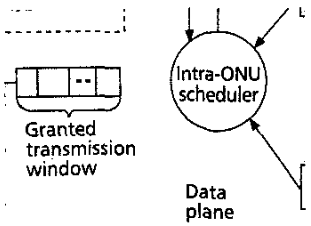


Ethernet PONs: A Survey of Dynamic Bandwidth Allocation (DBA) Algorithms

Michael P. McGarry, Arizona State University
Martin Maier, Centre Tecnològic de Telecomunicacions de Catalunya
Martin Reisslein , Arizona State University

Abstract

Optical networks are poised to dominate the access network space in coming years. Ethernet passive optical networks, which leverage the ubiquity of Ethernet at subscriber locations, seem destined for success in the optical access network. In this article we first provide a brief introduction to Ethernet passive optical networks, followed by a discussion of the problem of dynamic bandwidth allocation. We then introduce a framework for classifying dynamic bandwidth allocation schemes and provide a comprehensive survey of the dynamic bandwidth allocation methods proposed to date. We conclude with a side by side comparison of the schemes based on their most prominent characteristics, and outline future developments of dynamic bandwidth allocation schemes.

Introduction

Access networks connect business and residential subscribers to the central offices of service providers, which in turn are connected to metropolitan area networks (MANs) or wide area networks (WANs). Access networks are commonly referred to as the *last mile* or *first mile*; the latter term emphasizes their importance to subscribers. In today's access networks, telephone companies deploy digital subscriber loop (xDSL) technologies, and cable companies deploy cable modems. Typically, these access networks are hybrid fiber coax (HFC) systems with an optical-fiber-based feeder network between central office and remote node, and an electrical distribution network between remote node and subscribers. These access technologies are unable to provide enough bandwidth to current high-speed Gigabit Ethernet local area networks (LANs) and evolving services (e.g., distributed gaming or video on demand). Future first-mile solutions have to not only provide more bandwidth but also meet the cost sensitivity constraints of access networks arising from the small number of cost sharing subscribers.

In so-called FTTx access networks the copper-based distribution part of access networks is replaced with optical fiber; for example, fiber to the curb (FTTC) or home (FTTH). In doing so, the capacity of access networks is sufficiently increased to provide broadband services to subscribers. Due to the cost sensitivity of access networks, these all-optical FTTx systems are typically unpowered and consist of passive optical components (e.g.,, splitters and couplers). Accordingly, they are called *passive optical networks* (PONs). PONs attracted a great deal of attention well before the Internet spurred bandwidth growth. The Full Service Access Network (FSAN) Group (http://www.fsanweb.org) was the driving force behind the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) G.983 broadband PON (BPON) specification, which defines a PON with asynchronous transfer mode (ATM) as its native protocol data unit (PDU). ATM suffers from several shortcomings. Due to the segmentation of variable-size IP packets into fixed-size cells, ATM imposes a cell tax overhead. Moreover, a single corrupted or dropped ATM cell requires retransmission of the entire IP packet even though the remaining ATM cells belonging to the corresponding IP packet are received correctly, resulting in wasted bandwidth and processing resources. Finally, ATM equipment (switches, network cards) is quite expensive.

Recently, Ethernet PONs (EPONs) have gained a great amount of interest in both industry and academia as a promising cost-effective solution for next-generation broadband access networks, as illustrated by the formation of several fora and working groups, including the EPON Forum (http://www.ieeecommunities.org/epon), the Ethernet in the First Mile Alliance (http://www.efmalliance.org), and the IEEE 802.3ah working group (http://www.ieee802.org/3/efm). EPONs carry data encapsulated in Ethernet frames, which makes it easy to carry IP packets and eases interoperability with installed Ethernet LANs. EPONs represent the convergence of low-cost Ethernet equipment (switches, network interface cards [NICs]) and low-cost fiber architectures. Furthermore, given the fact that more than 90 percent of today's data traffic originates from and terminates in Ethernet LANs, EPONs appear to be a natural candidate for future first-mile solutions. The main standardization body behind EPON is the

