3RD QUARTER 2008, VOLUME 10, NO. 3

IEEE COMMUNICATIONS SURVEYS

www.comsoc.org/pubs/surveys

ETHERNET PASSIVE OPTICAL NETWORK ARCHITECTURES AND DYNAMIC BANDWIDTH ALLOCATION ALGORITHMS

MICHAEL P. MCGARRY, UNIVERSITY OF AKRON
MARTIN REISSLEIN, ARIZONA STATE UNIVERSITY
MARTIN MAIER, INSTITUT NATIONAL DE LA RECHERCHE SCIENTIFIQUE (INRS)

ABSTRACT

We compile and classify the research work conducted for Ethernet passive optical networks. We examine PON architectures and dynamic bandwidth allocation algorithms. Our classifications provide meaningful and insightful presentations of the prior work on EPONs. The main branches of our classification of DBA are: grant sizing, grant scheduling, and optical network unit queue scheduling. We further examine the topics of QoS support, as well as fair bandwidth allocation. The presentation allows those interested in advancing EPON research to quickly understand what already was investigated and what requires further investigation. We summarize results where possible and explicitly point to future avenues of research.

Over the past decade, the telecommunications infrastructure has transitioned from a copper-based plant to a fiber-based plant. The transition began with the wide area networks (WANs) that provide connectivity between cities and progressed through the metropolitan area networks (MANs) that provide connectivity between service provider locations within a metropolitan area. At the same time, local area networks (LANs) that interconnect nodes within an individual location have seen average bit rates migrate from 10 Mb/s to 1 Gb/s over copper cabling. Although significant bandwidth improvements occurred in the service provider networks (i.e., WANs and MANs), as well as at the subscriber premises (i.e., LANs), the link between the private customer networks and the public service provider networks did not experience the same level of progress. This so-called access network that provides the link between the private and public networks still relies on an aging copper infrastructure. The xDSL and cable modem technology developments made marginal improvements in bandwidth capacity but failed to open the bottleneck that exists in access networks.

A fiber infrastructure is required in the access networks to provide higher bit rates, as well as more flexibility. From the service provider perspective, access network links have different revenue dynamics than links in the WAN and MAN. Whereas WAN and MAN links carry the bit streams of many revenue generating customers, access network links carry a single or only a few revenue generating bit streams. For this reason, access networks are very sensitive to cost. Cost issues are slowing the deployment of a new physical plant in the access networks.

Deploying a passive optical network (PON) between service providers and customer premises can provide a cost efficient and flexible infrastructure that will provide the required bandwidth to customers for many years to come. PONs are a network in which a shared fiber medium is created using a passive optical splitter/combiner in the physical plant. Sharing the fiber medium means reduced cost in the physical fiber deployment, and using passive components in the physical plant means reduced recurring costs by not maintaining remote facilities with power. These reduced costs make PONs an attractive choice for access networks, which are inherently cost sensitive.

At a top level, PONs are classified by the used link-layer protocol. Whereas an asynchronous transfer mode (ATM) PON (APON) uses ATM, an Ethernet PON (EPON) uses Ethernet, and a gigabit PON (GPON) uses the GPON encapsulation method (GEM) in addition to ATM cells to support Ethernet. The International Telecommunication Union (ITU) has generated standards for APONs: G.983 broadband PON (BPON), as well as GPONs: G.984 gigabit-capable PON (GPON). The IEEE has generated a standard for EPONs: IEEE 802.3ah Ethernet in the first mile. Given the fact that

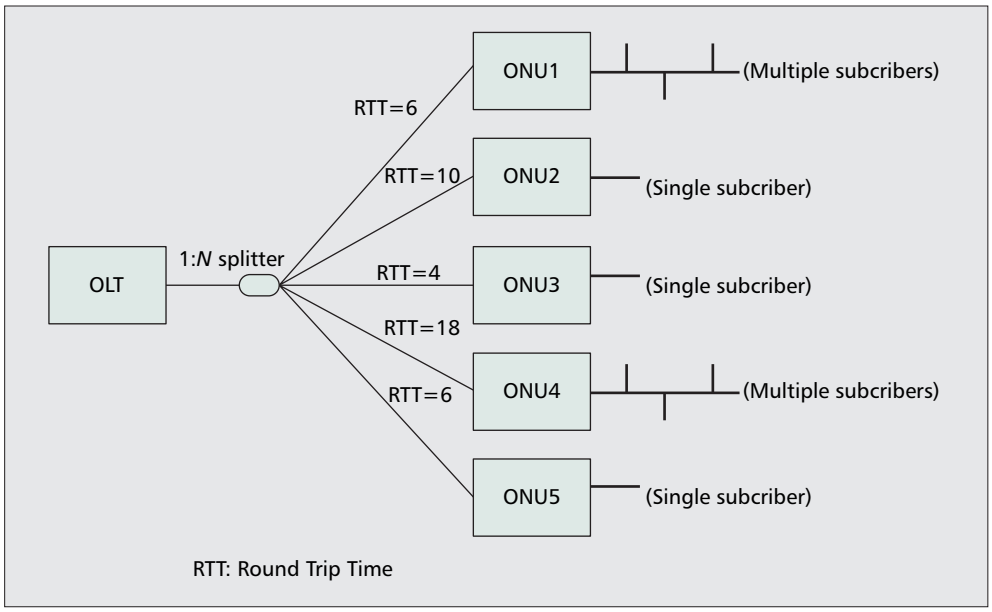


Figure 1. *Network architecture of a PON with one optical line terminal (OLT) and* N = *5 optical network units (ONUs), each with a different round-trip time (RTT).*

90 percent of data traffic originates and terminates in Ethernet frames, using an EPON can reduce the adaptation required to move data between the LAN and the access network. Furthermore, ATM creates inefficiencies in data transport as a result of its fixed data unit that requires most data packets to be segmented and reassembled at the end points of the network. This segmentation and reassembly results in higher processing delays, as well as reduced efficiency of error recovery techniques. For these reasons, EPONs appear to be more promising than APONs for data dominated networks. GPONs, on the other hand, by using GEM instead of ATM, avoid the inefficiency of segmentation and reassembly.

In this article, we review and classify the existing research on EPONs. The focus is on EPON architectures and dynamic bandwidth allocation (DBA) for EPONs, and our classifications provide insight into areas that are open for further investigation. For a survey on EPON security issues, which are not covered in this article, see [1]. This article provides a comprehensive and up-to-date EPON research survey as of spring 2007. The status and the main directions of this research as of early 2004 were presented in [2]. The EPON research area has been very active over the last few years, resulting in a dramatically expanded and more intricate body of EPON research. Therefore, a fundamentally new classification and survey of this area is required and provided in this article.

We review the standard PON architecture and two alternative architectures that were proposed. We review and classify all of the research done on the problem of DBA for EPONs. We classify this work in a meaningful way that provides insight to researchers currently working on EPONs and those considering working on EPONs. We discuss medium access control (MAC) protocols for the two alternative PON architectures. Finally, we conclude the article with a discussion of avenues of further investigation.

PON Architecture

A PON generally has a physical tree topology, where one optical line terminal (OLT) residing at the central office of the service provider connects to several optical network units (ONUs) in the field. The OLT is connected to the ONUs with a feeder fiber that is subsequently split using a 1 : *N* optical splitter/combiner to enable the ONUs to share the optical fiber. This is illustrated in Fig. 1. The transmission direction from OLT to ONU is referred to as *downstream* and operates as a broadcast medium. The transmission direction from the ONUs to the OLT is referred to as *upstream*. The upstream signals propagate from ONU to OLT but are not reflected back to each ONU; therefore, the PON is not a broadcast medium in the upstream direction. The EPON is a multipoint-to-point [3] medium, where the ONUs cannot detect each other's transmission because the upstream optical signal is not received by the ONUs. However, ONUs share the same fiber; hence, their transmissions can collide, and contention resolution must be performed.

To avoid collisions in the upstream direction, time division multiplexing (TDM) or wavelength division multiplexing (WDM) can be used [4]. WDM provides a large amount of bandwidth to each user, but requires that each ONU use a unique wavelength, which presents inventory challenges for service providers that must stock many different ONU types. TDM allows all ONUs to share a single wavelength, thus, reducing the number of transceivers at the OLT and allowing for a single ONU type. First generation PONs use wavelengths to separate the upstream and downstream channels but use TDM to avoid upstream transmission collisions between ONUs. Due to the topology of the PON, MAC protocols that rely on connectivity between all nodes cannot be utilized. A PON allows for connectivity from the OLT to all ONUs in the downstream and from each ONU to the OLT in the upstream (i.e., only the OLT has connectivity to all nodes). This connectivity pattern dictates the use of a centralized MAC protocol residing at the OLT. This leads to a polling-based MAC, where the OLT polls ONUs and grants them access to the shared PON medium.

Broadcast PON

An alternative PON architecture proposed in [5, 6] requires reflection of the upstream signal back to the ONUs, as illustrated in Fig. 2. Splitter 1 splits the upstream signal back to the ONUs, and splitter 2 splits the downstream signal as in the standard PON architecture. This creates a broadcast network that enables a decentralized medium access control protocol (e.g., carrier sense multiple access with collision

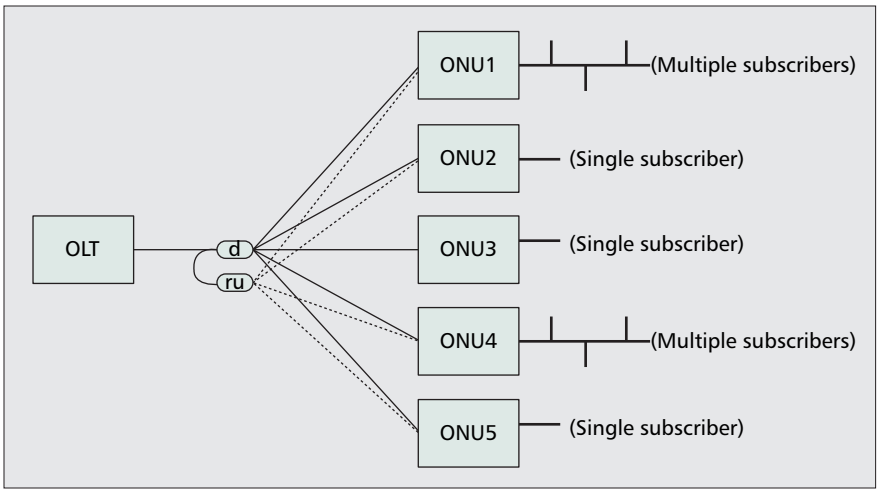


Figure 2. *Broadcast PON Architecture: Downstream OLT to ONUs transmissions are copied by splitter "d" to all ONUs, while each upstream ONU to OLT transmission is reflected by splitter "ru" back to all ONUs, thus creating a broadcast network for both upstream and downstream transmissions. The dashed lines represent the extra fibers used to carry the reflected upstream signal back to the ONUs.*

detection [CSMA/CD]) to be employed. Unfortunately, there are economic downsides to this approach. The ONUs become more expensive because they must:

- Contain higher power lasers to overcome the loss incurred by splitting their upstream signals to reflect back to other ONUs.
- Contain an extra receiver for the upstream wavelength.
- Have more intelligence to participate in the medium access arbitration.
- Have an extra fiber for the reflected upstream signal.

