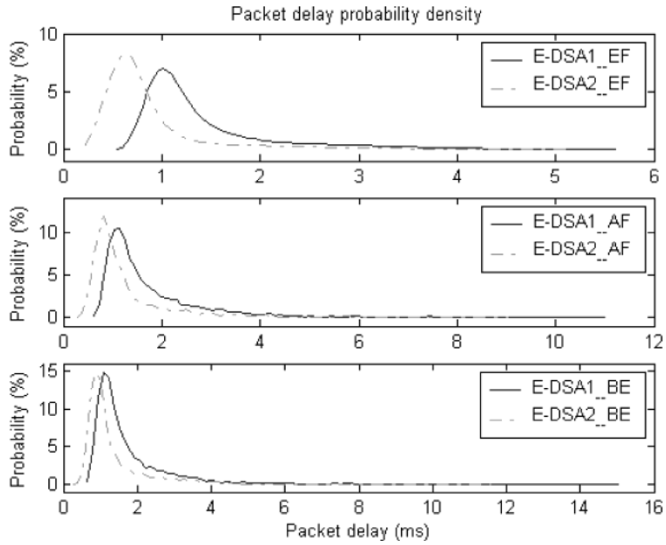


Fig. 12. Probability density of packet delay from ONU nodes.

into their appropriate priority queues [7]. Moreover, a priority-based scheduler is then used for scheduling packet transmissions. This scheduler is now explained further.

To identify and thereby even reduce the queueing delay of the packets traversing across both levels (i.e., packets wait in both ONU and sub-OLT queues before getting to the OLT), we introduce three extra queues in every P-EPON sub-OLT. These are used to buffer and distinguish the respective traffic classes (EF, AF, and BE) offered by ONUs from the ones offered by the local users of this sub-OLT. Namely, each of these three extra queues is granted a *higher* service priority than the queue that buffers packets of the same service class offered by the local users of this sub-OLT. Fig. 5 shows the queue structure of each sub-OLT.

At the intra-ONU and intrasub-OLT level, a two-stage queueing-based priority scheduler is applied in E-DSA1. This scheduler is driven by the following policy: only packets that were reported by the REPORT message of the last transmission cycle may be transmitted in the current transmission window according to their service priorities. This is required because the granted transmission window cannot exceed the bandwidth requested by the REPORT message. With this "cutoff" set by the last REPORT message, we can prevent higher priority queues from unreasonably monopolizing the granted transmission window, which incurs light load penalty for low priority packets. The operation of intra-ONU and intrasub-OLT scheduling is illustrated in Fig. 6(a) and (b).

In E-DSA2, because of the presence of predicted bandwidth requests at the sub-OLTs, in the last REPORT message, the granted bandwidth for the current transmission window is not necessarily only for the reported packets. After scheduling all of the reported packets, the scheduler, therefore, should schedule the unreported packets according to their service priorities. The intrasub-OLT scheduling for E-DSA2 is illustrated in Fig. 6(c). In addition, the pseudocodes in Figs. 7 and 8 explain these two DBA schedulers for E-DSA1 and E-DSA2 with more details.

Moreover, E-DSA2 also provides provision to mitigate bandwidth waste resulting from packet fragmentation. Namely, lower priority packets with smaller sizes are allowed to utilize the residual bandwidth of the transmission window that cannot accommodate the next higher priority packet with larger size, see [11] for details.

IV. Performance Evaluation

The network model in our simulations uses a P-EPON with one OLT, eight sub-OLT nodes, and 32 ONU nodes. Besides its local users, each sub-OLT is connected to four ONU nodes. The network traffic is distributed as follows: 20% from the local users of the sub-OLT nodes and 80% from the end users of ONU nodes. For the traffic model of variable bit rate services, previous studies [10] showed that the actual traffic can be characterized by self-similarity and *long-range dependence* (LRD). Hence, this model is used to generate bursty traffic such as AF and BE in our simulation and the packet sizes are uniformly distributed