0163-6804/04/$20.00 © 2004 IEEE

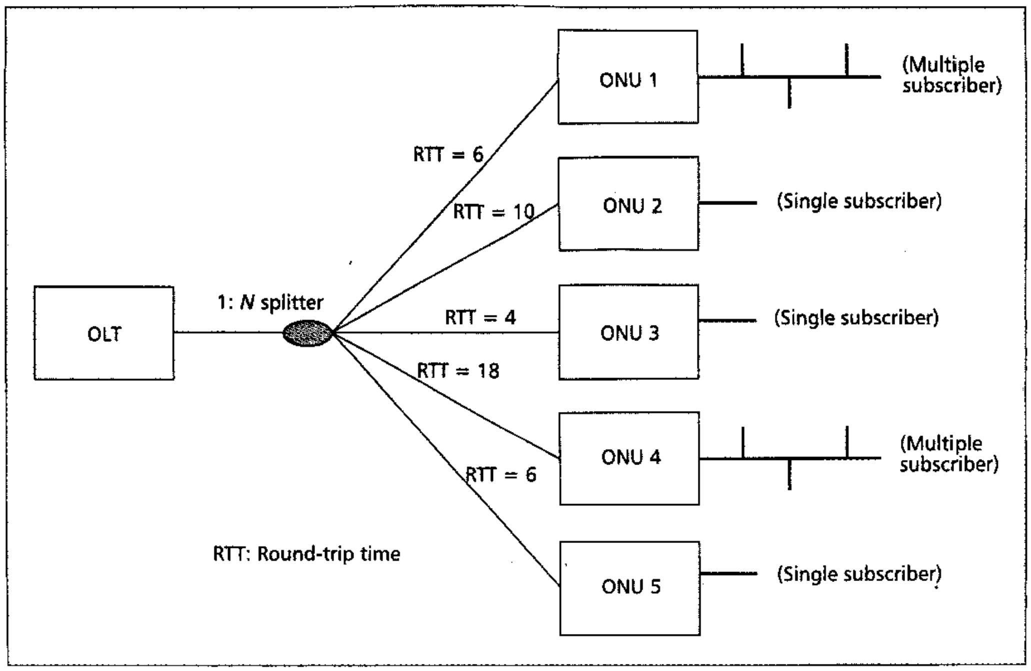


FIGURE 1. Network architecture of an EPON with one optical line terminal (OLT) and $N = 5$ optical network units (ONUs), each with a different round-trip time (RTT).

IEEE 802.3ah Task Force. This task force is developing the so-called multipoint control protocol (MPCP), which arbitrates channel access among central office and subscribers. MPCP is used to dynamically assign the upstream bandwidth (subscriber to service provider), which is the key challenge in access protocol design for EPONs. Note that MPCP does not specify any particular dynamic bandwidth allocation (DBA) algorithm. Instead, it is intended to facilitate the implementation of DBA algorithms.

To understand the importance of DBA in EPONs, note that the traffic on individual links in the access network is quite bursty. This is in contrast to MANs or WANs, 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 conditions due to a small number of ON/OFF sources. Due to this bursty nature the bandwidth requirements vary widely with time. Therefore, static allocation of bandwidth to individual subscribers (or sets of subscribers) in an EPON is very inefficient. Employing a DBA algorithm that adapts to instantaneous bandwidth requirements is much more efficient, capitalizing on the benefits of statistical multiplexing. Hence, DBA is a critical feature for EPON design. In this article we provide a comprehensive survey of the DBA algorithms examined to date; these algorithms are candidates for implementation in future EPONs.

The remainder of the article is organized as follows. First we describe the EPON architecture. Next we highlight the main features of MPCP. Then we provide a taxonomy and a survey of the DBA algorithms for EPONs. The final section concludes the article by providing a comparison of the various EPON DBA algorithms and outlining avenues for future development.

EPON Architecture

Typically, EPONs (and PONs in general) have a physical tree topology with the central office located at the root and the subscribers connected to the leaf nodes of the tree, as illustrated in Fig. 1. At the root of the tree is an optical line terminal (OLT), which is the service provider equipment residing at the central office. The EPON connects the OLT to multiple optical network units (ONUs) (the customer premises equipment) through a 1:*N* optical splitter/combiner. An ONU can serve a single residential or business subscriber, referred to as fiber to the home/business (FTTH/B), or multiple subscribers, referred to as fiber to the curb (FTTC). Each ONU buffers data received from the attached subscriber(s). In general, the round-trip time (RTT) between OLT and each ONU is different. Due to the directional properties of the optical splitter/combiner, the OLT is able to broadcast data to all ONUs in the downstream direction. In the upstream direction, however, ONUs cannot communicate directly with one another. Instead, each ONU is able to send data only to the OLT. Thus, in the downstream direction an EPON may be viewed as a point-to-multipoint network and in the upstream direction as a multipoint-to-point network [1]. Due to this fact, the original Ethernet media access control (MAC) protocol does not operate properly since it relies on a broadcast medium. (Instead, the MPCP arbitration mechanism is deployed, as discussed in the subsequent section.)