Further, the large bandwidth-propagation delay product of the optical access network limits the feasibility of this type of architecture.

Two-Stage PON

A two-stage PON architecture [7] can enable a PON to accommodate a higher number of ONUs compared to a single-stage PON. Two-stage PONs help to increase the reach of the PON. In the first stage, some ONUs act as sub-OLTs for other ONUs, as illustrated in Fig. 3. These sub-OLTs regenerate the optical signal in the upstream and downstream, as well as aggregate the traffic of their child ONUs. This allows a single OLT in a central office to potentially reach a larger number of ONUs because the sub-OLTs act as optical switches, mitigating optical power budget concerns that arise when increasing the number of ONUs.

Dynamic Bandwidth Allocation

DBA generally is defined as the process of providing statistical multiplexing among ONUs. To understand the importance of statistical multiplexing in PONs, note that the data traffic on the individual links in the access network is quite bursty. This is in contrast to metropolitan or wide area networks where the bandwidth requirements are relatively smooth due to the aggregation of many traffic sources. In an access network, each link represents a single or small set of subscribers with very bursty traffic due to a small number of bursty sources (e.g., Web data and packetized video). Because of this bursty traffic, the bandwidth requirements vary widely with time. Therefore, the static allocation of bandwidth to the individual subscribers (or sets of subscribers) in a PON is typically inefficient [8]. Statistical multiplexing that adapts to instantaneous bandwidth requirements is typically more efficient. The DBA that operates at the OLT is responsible for providing statistical multiplexing.

The OLT requires instantaneous bandwidth requirement information from each ONU to make access decisions. Having this precise information is not possible due to the non-zero propagation delays on a PON, typically up to 100 μsec, which is significantly larger than the transmission time of the maximum size Ethernet frame: 12.3 μsec. The ONUs must report their instantaneous queue sizes in a control frame and propagate this through the PON to the OLT.

A PON is a remote scheduling system [9] that suffers the following problems:

- Significant queue switchover overhead [9] (in the case of PONs, this is due to the guard times between ONU transmissions). Guard times between ONU transmissions are required to enable the previously transmitting ONU to power off its laser to prevent spurious transmission while the next ONU transmits; the next ONU to power on its laser in preparation for transmission; and the OLT to adjust its receiver to account for power-level differences in transmissions from different ONUs due to their

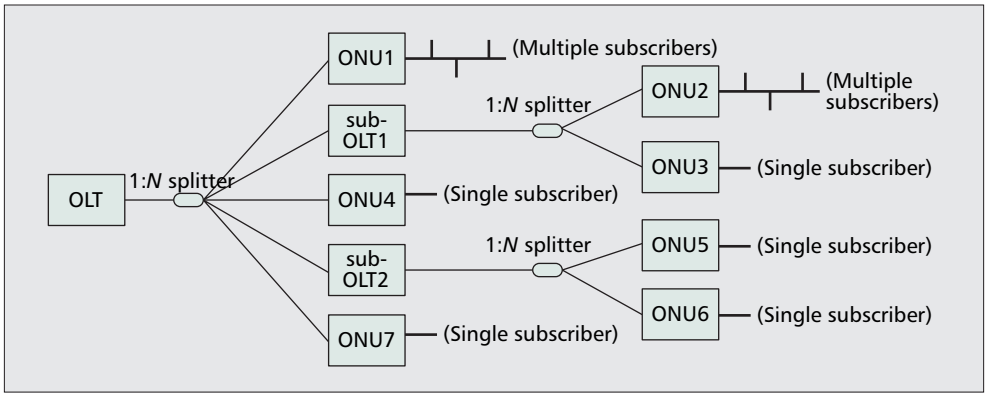


Figure 3. *Two-Stage PON Architecture: Certain ONUs act as sub-OLTs that regenerate the optical signal for ONUs in a second stage, thereby allowing for an increase in the total number of served ONUs.*

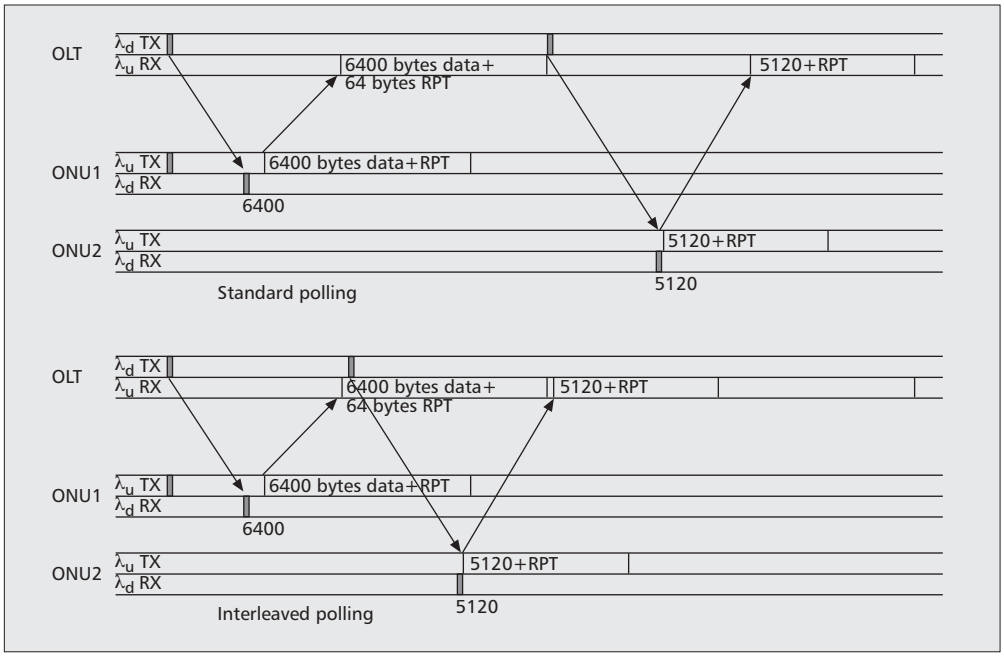


Figure 4. *With interleaved polling, the OLT can issue the grants on the downstream wavelength channel λ_d such that successive upstream transmissions on channel λ_u are separated in time by only a guard time interval compared to a round trip time with standard polling.*

different distances from the OLT. The queue switchover overhead should be mitigated by using a cyclic PON scheduler that issues one grant to an ONU per cycle.

- Large control plane propagation delay as a result of distances between OLT and ONUs. Interleaved polling is used to mitigate the large propagation delays on PONs [3]. With interleaved polling, the next ONU to be polled is issued a message giving transmission access while the previous ONU is still transmitting. This message, referred to as a *grant*, contains the start time of the transmission window, as well as the length (duration) of the transmission window. Figure 4 illustrates the difference between polling with and without interleaving.
- Limited control plane bandwidth, which is mitigated by short control plane messages.

A cyclic interleaved polling MAC called interleaved polling with adaptive cycle time (IPACT) [3] mitigates all of these issues.

The process of DBA consists of two parallel but potentially overlapping problems: grant sizing and grant scheduling (or inter-ONU scheduling). Grant sizing determines the size of a grant, that is, the length of the transmission window assigned to an ONU for a given grant cycle. Grant scheduling determines the order of ONU grants for a given cycle. Although the focus of this section is on DBA for EPONs, most of the results extend to other PONs as well. Specifically, the results that are not tied to the MultiPoint Control Protocol (MPCP) or Ethernet frame can be extended beyond EPONs to BPONs and GPONs.

Figure 5 shows our taxonomy for dynamic bandwidth allocation. We use this taxonomy as a framework for discussing

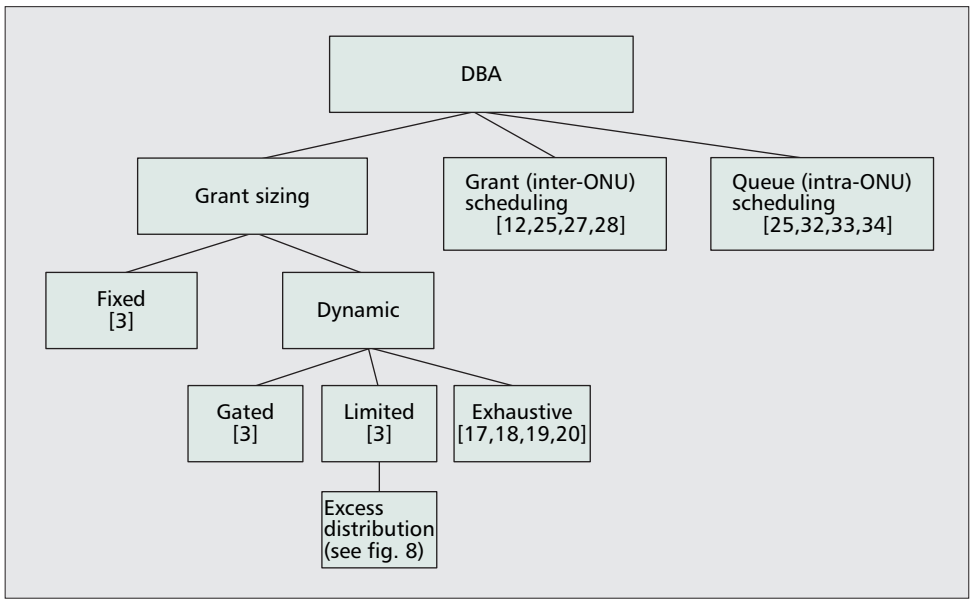


Figure 5. *Dynamic bandwidth allocation taxonomy.*

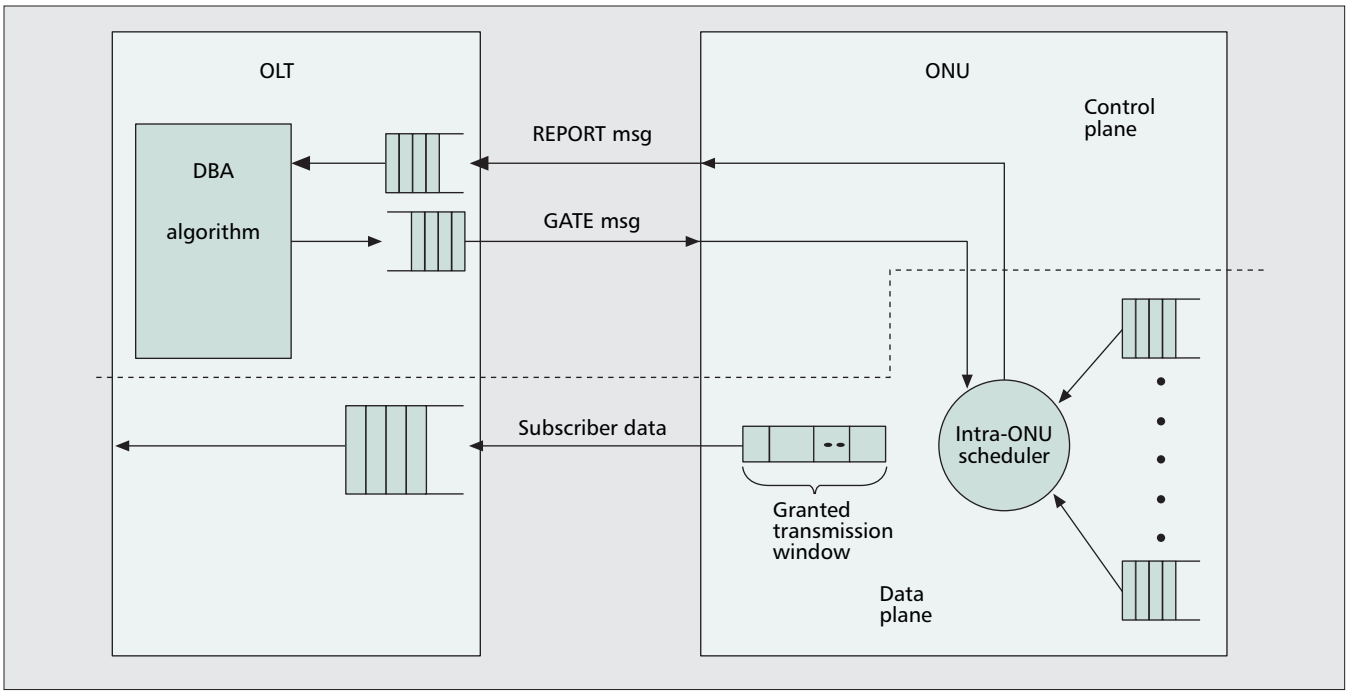


Figure 6. *MPCP operation: Two-way messaging assignment of time slots for upstream transmission between ONU and OLT.*