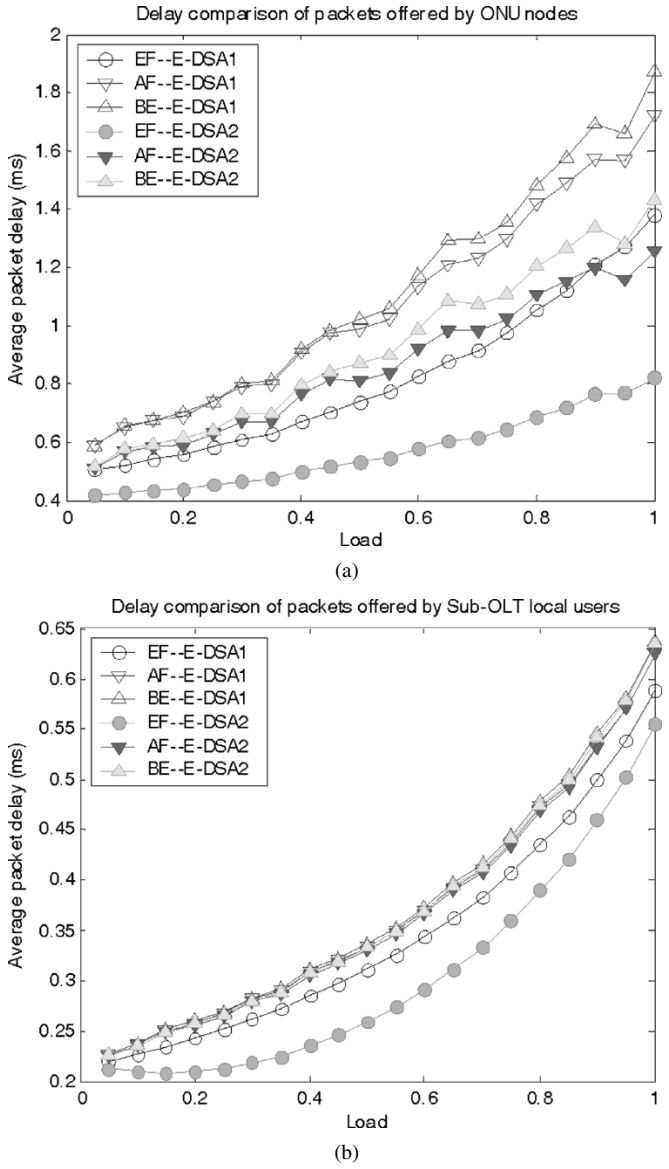


Fig. 13. (a) ONU's average packet delay (access distance 20 km). (b) Sub-OLT average packet delay (access distance 20 km).

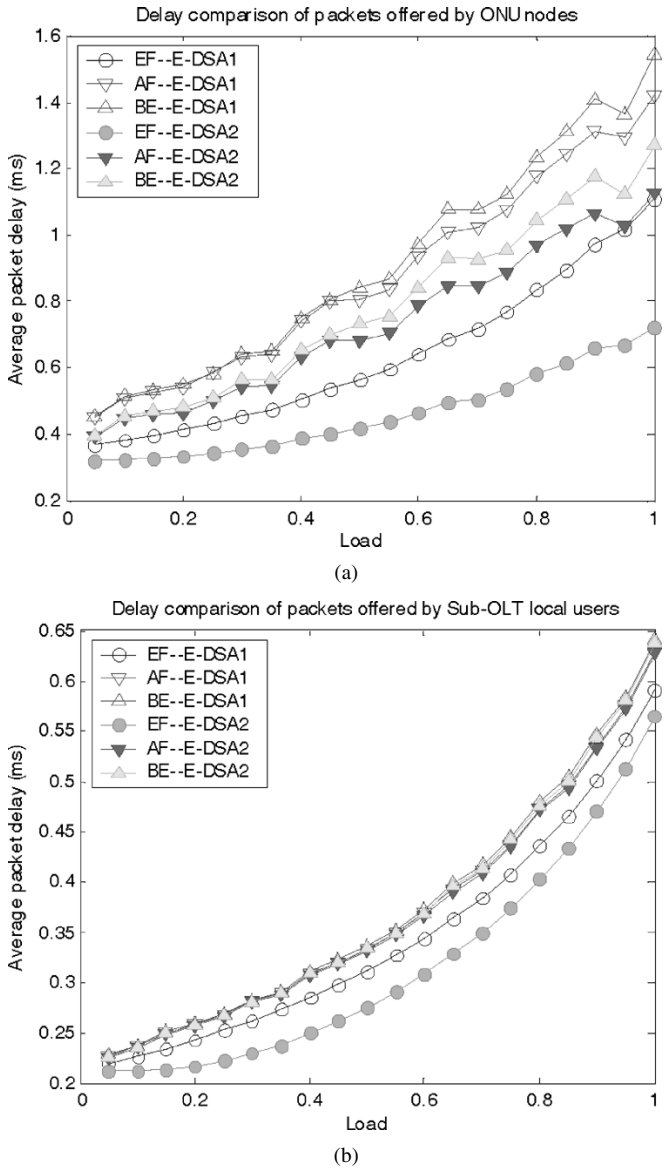


Fig. 14. (a) ONU's average packet delay (access distance 10 km). (b) Sub-OLT average packet delay (access distance 10 km).