In the upstream direction, all ONUs share the transmission medium. To avoid collisions, several approaches can be used. Wavelength-division multiplexing (WDM) is currently considered cost prohibitive since the OLT would require a tunable receiver or a receiver array to receive data on multiple wavelength channels, and each ONU would need to be equipped with a wavelength-specific transceiver. At present, time-division multiplexing (TDM) is considered a more cost-effective solution. With TDM a single transceiver is required at the OLT, and there is just one type of ONU equipment [2, 3]. Note that this does not prevent EPONs from being upgraded to multiple wavelength channels (WDM) in the future. Given the aforementioned different connectivity in upstream and downstream directions of EPONs, the OLT appears to be the best suited node to arbitrate the time sharing of the channel.

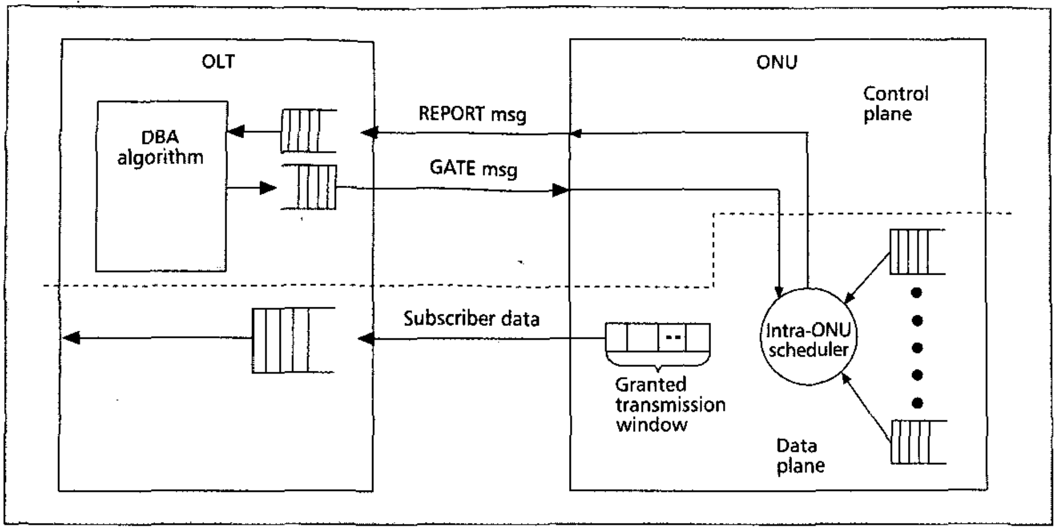


FIGURE 2. MPCP operation: Two-way messaging assignment of time slots for upstream transmission between ONU and OLT.

MULTIPOINT CONTROL PROTOCOL

The MPCP arbitration mechanism developed by the IEEE 802.3ah Task-Force is used to dynamically assign nonoverlapping upstream transmission windows (time slots) to each ONU. Besides auto-discovery, registration, and ranging (RTT computation) operations for newly added ONUs, MPCP provides the signaling infrastructure (control plane) for coordinating data transmissions from the ONUs to the OLT.

The basic idea is that the upstream bandwidth is divided into bandwidth units via TDM. These units are assigned to the ONUs as determined by the OLT according to the DBA algorithm in use. The OLT has control over the assignment of these units of bandwidth. These units can be assigned on the fly as needed or can be reserved in advance. For efficiency reasons, any reserved units or fraction of units of bandwidth that go unused can in general be re-assigned on the fly by the OLT to other ONUs that could make use of it. As shown in Fig. 2, MPCP uses two types of messages to facilitate arbitration: REPORT and GATE. Each ONU has a set of queues, possibly prioritized, holding Ethernet frames ready for upstream transmission to the OLT. The REPORT message is used by an ONU to report bandwidth requirements (typically in the form of queue occupancies) to the OLT. A REPORT message can support the reporting of up to 13 queue occupancies of the corresponding ONU. Upon receiving a REPORT message, the OLT passes it to the DBA algorithm module. The DBA module calculates the upstream transmission schedule of all ONUs such that channel collisions are avoided. Note that MPCP does not specify any particular DBA algorithm. MPCP simply provides a framework for the implementation of various DBA algorithms. After executing the DBA algorithm, the OLT transmits GATE messages to issue transmission grants. Each GATE message can support up to four transmission grants. Each transmission grant contains the transmission start time and transmission length of the corresponding ONU. Each ONU updates its local clock using the timestamp contained in each received transmission grant. Thus, each ONU is able to acquire and maintain global synchronization. The transmission start time is expressed as an absolute timestamp according to this global synchronization. Each ONU sends backlogged Ethernet frames during its granted transmission window using its local intra-ONU scheduler. The intra-ONU scheduler schedules the packet transmission from the various local queues. The transmission window may comprise multiple Ethernet frames; packet fragmentation is not allowed. As a consequence, if the next frame does not fit into the current transmission window, it has to be deferred to the next granted transmission window.