the research on dynamic bandwidth allocation for EPONs. We discuss the MPCP defined in the IEEE 802.3ah standard. This protocol defines the control plane used in EPONs to coordinate medium access. We discuss two fundamentally different problem-solving approaches to DBA; we call these approaches DBA frameworks. We discuss the research work done for the problem of grant sizing. We discuss the research on grant scheduling. We discuss intra-ONU scheduling, that is, arbitrating the different queues in a given ONU. We discuss the topic of quality of service (QoS), and we also discuss the issue of fairness, which is touched on later. All of these sections assume the standard PON architecture of Fig. 1. Discussions of protocols for broadcast and two-stage PON architectures are deferred to later.

MultiPoint Control Protocol

To facilitate the discovery and registration of ONUs, as well as medium access control, the IEEE 802.3ah task force designed the MPCP. The MPCP consists of five messages. REGISTER REQ, REGISTER, and REGISTER ACK are used for the discovery and registration of new ONUs. REPORT and GATE are used for facilitating centralized medium access control. The REPORT message is used to report the instantaneous queue occupancies at an ONU to the OLT. This REPORT message also can contain queue occupancies at certain threshold levels as opposed to only the full occupancy. This threshold queue reporting allows the OLT flexibility in determining the size of a granted transmission window. The GATE message is used by the OLT to grant non-overlapping transmission windows to the ONUs.

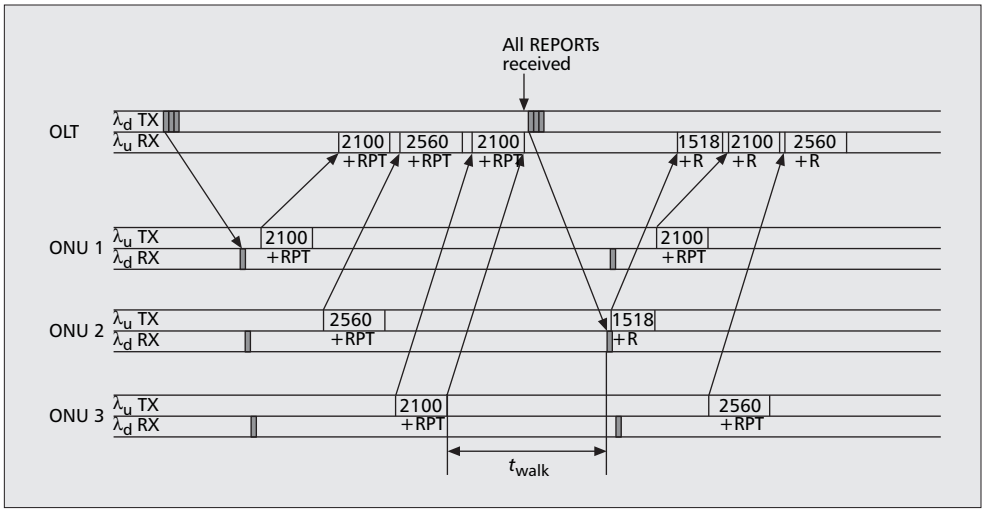
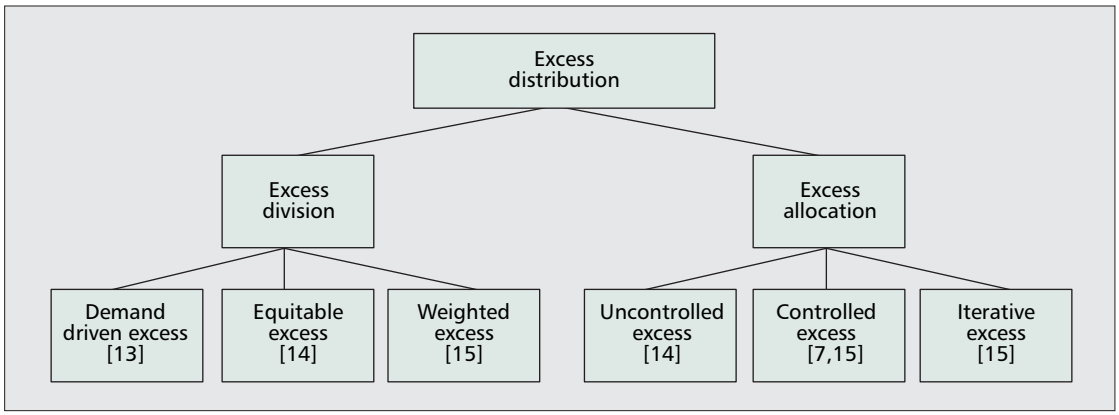


Figure 7. *In interleaved polling with stop, the OLT waits to receive the REPORT message from the last ONU in a cycle before polling the first ONU in the next cycle. This allows the OLT to make DBA decisions based on the REPORTs from all ONUs. As a result, a walk time of at least an RTT is incurred between grant cycles.*



■ **Figure 8.** *Excess distribution taxonomy.*

Figure 6 illustrates the infrastructure for facilitating DBA. REPORT messages flow upstream to report queue occupancies, and GATE messages flow downstream to grant upstream transmission windows. Upon receiving queue occupancy information by means of REPORT messages, the OLT — using a DBA algorithm — makes MAC decisions for the next cycle and communicates these decisions to the ONUs through GATE messages.

DBA Frameworks

With interleaved polling, the DBA granting cycles of the individual ONUs are interleaved, and the OLT typically makes grant decisions based on individual ONU REPORT messages. That is, the OLT typically does not wait until REPORT messages are received from all ONUs before making grant decisions. This prohibits the OLT from making DBA decisions that consider the bandwidth requirements of all ONUs. As a result, it is very difficult for the OLT to make fair access decisions. An alternative approach is the so-called interleaved polling with stop (Fig. 7) in which the OLT stops and waits between granting cycles for all ONU REPORT messages to be received before making DBA decisions. This affords the OLT the opportunity to provide a fair distribution of bandwidth. The trade-off is that interleaved polling with stop decreases the bandwidth utilization by forcing an idle period, t_{walk}, of the one-way propagation delay from the last polled ONU to the OLT plus the one-way propagation delay from the OLT to the first polled ONU, as illustrated in Fig. 7, that is, the walk time is typically equal to the average round trip time (RTT). Depending on the length of the granting cycle, this walk time can become a significant portion of the available bandwidth. For example, with a cycle length of 1.5 msec and an RTT on the order of 50 μsec, at least 3.33 percent of the available bandwidth is wasted on the walk time. For a 750 μsec cycle time, the walk times would consume 6.66 percent of the available bandwidth.

Interleaved polling and interleaved polling with stop also can be compared by their problem-solving approach to DBA. Interleaved polling without stop requires an online problem-solving approach to DBA, that is, the OLT makes DBA decisions with incomplete knowledge of the bandwidth requirements of all the ONUs. Whereas, interleaved polling with stop allows for an offline problem-solving approach to DBA, that is, the OLT makes DBA decisions with full knowledge of the bandwidth requirements of all the ONUs. We refer to these DBA problem-solving approaches as the online and offline DBA frameworks, respectively.

Grant Sizing

Grant sizing can be divided into four major categories:
- Gated
- Limited
- Limited with excess distribution
- Exhaustive using queue size prediction

Let G_i be the grant size for the current cycle for ONU i, R_i the queue size reported in the most recently received REPORT message from ONU i, $G_i^{\max}$ the limit on the maximum grant size for ONU i, E_i the share of the excess bandwidth in a cycle allocated to ONU i, and P_i the predicted queued traffic between the time of the REPORT transmission from ONU to OLT and the end of the granted transmission window to ONU i for the next cycle. The general equation for grant sizing would then be: $G_i = f(R_i, G_i^{\max}) + E_i + P_i$. We now discuss different techniques used for this general function $f(\cdot)$ and for determining E_i and P_i.

Fixed — In the fixed grant-sizing scheme, the grant size is fixed for an ONU every cycle. The function for G_i is simply $G_i = G_i^{\max}$. Simulation results [3] have shown that the fixed grant-sizing severely underperforms the dynamic grant-sizing techniques described below. An analysis in [8] confirms the simulation results.

Gated — In the gated grant-sizing technique, the grant size for an ONU is simply the queue size reported by that ONU, the function for G_i is $G_i = R_i$. This scheme provides low average delay but does not provide adequate control to ensure fair access between ONUs. In-depth delay analyses of the gated scheme can be found in [10] and [11].

Limited — In the limited grant-sizing technique [3], the grant size is set to the reported queue size up to a maximum grant size for that ONU. The function for G_i for the limited scheme is $G_i = \min(R_i, G_i^{\max})$. This grant-sizing scheme prevents any ONU from monopolizing the shared link. Simulation results [3] have shown that there is no average packet delay difference between gated and limited grant sizing. However, limited grant sizing can assist in providing fair access between ONUs by limiting the size of the grant to $G_i^{\max}$ and thereby preventing an ONU from monopolizing the link. Let t_{cycle} be the length of a grant cycle and t_{guard} be the guard time between grants. Under high traffic load, $t_{\text{cycle}}^{\max} = \Sigma_i(G_i^{\max} + t_{guard})$ (i.e., the maximum grant cycle length is defined by the maximum grant sizes). A large maximum grant cycle results in larger delays, whereas a short maximum grant cycle reduces the channel utilization due to increased guard times [12].