between 64 and 1518 bytes. Meanwhile, for high priority services, i.e., EF, nonbursty traffic is modeled using Poisson distributions via a fixed packet size of 70 bytes. Since the actual traffic in access networks mainly consists of bursty streams (i.e., those generated by web-browsing, file transfer, video streams, and the like), the EF proportion is limited to 20% of the total traffic. The remaining 80% is then equally split between the AF and BE classes, i.e., 40% for each. Details for self-similar trace generation can be found in [16].

The link capacities of level-1 and level-2 sections are 10 and 1 Gb/s respectively. The host-slave distances of level-1 and level-2 sections are all equally set at 20 km, the typical value for the maximum operation distance of EPON [8]. In ITU-T Recommendation G.114, the *one way transmission time requirements* states that the delay for voice traffic in access network should be less than 1.5 ms. Here, many studies have verified that for 1 Gb/s OLT-ONU link capacity, 2 ms maximum transmission cycle time is an acceptable value [9], [13]. The guard time separating transmissions of two nodes (ONU or sub-OLT) is further set to 1 μs. The size of shared buffering space at each ONU node is 10 Mbits and each sub-OLT node is 20 Mbits. To implement a DBA using MPCP control messages, one should be aware that the buffering space for each node must match the ingress traffic in order to avoid serious packet loss or displacement. The occupancy policy for each buffering space is as follows: when a higher priority packet arrives and the available buffering space cannot accommodate it, the node (ONU or sub-OLT) will check all the lower priority queues beginning from the lowest one. If this packet can be accepted by displacing some or all of the lower priority packets, it will be accepted. For example, if an arriving AF packet can be accepted by displacing the only five packets buffered in the BE queue, it will be accepted. Otherwise, this AF packet will be dropped and the five BE packets are left untouched. Finally, the transmission delay of a packet in our simulation is defined as the time interval between its entry into the network and

departure from a sub-OLT node. We take into consideration the packet propagation, transmission, queueing delay, as well as the 96 bits interframe gap (IFG) and 64 bits preamble in front of each Ethernet packet in the simulation.

Fig. 9(a) and (b) shows the average packet delays for the E-DSA1 and E-DSA2 schemes (i.e., with/without predicted ONU's traffic, respectively). The results show that E-DSA2 outperforms E-DSA1 for all classes of traffic offered by ONU nodes in terms of average packet delay (reduced by 20% for most loads). Note that the slight delay reduction of EF traffic offered by sub-OLT local users in E-DSA2, compared with E-DSA1, results mainly from the inaccurate bandwidth prediction with T_{TCT} of the sub-OLT nodes (note the reference time was estimated as equal to T_{TCT}). The reason here is that the unreported packets from the above traffic flow always have the first chance to access any positive offset introduced by inaccurate predictions.

The improved packet delay performance in E-DSA2, however, is not with the cost of throughput degradation in level-1 section of P-EPON. This is verified in Fig. 10. The slightly higher throughput of E-DSA2 (92.5%) over E-DSA1 (92%) is raised by the mitigated packet fragmentation in E-DSA2. Additionally, since E-DSA2 exempts some packets from waiting in the buffer of sub-OLT nodes before being forwarded to the OLT, it also reduces the buffer occupancy in every sub-OLT node. This certainly relieves the packet loss with a limited buffering space. Fig. 11 gives an example of the average buffer occupancy in a sub-OLT node, which confirms the betterment introduced by E-DSA2.

In Fig. 12, we can see for packets offered by ONU nodes, E-DSA2 reduces their service delay just by "shifting" the delay distribution toward zero without introducing more severe delay variation. As a result, the maximum packet delay is reduced as well. For example, at full load the maximum packet delay of 5.641, 11.0664, and 15.1363 ms, respectively, for EF, AF, and BE services in E-DSA1 are moved to 5.2129, 9.1048, and 11.6536 ms in E-DSA2, as shown in Fig. 12.