DYNAMIC BANDWIDTH ALLOCATION ALGORITHMS

In this section we survey the DBA algorithms proposed to date for EPONs. The discussed DBA algorithms can be used in the DBA module of the above described MPCP arbitration mechanism to calculate the collision-free upstream transmission schedule of ONUs and generate GATE messages accordingly. The taxonomy used in our survey is depicted in Fig. 3. We categorize the DBA algorithms for EPONs into algorithms with statistical multiplexing and algorithms with quality of service (QoS) assurances. The latter are further subdivided into algorithms with absolute and relative QoS assurances. In the following, we discuss the DBA algorithms of each class in greater detail.

STATISTICAL MULTIPLEXING METHODS

Interleaved Polling with Adaptive Cycle Time — In the interleaved polling with adaptive cycle time (IPACT) approach, the OLT polls the ONUs individually and issues transmission grants to them in a round-robin fashion [4]. The grant window size of each ONU's first grant, $G(1)$, is set to some arbitrary value. After n cycles, the backlog (in bytes) in each ONU's transmission buffer, $Q(n)$ (reported queue size), is piggybacked to the current data transmission from the corresponding ONU to the OLT during its grant window $G(n)$. The backlog $Q(n)$ is measured at the instant when the ONU generates the request message, which is piggybacked to the data transmission in cycle n. This backlog $Q(n)$ is used to determine the grant window size of the next grant $G(n + 1)$ of the ONU. In doing so, bandwidth is dynamically assigned to ONUs according to their queue occupancies. If a given ONU's queue is empty, the OLT still grants a transmission window of zero byte to that ONU such that the ONU is able to report its queue occupancy for the next grant. IPACT deploys in-band signaling of bandwidth requests by using escape characters within Ethernet frames instead of sacrificing an entire Ethernet frame for control (as in MPCP), resulting in a reduced signaling overhead. The OLT keeps track of the RTTs of all ONUs. As a result, the OLT can send out a grant to the next ONU in order to achieve a very tight guard band between consecutive upstream transmissions, resulting in improved bandwidth utilization. The guard band between two consecutive upstream transmissions is needed to compensate for RTT fluctuations and to give the OLT enough time to adjust its receiver to the transmission power level of the next ONU.

In IPACT each ONU is served once per round-robin polling cycle. The cycle length is not static but adapts to the instantaneous bandwidth requirements of the ONUs. By using a maximum transmission window (MTW), ONUs with high traffic volume are prevented from monopolizing the bandwidth. The OLT allocates the upstream bandwidth to ONUs in one of the following ways:

- *Fixed service*: This DBA algorithm ignores the requested window size and always grants the MTW size. As a result, the cycle time is constant.
- *Limited service*: This DBA algorithm grants the requested number of bytes, but no more than the MTW.
- *Credit service*: This DBA algorithm grants the requested window plus either a constant credit or a credit that is proportional to the requested window.
- *Elastic service*: This DBA algorithm attempts to overcome the limitation of assigning at most one fixed MTW to an ONU in a round. The maximum window granted to an ONU is such that the accumulated size of the last N grants does not exceed N MTWs, where N denotes the number of ONUs. Thus, if only one ONU is backlogged, it may get a grant of up to N MTWs.

The simulation results reported in [4] indicate that both the average packet delays and average queue lengths with the IPACT method with either limited, credit, or elastic service DBA were almost two orders of magnitude smaller than fixed service DBA (fixed service is a static bandwidth allocation) under light traffic loads. Under heavy loads, the average packet delays and average queue lengths for all four types of service were similar. Generally, limited, credit, and elastic service DBA all provided very similar average packet delays and average queue lengths.