The limited scheme suffers from two detriments. First, the queue is underserved if $G_i^{\max} < R_i$. Second, i bandwidth is

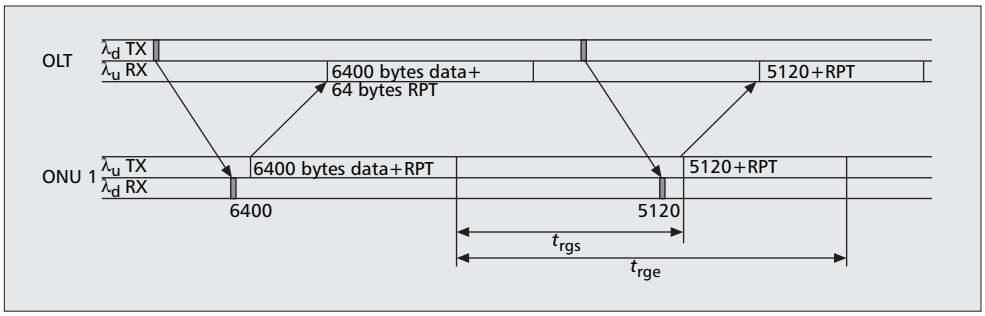


Figure 9. *Illustration of queue waiting times.*

wasted if the request is not fully satisfied, and the end of the grant does not accommodate the next Ethernet frame. In the best case, this head-of-line (HOL) Ethernet frame that does not fit into the remainder of the grant is only 64 bytes, which results in up to 63 bytes of wasted grant. The worst case is if the HOL packet is the maximum length 1518 bytes. This results in up to 1517 bytes of wasted grant.

Limited with Excess Distribution — Limited with excess distribution [13] augments the limited grant-sizing scheme to improve statistical multiplexing while still preventing ONUs from monopolizing the link. In general, ONUs are partitioned into two groups: underloaded ONUs and overloaded ONUs. Underloaded ONUs are those whose REPORTed queue size is less than or equal to the maximum grant size, that is, $R_i \le G_i^{\max}$. Overloaded ONUs are those whose REPORTed queue size is larger than their maximum grant size, that is, $R_i > G_i^{\max}$. In the excess distribution schemes, the overloaded ONUs share the unused or excess bandwidth left over from underloaded ONUs. The grant for an overloaded ONU then becomes $G_i = G_i^{\max} + E_i$. The total excess bandwidth is defined i to be the sum of the differences between the maximum grant size and the REPORTed queue size of all the underloaded ONUs. Let E_{total} be the total excess bandwidth for a cycle, $\mathcal{U}$ the set of underloaded ONUs, $\mathcal{O}$ the set of overloaded ONUs, and E_i the excess assigned to overloaded ONU i. The total excess is calculated as $E_{\text{total}} = \Sigma_{i\in\mathcal{U}} (G_i^{\max} - R_i)$. The computation of the total excess bandwidth requires the OLT to wait for all ONU REPORT messages, that is, requires the use of interleaved polling with stop or the offline DBA framework. A hybrid DBA framework that allows underloaded ONUs to be granted before the stop and overloaded ONUs to be granted after the stop [13] can help to mitigate the inefficiencies of the offline DBA framework. After it is computed, the E_{total} is divided between all the overloaded ONUs.

Excess distribution can be divided into excess division and excess allocation. Excess division divides E_{total} among the overloaded ONUs and excess allocation can, if used efficiently, redistribute excess credits that are unused by some overloaded ONUs. Figure 8 shows the taxonomy of excess distribution schemes (including excess division and allocation).

One approach to excess division (referred to as DBA1 in [13]) divides the excess according to demand, that is, DBA1 divides the excess according to relative request size following the formula:

$$E_i = \frac{R_i}{\Sigma_{j\in\mathcal{O}} R_j} \cdot E_{\text{total}}$$

We refer to this approach as demand-driven excess (DDE) division. Because each ONU's share of the excess is completely determined by its request size, the larger the ONU request size relative to the other ONUs, the more excess bandwidth it receives. This provides statistical multiplexing but is not necessarily fair. The fair excess [14] or equitable excess (EE) division method divides E_{total} equally among the overloaded ONUs. Let M be the total number of overloaded ONUs, EE divides the excess according to the formula:

$$E_i = \frac{1}{M} \cdot E_{\text{total}}.$$

This approach gives all ONUs an equal piece of the total excess, implying fairness. The weighted excess (WE) division method [15] uses ONU priority weights to divide the excess bandwidth. The total excess is divided among overloaded ONUs according to their weights:

$$E_i = \frac{w_i}{\sum_{j\in\mathcal{O}} w_j} \cdot E_{\text{total}}.$$

This method allows the ONUs to have differing priorities with respect to their fair share of the excess bandwidth.

After the excess is divided among the overloaded ONUs — according to DDE, EE, or WE — it is possible for the size of the grant to be larger than the request, that is, $G_i + E_i > R_i$, which results in wasted bandwidth. The excess division can be augmented by an excess allocation algorithm that sizes the grant so that it does not exceed the request, as well as redistributes unused excess credits.

The uncontrolled excess (UE) allocation method [14] assigns overloaded ONUs their share of the excess without regard to their request size. Therefore, bandwidth is wasted as some overloaded ONUs receive grants that are larger than their requests. The controlled excess (CE) allocation method [7, 15] provides better utilization by avoiding wasted bandwidth under certain conditions. Specifically, if the total excess demand from the overloaded ONUs $E_{\text{demand}} = \Sigma_{j\in\mathcal{O}} (R_j - G_j^{\max})$ is less than the total excess E_{total}, then each ONU is granted its full request R_i. This avoids wasted bandwidth for the situation when the total request does not exceed the maximum grant cycle size $\Sigma_i G_i^{\max}$.

The iterative excess (IE) allocation method [15] follows an iterative grant sizing approach to maximize the number of satisfied overloaded ONUs. To avoid sizing a grant larger than the request, the grant size for each ONU is determined as:

$$G_i = \begin{cases} R_i & : \text{ if } R_i \le G_i^{\max} + E_i \\ G_i^{\max} + E_i & : \text{ if } R_i > G_i^{\max} + E_i \end{cases}. \quad (1)$$

Initially, all of the overloaded ONUs are in a list. E_i is computed for each overloaded ONU according to one of the excess division methods. One by one an overloaded ONU's grant size is computed according to the above formula. If an

ONU is satisfied (i.e., $R_i \le G_i^{max} + E_i$), it is issued a grant and removed from the list. Unsatisfied ONUs remain in the list and participate in future iterations. After an iteration through the list, E_{total} is recomputed by removing the excess used by the satisfied ONUs. This allows the excess bandwidth unused by those satisfied ONUs to be made available to the unsatisfied overloaded ONUs. E_i is recomputed for the overloaded ONUs that remain in the list, and another iteration takes place. The iterations continue until there are no satisfied ONUs. On this final iteration, the unsatisfied ONUs simply are allocated their fair share of the total remaining excess (i.e., E_i). The iterative grant-sizing approach mitigates wasted bandwidth by not over-assigning bandwidth to ONUs and maximizes the number of satisfied ONUs to lower the amount of unused slot remainder. Further, it provides a more efficient distribution of the excess bandwidth.

Exhaustive Service System Using Queue Size Prediction — Queue size prediction is concerned with estimating the traffic that is generated during the period between the REPORT message transmission by the ONU and the beginning of the gated transmission window. Let t_{rgs} denote this time between the REPORT transmission at the ONU and the start of the next grant for that ONU. A service system that accommodates the traffic included in t_{rgs} is referred to as a partially gated service system [16]. Alternatively, additionally the traffic generated during the granted transmission window could be predicted. Let t_{rge} denote the time between the REPORT transmission by the ONU and the end of the next grant at that ONU. This results in an exhaustive service system [16]. Figure 9 illustrates these time periods. Let Q_i be the actual amount of traffic that queued up during t_{rgs} or t_{rge}. The goal of queue size prediction is to get P_i as close to Q_i as possible. For constant bit rate (CBR) traffic, which has a constant and therefore predictable rate of traffic generation, this is rather simple. Multiplying the constant rate of the CBR traffic in bits/sec by t_{rge} [13, 17] is a sufficient predictor for CBR traffic. For bursty variable bit rate (VBR) traffic, the queue size prediction is more challenging.

Elementary schemes for queue size prediction for bursty sources are [3]: constant credit and linear credit. In the constant credit scheme, the OLT adds some constant credit to the grant size. Let ψ be this credit, then $P_i = \psi$. In the linear credit scheme, the credit adapts to the size of the request. Let γ be the fraction of the request used as the credit, that is, $P_i = \gamma R_i$. The idea is that the request size gives some indication as to how much traffic will arrive in the waiting period, that is, t_{rgs} or t_{rge}.

Using control theory [18] to drive the gap between predicted and actual queued traffic to zero, an ONU reports the difference between the grant, G_i, and the actual data queued at the start of the granted transmission window. Let G_i^{prev} be the granted transmission window size, R_i^{prev} be the data queued at the time of the report, and Q_i^{prev} be the data queued during t_{rgs}, all for the previous grant cycle. Let δ_i be the difference reported by the ONU, that is, $\delta_i = G_i^{prev} - (R_i^{prev} + Q_i^{prev})$. Let α be a control gain parameter, then: $G_i = G_i^{prev} - \alpha\delta_i$.

Control theoretic approaches are used for modifying the control gain parameter to stabilize δ_i. Simulation results [18] show that this scheme has almost an order of magnitude lower packet delay compared to IPACT with gated grant sizing. This difference is attributed to much better queue size prediction. The results provide limited insight because they did not explicitly show that this control theoretic approach drove δ_i closer to zero than gated grant sizing.

A simple one step back linear predictor [19] can be used for prediction, that is, predictions are based on the actual data received during the previous waiting period. Let t_n^{prev} be the time of the previous cycle; the formula used for prediction is,

$$P_i = \frac{t_{rgs}}{t_{cycle}^{prev}} \cdot R_i.$$

This formula is identical to the linear credit scheme with

$$\gamma = \frac{t_{rgs}}{t_{cycle}^{prev}}.$$

Simulation results show that a one step back linear predictor provides lower packet delay for expedited forwarding traffic compared to fixed bandwidth allocation and limited bandwidth allocation.

A higher order linear predictor [20] for predicting traffic during the waiting period at an ONU also could be used. This linear predictor has its weights updated by means of the Least Mean Square (LMS) algorithm (see Section 3.5 in [21]). The linear predictor attempts to predict Q_i using information about a number, L, of previous Q_i values. Because it was shown in [22] that prediction of self-similar traffic is best performed using short-term correlation rather than long-term correlation, simulations were conducted using $L = 4$. An alternative approach could use this value as a starting point of a search to find the optimal value of L. Mathematical analysis proves that an increase in the accuracy of the predictor leads to a decrease in average delay.

In the simulation results presented in [20], the higher order linear traffic prediction scheme is compared against fixed grant sizing, limited grant sizing, and limited grant sizing with excess distribution. In limited grant sizing with excess distribution, the underloaded ONUs are scheduled immediately upon receipt of the REPORT, whereas overloaded ONUs are scheduled after REPORTs are received from all ONUs. The results show a reduction of average packet delay, as well as loss probability. However, it is unclear how much of the difference is attributable to traffic prediction and how much to the difference in scheduling, either waiting for all REPORTs or immediately scheduling each ONU.