Finally, Fig. 14(a) and (b) shows the performance of E-DSA1 and E-DSA2 for the case where level-2 section distances are set at 10 km (Fig. 13(a) and (b) shows the performance for the 20 km case). Again, we can see that the E-DSA2 scheme performs well without hurting the QoS provisions for sub-OLT local users.

V. Conclusion

In this paper, we reviewed the overall architectures and bandwidth allocation algorithms of emerging EPON designs and proposed a novel two-stage enhancement, penetrated-EPON (P-EPON). This scheme utilizes an intermediate state to help increase universality and boost distance and coverage within the access domains. A comprehensive DBA algorithm for P-EPON is also tabled and its throughput-delay performance studied using simulation techniques for various realistic network settings. Overall, these findings help validate the effectiveness of the proposed architecture and its new DBA algorithms.

References

[1] K. G. Koffman and A. M. Odlyzko, "Internet growth: Is there a "Moore's law" for data traffic?," in *Handbook of Massive Data Sets*. Norwell, MA: Kluwer, 2001.

[2] B. Mukherjee, "WDM optical communication networks: Progress and challenges," *IEEE J. Sel. Areas Commun.*, vol. 18, no. 10, pp. 1810–182, Oct. 2000.

[3] B. Lung, "PON architecture 'Future proofs' FTTH," *Lightwave*, vol. 16, pp. 104–7, Sep. 1999.

[4] N. Ghani, S. Dixit, and T.-S. Wang, "On IP-over-WDM integration," *IEEE Commun. Mag.*, vol. 38, no. 3, pp. 72–84, Mar. 2000.

[5] Alloptic. "Ethernet Passive Optical Networks," Whitepaper. International Engineering Consortium (IEC). [Online]. Available: http://www.iec.org

[6] G. Kramer and G. Pesavento, "Ethernet passive optical network (EPON): Building a next-generation optical access network," *IEEE Commun. Mag.*, vol. 40, pp. 66–73, Feb. 2002.

[7] C. Assi, Y. Ye, S. Dixit, and M. A. Ali, "Dynamic bandwidth allocation for quality-of-service over Ethernet PONs," *IEEE J. Sel. Areas Commun.*, vol. 21, no. 9, pp. 1467–1477, Nov. 2003.

[8] G. Kramer and B. Mukherjee, "Ethernet PON (EPON): Design and analysis of an optical access network," *Photonic Netw. Commun. J.*, vol. 3, pp. 307–319, Jul. 2001.

[9] G. Kramer, B. Mukherjee, and G. Pesavento, "Interleaved polling with adaptive cycle time (IPACT): A dynamic bandwidth distribution scheme in an optical access network," *IEEE Commun. Mag.*, vol. 40, pp. 74–80, Feb. 2002.

[10] H. D. J. Jeong, D. McNickle, and K. Pawlikowski, "Generation of self-similar time series for simulation studies of telecommunication networks," presented at the 1st Western Pacific/3rd Australia-Jpn. Workshop Stochastic Models, Christchurch, New Zealand, Sep. 1999.

[11] N. Ghani, A. Shami, C. Assi, and M. Y. A. Raja, "Intra-ONU bandwidth scheduling in Ethernet passive optical networks," *IEEE Commun. Lett.*, vol. 8, no. 11, pp. 683–686, Nov. 2004.

[12] IEEE 802.3ah Task Force Homepage [Online]. Available: http://www.ieee802.org/3/efm

[13] C. Assi, Y. Ye, and S. Dixit, "Support of QoS in IP-based Ethernet-PON," in *Proc. IEEE GLOBECOMM 2003*, San Francisco, CA, Dec. 2003, pp. 3737–3741.

[14] G. Kramer, A. Banerjee, N. K. Singhal, B. Mukherjee, S. Dixit, and Y. Ye, "Fair queueing with service envelopes (FQSE): A cousin-fair hierarchical scheduler for subscriber access networks," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 8, pp. 1497–1513, Oct. 2004.

[15] A. Shami, X. Bai, C. Assi, and N. Ghani, "Jitter performance in Ethernet passive optical networks," *IEEE J. Lightw. Technol.*, vol. 23, no. 4, pp. 1745–1753, Apr. 2005.