In summary, IPACT improves channel utilization efficiency by reducing the overhead arising from walk times (propagation delay) in a polling system. This is achieved by overlapping multiple polling requests in time. As opposed to static TDM systems, IPACT allows for statistical multiplexing and dynamically allocates upstream bandwidth according to the traffic demands of the ONUs within adaptive polling cycles. Furthermore, IPACT deploys an efficient in-band signaling approach that avoids using extra Ethernet frames for control. By using an MTW, throughput fairness among ONUs is achieved. On the downside, this original design for IPACT does not support QoS assurances or service differentiation by means of reservation or prioritization of bandwidth assignment. An IPACT extension to support multiple service classes was developed in [5], which we review later.

Control Theoretic Extension of IPACT — In IPACT, the ONU requests (reports) the amount of backlogged traffic $Q(n)$ as a grant for the next cycle. One drawback of this approach is that the request does not take into consideration the amount of traffic arriving at the ONU between generation of the request message in cycle n and arrival of the grant $G(n + 1)$ for the next cycle at the ONU. As a consequence, traffic arriving after generation of a request message is only taken into consideration in the next request message and hence experiences typically a queuing delay of one cycle in the ONU.

To overcome this queuing delay, a control theoretic extension to IPACT was proposed in [6]. In this extension the amount of traffic arriving at the ONU between two successive requests is estimated; this estimate is incorporated into the grant to the ONU. More specifically, the estimation works as follows. Recall that $Q(n - 1)$ denotes the amount of backlogged traffic in the ONU at the instant when the request of cycle $n - 1$, which is used by the OLT to calculate grant $G(n)$, is generated. Let $A(n - 1)$ denote the amount of traffic arriving at the ONU between generating the request for cycle $n - 1$ and receiving the grant for cycle n. With these definitions, the difference between the grant for cycle n and the amount of traffic backlogged in the ONU when the grant arrives is approximately $D(n) = G(n) - [Q(n - 1) + A(n - 1)]$. The OLT allocates bandwidth based on the size of the previous grant and the scaled version of the difference reported by the ONUs. More specifically, the grant for cycle $n + 1$ is calculated as $G(n + 1) = G(n) - \alpha \cdot D(n)$, where α is the gain factor. Using control theoretic arguments, it is shown in [6] that for piecewise constant traffic with infrequent jumps the system is asymptotically stable for $0 < \alpha < 2$.

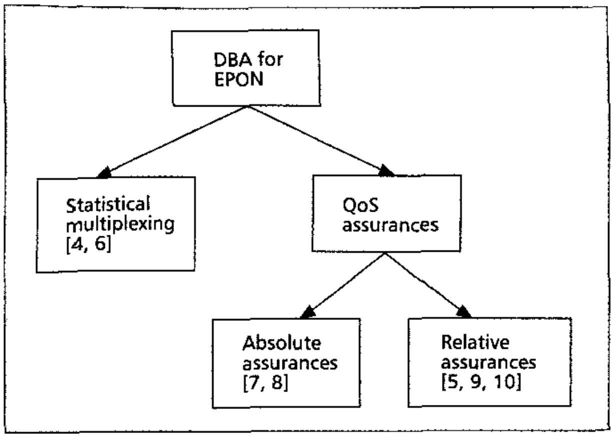

DBA for EPON
Statistical multiplexing [4, 6]
QoS assurances
Absolute assurances [7, 8]
Relative assurances [5, 9, 10]

FIGURE 3. Taxonomy of dynamic bandwidth allocation algorithms for EPONs.