Summary — The gated grant-sizing technique is the simplest, adapts to changing traffic demands, and eliminates wasted portions of a grant. However, it does not provide partially gated or exhaustive service, and it cannot prevent ONUs from monopolizing the upstream channel. This lack of control limits the ability to provide fair access and quality of service guarantees. The limited grant-sizing scheme can place limits on an ONU's access to the medium but can be inefficient in utilizing the upstream capacity. Further, the limited and any other non-gated grant-sizing scheme open up the possibility of a wasted portion of a grant due to lack of frame segmentation in Ethernet. When using a fixed-size protocol data unit (PDU), such as for a BPON, this is not a problem. Employing excess distribution can improve the efficiency of the limited scheme by allowing overloaded ONUs to take advantage of the bandwidth not used by underloaded ONUs in a granting cycle.

Queue size prediction can lower queuing delays by attempting to predict the precise instantaneous traffic demands of an ONU to allow for an exhaustive service discipline. The risk is in reduced throughput due to wasted portions of a grant that result from imprecise prediction. The bursty nature of data traffic in the access network makes precise prediction difficult. A slight over prediction will reduce throughput but decrease delay by at least accommodating all of the queued Ethernet

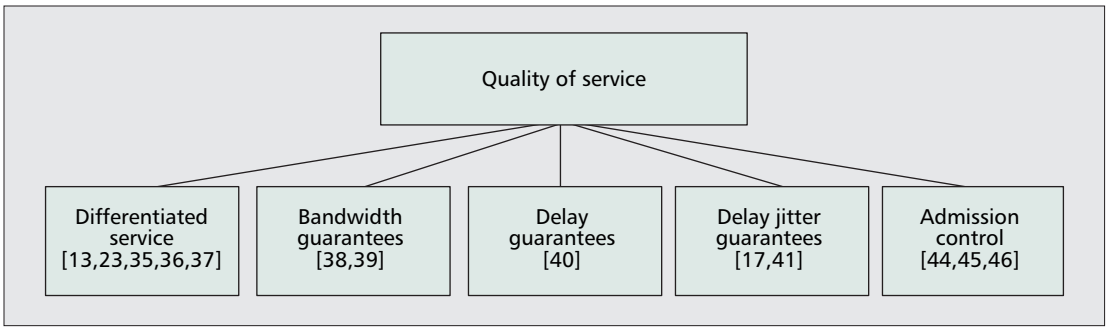


Figure 10. *Quality of service taxonomy.*

frames. Investigating non-linear neuro-computational methods for the queue size prediction problem potentially is an area for further research.

Grant Scheduling (Inter-ONU Scheduling)

Medium access control for an EPON contains two scheduling problems. The first is concerned with scheduling grants to each ONU, namely, inter-ONU scheduling. The second is concerned with scheduling the individual queues of Ethernet frames at the ONU for transmission within the granted transmission window, that is, intra-ONU scheduling. This division of the scheduling often is referred to as hierarchical scheduling [9, 23, 24]. We discuss inter-ONU or grant scheduling in this section and defer discussion of intra-ONU scheduling to the next section.

Since grant scheduling works at the inter-ONU level and is coupled with the process of grant sizing, it is performed at the OLT. Typically, to change the scheduling order from round robin, the OLT must wait for all REPORTed queue sizes from the ONUs and then determine the grant order. This requires the use of interleaved polling with stop or the offline DBA framework, as illustrated in Fig. 7.

ONU transmissions ordered longest queue first (LQF) [12, 25] or earliest packet first (EPF) [12] have been examined. LQF allows ONUs with the largest grant size to transmit first. This is the same as largest processing time (LPT) first scheduling in standard scheduling theory notation [26]. EPF allows ONUs with the earliest arriving HOL Ethernet frame to transmit first. To avoid the walk time between cycles when the OLT waits for all ONUs to REPORT before making any scheduling decisions, the scheduling can leave out the last or last few ONUs when scheduling. Simulation results [25, 27] using Poisson traffic show that both LQF and EPF provide lower average delay at medium loads compared to a round robin scheduler. At low and high loads, the average delay was the same as a round robin scheduler.

Implementing the limited with excess distribution grant sizing technique requires special consideration with respect to grant scheduling. To accumulate and distribute the total excess bandwidth, the OLT must wait to receive REPORT messages from all ONUs before issuing grants for the next cycle. This means a walk time, t_{walk}, is wasted between cycles. To mitigate this, lightly loaded ONUs can be scheduled immediately upon receiving their REPORT messages because they will not receive any excess [13]. However, as the traffic load increases to the point where all ONUs become overloaded, t_{walk} still is wasted between cycles. To avoid wasting t_{walk} in between cycles, an overloaded ONU can be scheduled immediately if there is no underloaded ONU available for scheduling when the channel becomes available [28].

Summary — Examining the performance of these existing scheduling schemes for self-similar traffic is an important topic for future research. Another avenue for future research is to employ scheduling theory to find a better scheduler.

ONU Queue Scheduling (Intra-ONU Scheduling)

Intra-ONU scheduling is concerned with scheduling the multiple queues of Ethernet frames at an ONU, for transmission within the ONU's granted transmission window. If the number of queues in an ONU is relatively small, this intra-ONU scheduling can be performed at the OLT. However, as the number of queues increases, scheduling is typically made hierarchical [9] with the inter-ONU scheduling at the root of the hierarchy in the OLT and one level of branches. The ONU contains the branch (i.e., intra-ONU) schedulers.

Low complexity is a key design goal for intra-ONU schedulers so that the cost of the ONUs is kept at a minimum. There are typically two classes of scheduling that service multiple queues of differing priority:

- Strict priority (SP) scheduling, which can be unfair
- Weighted fair queuing (WFQ) scheduling

SP scheduling creates unfairness when starving lower priority traffic due to unrestricted preemption. The ideal scheduler should allow statistical multiplexing, but guarantee a minimal portion of the available bandwidth to each priority queue (i.e., provide link sharing). Generalized processor sharing (GPS) [29] achieves these goals for the fluid traffic model, where packets are infinitesimally small. Unfortunately, in practical systems with finite-size packets, the ideal GPS link sharing is not directly applicable because a packet must monopolize the server (i.e., transmission link) while in service. WFQ [30] is a packet approximation of GPS whose deviation from the ideal case is bounded by the maximum packet size. WFQ calculates the start time of a packet under the ideal GPS system and based on this start time, computes the finish time under ideal GPS. Then, packets are transmitted in the order of the calculated finish time. The calculations of the ideal GPS times can be computationally intensive for ONUs. A few schemes were proposed to simplify these calculations at the expense of approximation accuracy to the ideal GPS.

Start-time fair queuing (SFQ) [31] is one simplified version of WFQ. In SFQ, the calculation of server virtual time that is used to calculate the start time of a packet is reduced to the start time of the packet currently in service, greatly reducing the computational complexity. In contrast to WFQ, SFQ sorts packets by start time rather than finish time. An intra-ONU frame scheduler that employs a modified start-time fair queuing (M-SFQ) algorithm [32, 33] can be used as a low complexity alternative to SFQ. M-SFQ further simplifies the scheduling by calculating the start time only for the HOL packets.

Simulation results comparing M-SFQ to strict priority scheduling indicate that M-SFQ provides the same average delay for the expedited forwarding class (delay sensitive traffic), higher average delay for assured services class 1 (high-

speed video), and lower average delay for assured services class 2 (low-speed pre-recorded video). Further, the average delay for class 2 is lower than for class 1, which seems undesirable, but is not commented on in the simulation study. One can conclude that M-SFQ provides improved delay and throughput for assured services class 2 at the expense of worse performance of assured services class 1. It is not clear how this displays the strengths of M-SFQ over strict priority scheduling. M-SFQ achieves better inter-class isolation [33] but also treats class 2 traffic better than class 1 traffic, which is undesirable.

An additional shortcoming of M-SFQ is that it tends to starve best-effort traffic to provide better QoS to the assured services and expedited forwarding classes [34]. A modified version of deficit weighted round robin (M-DWRR) was proposed and examined in [34] to address this shortcoming. M-DWRR maintains a credit deficit counter for each class and considers only the HOL packets, ensuring low computational complexity. In a first scheduling pass, M-DWRR offers bandwidth to each class queue according to the class weight (which is reflected in the deficit credit counter). The bandwidth that is not required by the individual queues is redistributed in a second scheduling pass, which has some resemblance to the excess distribution mentioned earlier, but is conducted internally by the ONU. Simulation results indicate that M-DWRR ensures fairness according to the chosen weights for the different service classes, including best-effort traffic. Also, overall throughput with M-DWRR is about 10 percent higher than with M-SFQ due to eliminating best-effort traffic starvation.

A non-work-conserving scheduling discipline, called priority with insertion scheduling (PIS) [25], that transmits real-time packets when their delay-bound will be exceeded is yet another ONU queue scheduling approach. PIS allows non-real-time traffic to gain access to the medium, as long as the real-time traffic can be delayed without detriment.

Summary — To keep ONU costs low, the complexity of the ONUs should be kept low. Therefore, the ideal intra-ONU scheduler provides quality of service guarantees through link sharing with low complexity. Alternatively, the intra-ONU scheduling can be performed at the OLT. This can result in potential scalability problems as the number of queues increases [9]. However, allowing the OLT to perform the ONU queue scheduling reduces ONU complexity concerns.

Quality of Service

EPONs are intended not only to carry best-effort data traffic, they also are expected to carry packetized voice and video that have strict bandwidth and delay requirements, as well as delay jitter sensitivity. We present in Fig. 10 the taxonomy for organizing the research work on quality of service guarantees for EPONs.

Differentiated Service — The simplest way to facilitate QoS is to provide differentiation of traffic and different service to each differentiated traffic class. The ONUs classify and separately buffer ingress traffic and can perform strict priority scheduling between the classes when deciding which frames to send during a gated transmission window. The use of strict priority scheduling is required for compliance with 802.1d bridging [23]. Standard strict priority scheduling results in a phenomenon referred to as the light load penalty [13, 23]. The individual queue sizes are REPORTed at the end of a grant. During the period t_{rge}, more high-priority traffic can arrive at the ONU, which can preempt the lower-priority traffic that was accounted for in the REPORT. If the grant sizing predicts this higher priority traffic, newly arriving during t_{rge}, then the grant will accommodate this traffic; otherwise, it will unfairly preempt the lower priority traffic that was accounted for in the REPORT. This problem occurs at low loads when the grants are typically small and are more likely used up by the newly arriving high-priority traffic. To alleviate this problem, a two-stage buffering scheme [23] should be used at the ONU. The two-stage buffering moves the frames that were accounted for in the REPORT scheduled into a single second-stage queue that is emptied first during the next grant. This scheme effectively enforces a strict priority scheduling performed at REPORT time, as opposed to the time of the grant, which will have a queue occupancy that is potentially different than that REPORTed. Simulation results [13, 23] demonstrate the existence of the light-load penalty and indicate that the two-stage buffering eliminates the light-load penalty at the expense of higher delay of high-priority traffic.

Strict priority scheduling can be extended to the PON level [35]. A DBA algorithm that uses a fixed-cycle length and divides this cycle between three priority classes on a strict priority basis is one approach. The OLT would send a separate grant for each priority class [13, 35].