[16] X. Bai and A. Shami. Modeling self-similar traffic for network simulation. [Online]. Available: http://www.eng.uwo.ca/people/ashami/Publications.htm

Abdallah Shami (S'01–M'03) received the B.E. degree in electrical and computer engineering from the Lebanese University, Beirut, Lebanon, in 1997, the M.S. degree in computer and electrical engineering from the Saint Joseph/Lebanese University, Beirut, Lebanon, in 1998, and the Ph.D. degree in electrical engineering from the Graduate School and University Center, City University of New York, New York, in September 2002.

In September 2002, he joined the Department of Electrical Engineering at Lakehead University, ON, Canada, as an Assistant Professor. Since July 2004, he has been with the University of Western Ontario, London, ON, Canada, where he is currently an Assistant Professor in the Department of Electrical and Computer Engineering. His current research interests are in the area of data/optical networking, EPON, optical packet switching, Internet protocol over wavelength-division multiplexing, and software tools.

Dr. Shami held the Irving Hochberg Dissertation Fellowship Award at the City University of New York and a GTF Teaching Fellowship.

Xiaofeng Bai (S'04) received the B.Eng. degree in communication engineering from Shandong University, Shandong, China, in 1997, the M.E.Sc. degree in electrical and computer engineering from the University of Western Ontario, London, ON, Canada, in 2005. He is currently working towards the Ph.D. degree in electrical and computer engineering at the University of Western Ontario.

His research interests are in the area of quality-of-service in broadband access networks, medium access control protocol design, and network traffic modeling.

Nasir Ghani (S'96–M'97–SM'01) has a wide range of experience in a broad range of networking-related fields and has held senior technical positions at Sorrento Networks, Nokia Research, IBM, and Motorola Codex. Currently, he is with the Electrical and Computer Engineering Department, Tennessee Tech University, Cookeville, TN, where his research focus includes metro/edge networks, access networks, optical networks, Internet protocol routing, and network survivability. He has published more than 40 refereed journal and conference papers, several book chapters, and has two patents granted.

Dr. Ghani was full conference Co-Chair for the International Society for Optical Engineers (SPIE) OPTICOMM 2003 and has served as a program committee member for the IEEE, SPIE, Association for Computing Machinery (ACM), and International Electrotechnical Commission (IEC) conferences. Currently, he is also a Co-Chair for the Optical Networking Symposium for the IEEE ICC 2006 and 2007, and also IEEE GLOBECOM 2006. He is an Associate Editor for the IEEE COMMUNICATIONS LETTERS and has also served as a Guest Editor for the *IEEE Network*.

Chadi M. Assi (M'03) received the B.S. degree in engineering from the Lebanese University, Beirut, Lebanon, in 1997 and the Ph.D. degree from the Graduate Center, City University of New York, New York, in April 2003. He was a Visiting Researcher at Nokia Research Center, Boston, MA, from September 2002 to August 2003, working on quality-of-service in optical access networks. He joined the Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, QC, Canada, in August 2003 as an Assistant Professor. His research interests are in the areas of optical networks control, provisioning and restoration, and ad hoc networks.

Dr. Assi received the Mina Rees Dissertation Award from the City University of New York in August 2002 for his research on wavelength-division-multiplexing optical networks.

Hussein T. Mouftah (M'76–SM'80–F'90) joined the School of Information Technology and Engineering (SITE), University of Ottawa, Ottawa, ON, Canada, in September 2002, as a Canada Research Chair (Tier 1) Professor of Optical Networks. He was with the Electrical and Computer Engineering Department, Queen's University (1979–2002), where he was prior to his departure a full Professor and Department Associate Head. He has three years of industrial experience mainly at BNR of Ottawa, now Nortel Networks (1977–1979). He is the author or coauthor of five books, 22 book chapters, and more than 700 technical papers and eight patents.

Dr. Mouftah was the recipient of the 1989 Engineering Medal for Research and Development of the Association of Professional Engineers of Ontario (PEO), and the Ontario Distinguished Researcher Award of the Ontario Innovation Trust (2002). He is the joint holder of the Best Paper Award for a paper presented at SPECTS 2002, and the Outstanding Paper Awards for IEEE HPSR 2002 and the IEEE ISMVL 1985. He served as Editor-in-Chief of the *IEEE Communications Magazine* (1995–1997), the IEEE ComSoc Director of Magazines (1998–1999), and Chair of the Awards Committee (2002–2003). He has been a Distinguished Speaker of IEEE ComSoc since 2000.