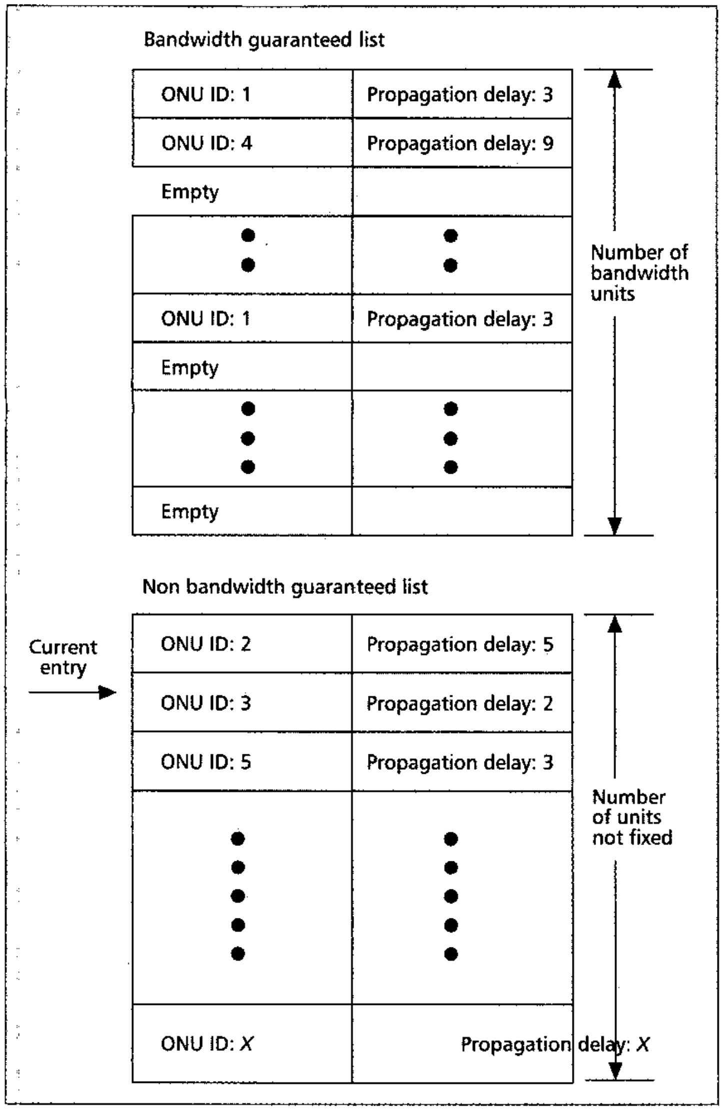
Note that this refinement to IPACT essentially views bandwidth assignment as an automatic control system with the goal of keeping the difference $D(n)$ close to zero. Proportional (P) control is proposed for this system with control gain α. The advantage of this control theoretic approach is that the grant size is typically closer to the size of the backlog at the instant of receiving the grant at the ONU. This in turn results in lower queuing delays. On the downside, the control system may require careful tuning to achieve a prompt response to changes in traffic load without creating oscillations in the system. This may be a challenging problem if the traffic load is highly variable.

ABSOLUTE QOS ASSURANCES

Bandwidth Guaranteed Polling — The Bandwidth Guaranteed Polling (BGP) method proposed in [7] divides ONUs into two disjoint sets: bandwidth guaranteed and best effort. Bandwidth guaranteed nodes are characterized by their service level agreement (SLA) with the network provider. The SLA specifies the bandwidth this node is to be guaranteed.

The total upstream bandwidth is divided into equivalent bandwidth units, whereby the bandwidth unit is chosen such that the total upstream bandwidth in terms of the bandwidth unit is larger than the number of ONUs. For instance, for a network with 64 ONUs and an upstream bandwidth of 1 Gb/s, the equivalent bandwidth unit may be chosen as 10 Mb/s, that is, the total upstream bandwidth corresponds to 100 bandwidth units. The OLT maintains two entry tables, one for bandwidth guaranteed ONUs (ONUs with IDs 1 and 4 in Fig. 4) and one for best effort ONUs (ONUs with IDs 2, 3, and 5 in Fig. 4). Each table entry (row) has two fields: ONU ID and propagation delay from ONU to OLT. The table for bandwidth guaranteed ONUs has as many rows (entries) as there are bandwidth units in the total upstream bandwidth. In the above example, the bandwidth guaranteed ONU table has 100 rows. The table for best effort nodes is not fixed in size. Entries in the bandwidth guaranteed ONU table are established for each bandwidth guaranteed ONU based on its SLA. If an ONU requires more than one bandwidth unit, these units are spread evenly through the table, as illustrated in Fig. 4 for ONU with ID 1, which is guaranteed a bandwidth of two bandwidth units (20 Mb/s in the example). Rows in the guaranteed bandwidth ONU table that are not occupied can be dynamically assigned to best effort nodes. The OLT polls the best effort ONUs during the rows that are not used by the bandwidth guaranteed ONUs in the order they are listed in the best effort table.