A two-layer DBA (TLBA) scheme [36] for differentiated services also could be used. In the first layer, the OLT decides the cycle partitioning between the classes of service (class-layer allocation), and in the second layer, the partition for each class is further divided for each ONU (ONU-layer allocation). Within a class, all ONUs share the bandwidth according to a max-min fairness policy. To keep any class from monopolizing the available bandwidth in a frame, a per-class bandwidth threshold is enforced. The bandwidth threshold guarantees a minimum bandwidth for a class under heavy load. Any remaining bandwidth from classes that request less than their threshold is divided among classes that request more than their threshold. The division of that remaining bandwidth is handled through weights that are assigned to each class. For ONU buffer management, weighted random early detection (RED) can be used. Simulation results [36] indicate that TLBA can divide the bandwidth as set through class weights when under heavy load. This allows for bandwidth guarantees for each class. The results also show that under lower loads, TLBA allows for effective utilization of the medium.

Class-of-service oriented packet scheduling (COPS) [37] is another method to provide differentiated services. COPS regulates the traffic of each ONU, as well as each class-of-service (CoS) using two sets of credit pools, one per ONU and one per CoS. Granting begins with the highest CoS and ends with the lowest CoS. In the first round of granting, each ONU with traffic for the current CoS is granted up to the number of credits stored for that ONU, as well as that CoS. To mitigate the unused slot remainder for those grants that cannot be fully satisfied, a threshold queue-reporting scheme is used. At the end of the first round, the unused credits are pooled together and in the second round, these unused credits are distributed to the CoS-ONU pairs that were not fully satisfied. Simulation results [37] show that COPS can provide lower average and maximum delay for all but the highest CoS as compared to IPACT with limited grant sizing (IPACT-LS). The highest CoS experiences slightly higher average delay under COPS as compared to IPACT-LS.

All of the above schemes suggest differentiating traffic at the ONU into classes, separately reporting queue sizes of each class, and allowing the OLT to provide individual grants to each class. The schemes differ in how they determine the grant sizes for each class. It is also apparent that two-stage buffering is required to keep higher priority traffic arriving

during t_{rge} from preempting the lower priority traffic that was accounted for in the REPORT.

Bandwidth Guarantees — A DBA algorithm for EPONs called Bandwidth Guaranteed Polling (BGP) [38] can be used for providing bandwidth guarantees. The BGP algorithm acts as a compromise between fixed TDM and statistical multiplexing. In BGP, ONUs are divided into two disjoint sets:
- Bandwidth guaranteed ONUs
- Non-bandwidth guaranteed ONUs

The algorithm maintains two polling tables.

The first polling table divides a fixed length polling cycle into a number of bandwidth units. The required bandwidth of an ONU, as dictated by a service level agreement (SLA) with a service provider, determines the number of bandwidth units allocated in the polling table to that ONU. A bandwidth-guaranteed ONU with more than one entry in the polling table has these entries spreading through the table rather than appearing contiguously. This lowers the average queuing delay because these ONUs are polled more frequently. However, the increased polling frequency results in more grants per cycle and hence, more guard times between grants, and thus lower channel utilization. Further, fragmenting the grants can potentially lead to lower grant utilization because Ethernet frames cannot be fragmented to be transmitted across grant boundaries. Therefore, a frame that is too large to fit in the remainder of a bandwidth unit must wait for the next bandwidth unit, and a portion of the current bandwidth unit is wasted. Lower grant utilization further reduces the channel utilization. BGP employs a method to mitigate the grant utilization problem by allowing an ONU to communicate its actual use of a bandwidth unit to an OLT. If the unused portion of the bandwidth unit is sufficiently large, this portion is granted to a non-bandwidth guaranteed ONU. Otherwise, the next bandwidth-guaranteed ONU is polled. However, this approach is severely limited by the propagation delays (i.e., walk times) required for message exchange on an EPON.

In BGP, unused bandwidth units in the first polling table are given to non-bandwidth-guaranteed ONUs in their order of appearance in the second polling table. The second polling table, is constructed differently than the first. Entries in the second polling table are created as non-bandwidth-guaranteed ONUs request grants. This is in contrast to the first polling table, which represents a division of time on the upstream channel of the EPON. Simulation results presented in [38] show that, as one expects, ONUs with more entries in the polling table have lower queuing delay than those with fewer entries, and IPACT lies somewhere in the middle.

BGP can be augmented to include admission control for new bandwidth-guaranteed ONUs [39]. This admission control uses standard parameter-based admission control. There are two parameters that describe the resource requirements of ONUs:
- Bandwidth requirement (peak rate)
- Delay bound requirement

Using these parameters, the admission control determines whether the ONU is accepted as a bandwidth-guaranteed (BG) ONU or as a best-effort (non-BG) ONU.

Delay Guarantees — A DBA algorithm called Dual DEB-GPS Scheduler [40] potentially can help provide delay guarantees. This DBA uses deterministic effective bandwidth (DEB), to determine the scheduling weights used in a generalized processor sharing (GPS) scheduler. The scheduling is performed in two layers, hence the name Dual. The first layer performs class-level multiplexing at the OLT. The second layer performs source level multiplexing at the ONU.

Traffic arriving at the ONU is regulated by a leaky bucket mechanism. This leaky bucket enforces a traffic profile characterized by: burst size Ω, peak rate Φ, and average rate μ. These leaky bucket parameters are used to determine the DEB for a source. This DEB is directly used as a weight to determine the portion of the upstream bandwidth assigned to this traffic source. The DEB guarantees a particular delay bound. Let Δ be the desired delay bound, and $B_{\text{eff}}(\Delta)$ be the effective bandwidth to guarantee the delay bound. Then,

$$B_{\text{eff}}(\Delta) = \begin{cases} \dfrac{\Omega}{\left(\Delta + \dfrac{\Omega}{\Phi}\right)} & : \text{ if } 0 \le \Delta \le \Omega\left(\dfrac{1}{\mu} - \dfrac{1}{\Phi}\right) \\ \mu & : \text{ if } \Delta > \Omega\left(\dfrac{1}{\mu} - \dfrac{1}{\Phi}\right). \end{cases} \quad (2)$$

The OLT divides each bandwidth cycle according to the $B_{\text{eff}}(\Delta)$ weights; the remaining bandwidth in a cycle is divided equally among all best-effort sources (i.e., those that do not require any delay bounds). The OLT generates the grants per ONU based on the weights and the information about which sources belong to each ONU. The ONU is then responsible for performing the intra-ONU scheduling that determines how the different sources fill the granted transmission window. DEB attempts to guarantee delay bounds by guaranteeing a certain bandwidth (i.e., B_{eff}).

Simulation results are presented [40] with ONUs having two QoS-aware traffic sources (QoS1 and QoS2) and two best-effort sources (BE1 and BE2). QoS1 has a delay bound of 1 msec and jitter bound of 0.2 msec. QoS2 has a delay bound of 2 msec and jitter bound of 0.4 msec. According to the presented results, QoS1 traffic experiences roughly a 2 to 2.5-millisecond delay at a load of 0.7 and higher (bound is 1 millisecond). QoS2 traffic experiences roughly a 3-millisecond delay at a load of 0.7 and higher (bound is 2 milliseconds). From the presented results, it seems Dual DEB-GPS provides low delay and jitter for traffic sources that require QoS but does not provide a guaranteed delay bound at higher loads.

Delay Jitter Guarantees — A scheme called the Hybrid Slot Size/Rate algorithm (HSSR) [41] can stabilize packet delay variation in EPONs for jitter-sensitive high-priority traffic. HSSR not only uses a fixed cycle length but also fixes the position of jitter-sensitive high-priority traffic grants (fixed to the beginning of the frame). The lower priority traffic from ONUs occupies the remainder of the frame. HSSR causes more than one grant per cycle to an ONU, which reduces efficiency due to extra guard times. However, the reduced efficiency allows for guaranteeing packet-delay variation bounds for certain traffic. A portion of the fixed grant cycle is partitioned for jitter-sensitive traffic. Quasi non-intrusive ranging keeps the ranging and registration process of new ONUs from disturbing high-priority traffic. The ranging and registration responses from new ONUs are scheduled to occur during the best-effort portion of the fixed frame. The fixed frame is large enough to provide ample time for this process.

Simulation results show that HSSR offers lower average delay and packet-delay variation than the conventional scheme (i.e., no fixed position of high-priority traffic). Further, the packet-delay variation for HSSR is due solely to queuing delay and not from the scheduling by the DBA.

The Hybrid Granting Protocol (HGP) [17] can ensure QoS through minimizing jitter and guaranteeing bandwidth. It is a hybrid of two approaches to sizing grants, one uses the

REPORT message to size the grant; the other uses a queue prediction mechanism to size the grant to accommodate all queued traffic at the point the grant begins (i.e., accommodates traffic in t_{rgs}). The first approach is used for assured forwarding (AF) [42] and best-effort (BE) services, whereas the latter approach is used for expedited-forwarding (EF) [43] services that are assumed to have a constant bit rate and therefore can be estimated easily.

A scheduling cycle is then divided into two subcycles: EF subcycle and AF/BE subcycle. The EF subcycle carries the EF services for each ONU and the AF/BE subcycle carries the AF and BE services for each ONU. Hence, every scheduling cycle there are two grants for each ONU. REPORTing of the AF and BE queues in an ONU is delayed until the end of the EF grant for that ONU. This allows the OLT to obtain more up-to-date queue occupancy for the ONU because the DBA computation is performed after the EF subcycle.

By fixing the scheduling cycle size and fixing the position of the EF grant to each ONU, HGP can guarantee bandwidth to EF traffic and minimize the jitter experienced by the EF traffic. This QoS comes at the expense of more guard times per cycle because of the separate grant to each ONU for its EF traffic. To mitigate this inefficiency, a grant for AF/BE traffic is not sent if there is no pending traffic, eliminating the need for a guard time for the AF/BE grant.

HGP and HSSR [41] share the same frame structure and division. The novelty of HGP is the way in which the REPORTs are generated. Simulation results [17] show that HGP provides lower queuing delay at higher loads as compared to a regular EPON scheduler. At lower loads, the regular EPON scheduler provides lower queuing delay, which is attributed to the increased number of guard times per cycle with HGP.

Admission Control — As reviewed, a variety of inter- and intra-ONU scheduling solutions exist to provide QoS in EPON networks. These solutions should be effective not only in supporting QoS but also must be designed carefully to protect the requirements of already admitted traffic, as specified in the associated SLAs. Toward this end, admission control may become necessary to both support and protect the QoS requirements in EPON networks.

Research on admission control for EPON began only very recently [44, 45]. An admission control framework together with an appropriate DBA algorithm that is capable of supporting and protecting QoS of real-time traffic while guaranteeing a minimum bandwidth for best-effort traffic was introduced and studied in [46]. The examined admission control algorithm determines whether or not to admit a new real-time traffic stream based on its requirements and the utilization of the upstream wavelength channel. To achieve this, each polling cycle is divided into two subcycles. In the first subcycle, each ONU is assigned a guaranteed minimum bandwidth to support the respective QoS requirements of its streams. The second subcycle is used by the OLT to dynamically assign transmission windows to best-effort traffic of all ONUs. The proposed admission control proceeds in two steps. First, by using its assigned guaranteed bandwidth, each ONU performs local rate-based admission control according to the bandwidth requirements of newly arriving flows and current bandwidth availability. Second, flows that could not be locally admitted are reported to the OLT, which in turn tries to admit them in the second subcycle, provided sufficient unused bandwidth is available.

The performance of the proposed admission control was studied by means of simulation using a strict priority and a deficit-weighted round-robin based intra-ONU scheduling algorithm for real-time voice and video streams, as well as best-effort data traffic. The obtained results show that the considered admission control is able to satisfy the QoS requirements in terms of delay bound and throughput.

Summary — Differentiated services require differentiated queuing and reporting and different grant sizing and scheduling treatment for each class at the OLT. As the number of queues increases, scalability becomes an issue and a hierarchical approach should be followed. Grant reservations are required for providing bandwidth guarantees. A fixed and properly-sized cycle length with fixed position of delay and jitter-sensitive traffic can provide delay and jitter guarantees. More research must be conducted in the area of providing bandwidth, delay, and jitter guarantees on an EPON.

FAIRNESS

EPONs carry traffic from a diverse group of non-cooperative subscribers. This non-cooperation requires fairness mechanisms to ensure that all nodes receive their "fair" share of the network resources. Fairness is most effectively tackled by the grant-sizing process and requires the use of the offline DBA framework. The limited and limited with excess distribution grant-sizing schemes offer a method to provide fairness on an EPON. We now discuss other approaches to fairness.

A sibling fair scheduler [9], in the case of EPONs, is a scheduler that ensures fairness for queues within an ONU. A cousin fair scheduler [9] can guarantee fairness among all leaves (i.e., queues) regardless of grouping. An EPON requires a scheduler that is fair among all queues and therefore requires a cousin fair scheduler. Fair Queuing with Service Envelopes (FQSE) [9] is a hierarchical scheduling algorithm that is cousin fair.

A service envelope (SE) is a grant to a node in the scheduling hierarchy. A service envelope for a leaf node is a piecewise linear function of how satisfied a node is, whereby satisfaction is measured in terms of a satisfiability parameter (SP). The SP begins at 0 when the SE is at its minimum guaranteed bandwidth and increases linearly until the request is completely satisfied. The slope of the linear increase is determined by the weight of the node.

The SE of non-leaf nodes is calculated as an approximation of the sum of the SEs of all their children. The approximation is necessary to keep the number of knee points in the piecewise linear function from exceeding a prescribed maximum. Limiting the number of knee points is necessary to keep the request message that conveys the piecewise linear function within a fixed length.

A deficit round robin approach within an ONU can be used to allow queues to contend for the cumulative slot remainder. This guarantees that there is only one slot remainder per ONU as opposed to a slot remainder for each queue. As a result, unused slot remainders are reduced. Simulation and mathematical analysis [9] show that FQSE is capable of making bandwidth guarantees, as well as providing fairness between all queues regardless of which ONU they reside in (i.e., cousin fairness).

For open access EPONs, in which multiple service providers share a single EPON, fairness must be maintained among service providers, as well as subscribers. Dual service level agreements (Dual-SLAs) [47] can be used to manage the fairness for both subscribers and service providers. A Dual-SLA manages two sets of SLAs, one for the subscribers and one for the service providers. One of the sets of SLAs is selected as the primary. The primary SLA set is given priority over the secondary.

MAC Protocols for Alternative PON Architectures

MAC Protocols for a Broadcast PON

Full-utilization local-loop request contention multiple access (FULL-RCMA) [5] is an extension of RCMA that allows for transmission interleaving to tolerate the walk times on an EPON and provides a bounded cycle time for support of CBR traffic. RCMA is a hybrid between a token-passing and contention MAC protocol. A cycle of time is divided into a contention-based request period and a token-passing data period. During the request period, ONUs contend through request messages to be added to the token-passing list for the upcoming data period. One of the ONUs is designated the master and generates the token-passing list for the upcoming data period. ONUs, according to the order determined by the master, pick up the token to gain access to the medium for their transmission. If ONUs remain backlogged after receiving their transmission window, they set a "more data" bit to signify that they automatically should be added to the token-passing list for the next cycle. This helps to reduce the probability of request collisions. ONUs that are added to the token-passing list through request contention are given priority over ONUs that are added through setting the "more data" bit in the previous cycle. This allows new ONUs to gain prompt access to the network. Performance analysis shows that FULL-RCMA can provide higher upstream link utilization compared to IPACT or RCMA. However, a comparison of average packet delay between the schemes is not available.

For the MAC protocol proposed in [6], cycles are divided into a control period and a data period separated by a waiting time that is used to collect all control messages and to produce a schedule. The control period is divided using fixed time division multiple access (TDMA), one slot for each ONU. The ONUs independently compute the same transmission schedule given the control information, and the data period is divided between the ONUs according to this schedule. Because the control messages and data messages are separated, there are $2 \cdot N$ guard intervals per scheduling cycle as opposed to N guard intervals for a typical EPON, in which the control information is appended to the end of the data transmissions. This increased number of guard intervals degrades the channel utilization.

MAC Protocols for Two-Stage PONs

A DBA scheme for a two-stage PON called EPON Dynamic Scheduling Algorithm (EDSA-2) [7] takes advantage of some predictability of the aggregated traffic from the sub-OLTs. Because sub-OLTs aggregate traffic from several bursty sources, the sub-OLT traffic tends to be less bursty and hence, more predictable. Specifically, EDSA-2 predicts the traffic during t_{rgs} by assuming short-term rate stationarity from the aggregated ONUs. Differentiated services are provided by having a set of class queues for the local sub-OLT traffic, as well as a separate set of class queues for the aggregated ONU traffic. The aggregate ONU class queues are given priority over the corresponding class queues for the local sub-OLT traffic.

Simulation results that compare EDSA-2 to a QoS DBA that performs no traffic prediction (EDSA-1) show that EDSA-2 provides lower average packet delay for all classes of traffic from both standard ONUs and the sub-OLT ONUs. The lower delays for the sub-OLT traffic can be explained by the accommodation of traffic received during the granting period (i.e., t_{rgs}) through traffic prediction. However, it is not clear why the standard ONUs would experience lower delays.

Conclusion

In this article, we summarized and classified the existing research on EPONs. We introduced a meaningful framework that allows those interested in advancing EPON research to quickly understand the state-of-the-art and to identify areas requiring further study. We outlined the standard physical PON architecture, as well as two alternative architectures, broadcast PON and two-stage PON. We also examined and provided a meaningful taxonomy for dynamic bandwidth allocation. Using this taxonomy, we presented the existing work on dynamic bandwidth allocation. The major branches of the taxonomy are

- Grant sizing
- Grant scheduling
- Queue scheduling

We also surveyed the existing approaches for supporting quality of service and fairness. Finally, we presented a discussion of protocols for the alternative physical PON architectures. We conclude by outlining areas that we believe are in urgent need of further research.

The problem of ONU grant sizing has received significant attention from the research community, but there are still important open questions. Providing an exhaustive service discipline for EPONs can lower queuing delays by up to a cycle time. However, due to the nature of EPONs as a remote scheduling system, an exhaustive service discipline is not possible without queue-size prediction. Prediction of queued CBR traffic is straightforward as a result of its constant rate. VBR traffic, on the other hand, is difficult to predict. With the proliferation of VBR video through IPTV services, it would be worthwhile to explore schemes to predict its short-term bandwidth requirements. These schemes can be used for queue-size prediction to help lower the queuing delay for this delay-sensitive traffic. Data traffic is typically delay insensitive, so there is less of a need to reduce the queuing delays for this type of traffic.

The problem of distributing excess bandwidth has been explored in the context of the offline DBA framework (i.e., interleaved polling with stop). It would be of value to explore the possibility of providing fair excess bandwidth distribution in the online DBA framework (i.e., a purely interleaved granting system).

The topic of ONU grant scheduling has received some attention from the research community. However, we feel this topic can be explored further to uncover the best grant scheduler for an EPON. The topic of ONU grant scheduling anchored in scheduling theory has been studied extensively in [48, 49] within the context of multiple-channel EPONs. Similar exposition should be extended to single-channel EPONs.

With respect to providing QoS on EPONs, providing differentiated services has received significant attention from the research community. However, providing bandwidth, delay, and delay variation (i.e., jitter) guarantees requires further study. Providing guaranteed service across an EPON will be critical because the access network is required to be an integrated services network carrying packetized voice and video along with data traffic. The voice and video services will require some level of bandwidth, delay, and jitter guarantees for successful operation.

Emerging from the work on single-channel EPONs, researchers are beginning to extend the DBA problem to EPONs that employ more than one upstream and/or downstream channel [14, 50, 51, 52]. DBA for multi-wavelength EPONs represents a broad area for future research.

REFERENCES

[1] M. Hajduczenia *et al.*, "On EPON Security Issues," *IEEE Commun. Surveys and Tutorials*, vol. 9, no. 1, 1st Quarter 2007, pp. 68–83.

[2] M. McGarry, M. Maier, and M. Reisslein, "Ethernet PONs: A Survey of Dynamic Bandwidth Allocation (DBA) Algorithms," *IEEE Commun. Mag.*, vol. 42, no. 8, Aug. 2004, pp. S8–S15.

[3] G. Kramer, B. Mukherjee, and G. Pesavento, "IPACT: A Dynamic Protocol for an Ethernet PON (EPON)," *IEEE Commun. Mag.*, vol. 40, no. 2, Feb. 2002, pp. 74–80.

[4] G. Kramer, B. Mukherjee, and G. Pesavento, "Ethernet PON (ePON): Design and Analysis of an Optical Access Network," *Photonic Network Commun.*, vol. 3, no. 3, July 2001, pp. 307–19.

[5] C. Foh *et al.*, "FULL-RCMA: A High Utilization EPON," *IEEE JSAC*, vol. 22, no. 8, Oct. 2004, pp. 1514–24.

[6] S. Sherif *et al.*, "A Novel Decentralized Ethernet-Based PON Access Architecture for Provisioning Differentiated QoS," *IEEE/OSA J. Lightwave Tech.*, vol. 22, no. 11, Nov. 2004, pp. 2483–97.

[7] A. Shami *et al.*, "QoS Control Schemes for Two-Stage Ethernet Passive Optical Access Networks," *IEEE JSAC*, vol. 23, no. 8, Aug. 2005, pp. 1467–78.

[8] T. Holmberg, "Analysis of EPONs under the Static Priority Scheduling Scheme with Fixed Transmission Times," *Proc. IEEE Conf. Next Generation Internet Design and Engineering (NGI)*, Apr. 2006, pp. 192–99.

[9] G. Kramer *et al.*, "Fair Queuing with Service Envelopes (FQSE): A Cousin-Fair Hierarchical Scheduler For Subscriber Access Networks," *IEEE JSAC*, vol. 22, no. 8, Oct. 2004, pp. 1497–1513.

[10] S. Bhatia, D. Garbuzov, and R. Bartos, "Analysis of the Gated IPACT Scheme for EPONs," *Proc. IEEE ICC*, June 2006, pp. 2693–98.

[11] F. Aurzada *et al.*, "Delay Analysis of Ethernet Passive Optical Networks with Gated Service," Arizona State University Technical Report, Mar. 2007.

[12] J. Zheng and H. Mouftah, "Media Access Control for Ethernet Passive Optical Networks: An Overview," *IEEE Commun. Mag.*, vol. 43, no. 2, Feb. 2005, pp. 145–50.

[13] C. Assi *et al.*, "Dynamic Bandwidth Allocation for Quality-of-Service over Ethernet PONs," *IEEE JSAC*, vol. 21, no. 9, Nov. 2003, pp. 1467–77.

[14] A. Dhaini *et al.*, "Dynamic Wavelength and Bandwidth Allocation in Hybrid TDM/WDM EPON Networks," *IEEE/OSA J. Lightwave Tech.*, vol. 25, no. 1, Jan. 2007, pp. 277–86.

[15] X. Bai, A. Shami, and C. Assi, "On the Fairness of Dynamic Bandwidth Allocation Schemes in Ethernet Passive Optical Networks," *Computer Commun.*, vol. 29, no. 11, July 2006, pp. 2123–35.

[16] D. Bertsekas and R. Gallager, *Data Networks, 2nd Ed.*, Prentice Hall, 1991.

[17] A. Shami *et al.*, "Jitter Performance in Ethernet Passive Optical Networks," *IEEE/OSA J. Lightwave Tech.*, vol. 23, no. 4, Apr. 2005, pp. 1745–53.

[18] H.-J. Byun, J.-M. Nho, and J.-T. Lim, "Dynamic Bandwidth Allocation Algorithm in Ethernet Passive Optical Networks," *Electronics Letters*, vol. 39, no. 13, June 2003, pp. 1001–2.

[19] Y. Luo and N. Ansari, "Bandwidth Allocation for Multiservice Access on EPONs," *IEEE Commun. Mag.*, vol. 43, no. 2, Feb. 2005, pp. S16–S21.

[20] Y. Luo and N. Ansari, "Limited Sharing with Traffic Prediction for Dynamic Bandwidth Allocation and QoS Provisioning over Ethernet Passive Optical Networks," *OSA J. Opt. Net.*, vol. 4, no. 9, Sept. 2005, pp. 561–72.

[21] S. Haykin, *Neural Networks: A Comprehensive Foundation, 2nd Ed.*, Prentice Hall, 1999.

[22] S. Ostring and H. Sirisena, "The Influence of Long-Range Dependence on Traffic Prediction," *Proc. IEEE Int'l. Conf. Communications (ICC)*, vol. 4, June 2001, pp. 1000–1005.

[23] G. Kramer *et al.*, "Supporting Differentiated Classes of Service in Ethernet Passive Optical Networks," *OSA J. Opt. Net.*, vol. 1, no. 8, Aug. 2002, pp. 280–98.

[24] Y. Zhu, M. Ma, and T. Cheng, "Hierarchical Scheduling to Support Differentiated Services in Ethernet Passive Optical Networks," *Computer Networks*, vol. 50, no. 3, Feb. 2006, pp. 350–66.

[25] M. Ma, L. Liu, and T. H. Cheng, "Adaptive Scheduling for Differentiated Services in the Ethernet Passive Optical Networks," *Proc. 9th Int'l. Conf. Commun. Systems*, Sept. 2004, pp. 102–6.

[26] M. Pinedo, *Scheduling: Theory, Algorithms, and Systems, 2nd Ed.*, Prentice Hall, 2002.

[27] J. Zheng and H. Mouftah, "Adaptive Scheduling Algorithms for Ethernet Passive Optical Networks," *IEE Proc. Commun.*, vol. 152, no. 5, Oct. 2005, pp. 643–47.

[28] J. Zheng, "Efficient Bandwidth Allocation Algorithm for Ethernet Passive Optical Networks," *IEE Proc. Commun.*, vol. 153, no. 3, June 2006, pp. 464–68.

[29] A. Parekh and R. Gallager, "A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: The Single-Node Case," *IEEE/ACM Trans. Net.*, vol. 1, no. 3, June 1993, pp. 344–57.

[30] A. Demers, S. Keshav, and S. Shenker, "Analysis and Simulation of a Fair Queuing Algorithm," *Internetworking: Research and Experience*, vol. 1, no. 1, Sept. 1990, pp. 3–26.

[31] P. Goyal, H. M. Vin, and H. Cheng, "Start-Time Fair Queuing: A Scheduling Algorithm for Integrated Services Packet Switching Networks," *IEEE/ACM Trans. Net.*, vol. 5, no. 5, Oct. 1997, pp. 690–704.

[32] N. Ghani *et al.*, "Quality of Service in Ethernet Passive Optical Networks," *Proc. 2004 IEEE/Sarnoff Symp. Advances in Wired and Wireless Commun.*, Apr. 2004, pp. 161–65.

[33] N. Ghani *et al.*, "Intra-ONU Bandwidth Scheduling in Ethernet Passive Optical Networks," *IEEE Commun. Letters*, vol. 8, no. 11, Nov. 2004, pp. 683–85.

[34] A. Dhaini *et al.*, "Adaptive Fairness through Intra-ONU Scheduling for Ethernet Passive Optical Networks," *Proc. IEEE Int. Conf. Commun. (ICC)*, June 2006, pp. 2687–92.

[35] S.-I. Choi and J.-D. Huh, "Dynamic Bandwidth Allocation Algorithm for Multimedia Services over Ethernet PONs," *ETRI Journal*, vol. 24, no. 6, Dec. 2002, pp. 465–68.

[36] J. Xie, S. Jiang, and Y. Jiang, "A Dynamic Bandwidth Allocation Scheme for Differentiated Services in EPONs," *IEEE Commun. Mag.*, vol. 42, no. 8, Aug. 2004, pp. S32–S39.

[37] H. Naser and H. Mouftah, "A Joint-ONU Interval-Based Dynamic Scheduling Algorithm for Ethernet Passive Optical Networks," *IEEE/ACM Trans. Net.*, vol. 14, no. 4, Aug. 2006, pp. 889–99.

[38] M. Ma, Y. Zhu, and T. Cheng, "A Bandwidth Guaranteed Polling MAC Protocol for Ethernet Passive Optical Networks," *Proc. IEEE INFOCOM*, vol. 1, Mar. 2003, San Francisco, pp. 22–31.

[39] M. Ma, Y. Zhu, and T. Cheng, "A Systematic Scheme for Multiple Access in Ethernet Passive Optical Access Networks," *IEEE/OSA J. Lightwave Technology*, vol. 23, no. 11, Nov. 2005, pp. 3671–82.

[40] L. Zhang *et al.*, "Dual DEB-GPS Scheduler for Delay-Constraint Applications in Ethernet Passive Optical Networks," *IEICE Trans. Commun.*, vol. E86-B, no. 5, May 2003, pp. 1575–84.

[41] F. An *et al.*, "A New Dynamic Bandwidth Allocation Protocol with Quality of Service in Ethernet-Based Passive Optical Networks," *Proc. IASTED Int'l. Conf. Wireless and Optical Communications (WOC 2003)*, vol. 3, July 2003, pp. 165–69.

[42] J. Heinanen *et al.*, "Assured Forwarding PHB Group," RFC 2597 (Proposed Standard), June 1999, updated by RFC 3260; http://www.ietf.org/rfc/rfc2597.txt

[43] V. Jacobson, K. Nichols, and K. Poduri, "An Expedited Forwarding PHB," RFC 2598 (Proposed Standard), June 1999, obsoleted by RFC 3246; http://www.ietf.org/rfc/rfc2598.txt

[44] A. Dhaini *et al.*, "Admission Control in Ethernet Passive Optical Networks (EPONs)," *Proc. IEEE Int'l. Conf. Communications (ICC)*, Glasgow, Scotland, June 2007.

[45] A. Dhaini *et al.*, "Per-Stream QoS and Admission Control in Ethernet Passive Optical Networks (EPONs)," *IEEE/OSA J. Lightwave Tech.*, vol. 25, no. 7, July 2007, pp. 1659–69.

[46] C. Assi, M. Maier, and A. Shami, "Toward Quality-of-Service Protection in Ethernet Passive Optical Networks: Challenges and Solutions," *IEEE Network*, vol. 21, no. 5, Sept.–Oct. 2007.

[47] A. Banerjee, G. Kramer, and B. Mukherjee, "Fair Sharing

Using Dual Service-Level Agreements to Achieve Open Access in a Passive Optical Network," *IEEE JSAC*, vol. 24, no. 8, Aug. 2006, pp. 32–44.
[48] M. McGarry *et al.*, "Bandwidth Management for WDM EPONs," *OSA J. Opt. Net.*, vol. 5, no. 9, Sept. 2006, pp. 637–54.
[49] M. McGarry *et al.*, "Just-in-Time Online Scheduling for WDM EPONs," *Proc. IEEE ICC 2007*, June 2007.
[50] K. Kwong, D. Harle, and I. Andonovic, "Dynamic Bandwidth Allocation Algorithm for Differentiated Services over WDM EPONs," *Proc. 9th IEEE Int'l. Conf. Commun. Systems (ICCS)*, Sept. 2004, pp. 116–20.
[51] M. McGarry, M. Maier, and M. Reisslein, "WDM Ethernet Passive Optical Networks," *IEEE Commun. Mag.*, vol. 44, no. 2, Feb. 2006, pp. S18–S25.
[52] A. Dhaini, C. Assi, and A. Shami, "Quality of Service in TDM/WDM Ethernet Passive Optical Networks (EPONs)," *Proc. IEEE ISCC 2006*, June 2006, pp. 616–21.

BIOGRAPHIES

MICHAEL MCGARRY (mmcgarry@uakron.edu) received his BS in Computer Engineering from Polytechnic University, Brooklyn, NY in 1997. He received his M.S. and Ph.D. in Electrical Engineering from Arizona State University, Tempe, AZ in 2004 and 2007 respectively. He is an Assistant Professor at the University of Akron, Akron, OH. From 2007 through 2008 he was a Senior Staff Scientist at ADTRAN and an Adjunct Professor at Arizona State University. From 1997 through 2003 he was employed in industry by companies including PMC-Sierra and Yurie Systems (now part of Alcatel-Lucent). His research interests include congestion control and the optimization of MAC protocols for both optical access and mobile ad hoc networks.

MARTIN REISSLEIN (reissleing@asu.edu) received his Ph.D. in systems engineering from the University of Pennsylvania, Philadelphia, in 1998. He is an associate professor in the Department of Electrical Engineering at Arizona State University, Tempe. From July 1998 through October 2000 he was a scientist with the German National Research Center for Information Technology (GMD FOKUS), Berlin, and lecturer at the Technical University Berlin. He maintains an extensive library of video traces for network performance evaluation, including frame size traces of MPEG-4 and H.264 encoded video, at http://trace.eas.asu.edu.

MARTIN MAIER (maier@ieee.org) received his M.Sc. and Ph.D. degrees (both with distinction) in electrical engineering from the Technical University Berlin, Berlin, Germany, in 1998 and 2003, respectively. In the summer of 2003, he was a Postdoctoral Fellow at the Massachusetts Institute of Technology (MIT), Cambridge. He is an associate professor with the Institut National de la Recherche Scientifique (INRS), Montreal, Canada. His recent research activities aim at providing insights into technologies, protocols, and algorithms shaping the future of optical networks and their seamless integration with next-generation wireless networks. He was a visiting professor at Stanford University, California, October 2006 through March 2007.