The OLT begins polling ONUs using the information in the two tables. The OLT polls an ONU by sending a Grant message to grant a window of size G, which is initially set to one bandwidth unit. The ONU decides based on the size of its output buffer if it has enough data to fully utilize the granted transmission window. The ONU sends a reply to the OLT with the amount of the window it intends to utilize, B, and then transmits this amount of data. The OLT, upon receiving a reply from an ONU, checks the amount of the granted window the currently polled ONU intends to use. If B is zero, the OLT immediately polls the next ONU in the table. (Note that this wastes bandwidth during the RTT to that next ONU, whereas polling to the first ONU can be interleaved with the preceding data transmission to avoid wasting bandwidth.) If B is between zero and some threshold G_{reuse}, whereby $G - G_{reuse}$ specifies the minimum portion of the bandwidth unit that can be effectively shared, the OLT polls the next best effort ONU ready for transmission and grants it a transmission window $G - B$. Lastly, if B is larger than the threshold G_{reuse}, the OLT will not poll the next ONU until the current grant has passed.


Bandwidth guaranteed list
Number of bandwidth units
Non bandwidth guaranteed list
Current entry
Number of units not fixed

ONU ID: 1	Propagation delay: 3
ONU ID: 4	Propagation delay: 9
Empty	
• •	• •
ONU ID: 1	Propagation delay: 3
Empty	
• • •	• • •
Empty	

ONU ID: 2	Propagation delay: 5
ONU ID: 3	Propagation delay: 2
ONU ID: 5	Propagation delay: 3
• • • • •	• • • • •
ONU ID: X	Propagation delay: X

FIGURE 4. An illustration of bandwidth guaranteed polling tables.

The simulation results reported in [7] indicate that for bandwidth guaranteed ONUs with four or more entries, the delays were an order of magnitude smaller than with IPACT. However, for bandwidth guaranteed ONUs with only one entry as well as best effort ONUs, the delays were orders of magnitude larger than with IPACT under light loads and almost an order of magnitude larger than with IPACT under heavy loads. On the other hand, for bandwidth guaranteed ONUs with four or more entries, the queue lengths were similar to IPACT for light loads and orders of magnitude shorter than with IPACT under heavy loads. However, for bandwidth guaranteed ONUs with only one entry as well as best effort ONUs, the queue lengths were orders of magnitude larger than with IPACT under light loads and similar to IPACT under heavy loads. It was also found that the throughput with BGP tends to be lower than that with IPACT, especially at heavy loads.

Overall, the advantage of the bandwidth guaranteed polling approach is that it ensures an ONU's receiving the bandwidth specified by its SLA and that the spacing between transmission grants corresponding to SLAs has a fixed bound. The approach also allows for the statistical multiplexing of traffic into unreserved bandwidth units as well as unused portions of a guaranteed bandwidth unit (i.e., if an ONU does not have enough traffic to use all the bandwidth specified in its SLA). One drawback of table-driven upstream transmission grants of fixed bandwidth units is that the upstream transmission tends to become fragmented, with each fragment requiring a guard band, which tends to reduce throughput and bandwidth utilization.

Deterministic Effective Bandwidth — In [8], a system in which ONUs and OLT employ deterministic effective bandwidth (DEB) admission control and resource allocation in conjunction with generalized processor sharing (GPS) scheduling is developed. In this system, a given ONU maintains several queues, typically one for each traffic source or each class of traffic sources. A given queue is categorized as either a QoS queue or a best effort queue, depending on the requirements of the corresponding traffic source (class). A given traffic source feeding into a QoS queue is characterized by leaky bucket parameters. The leaky bucket parameters are traffic descriptors widely used in QoS networking; they give the peak rate of the source, the maximum burst the source can send at the peak rate, as well as the long run average rate of the source. A source also specifies the maximum delay it can tolerate. The leaky bucket traffic characterization together with the delay limit of the source (class) are used to determine whether the system can support the traffic in the QoS queues at all ONUs without violating delay bounds (and also without dropping any traffic at a QoS queue) using techniques derived from the general theory of deterministic effective bandwidths.

During the operation of the network, the OLT assigns grants to a given ONU based on the aggregate effective bandwidth of the traffic of the QoS queues at the ONU. Roughly speaking, a given ONU is assigned grants proportional to the ratio of the aggregate effective bandwidth of the traffic of the ONU to the total aggregate effective bandwidth of the traffic of all ONUs supported by the OLT. In turn, a given ONU uses the grants it receives to serve its QoS queues in proportion to the ratio of the effective bandwidth of the traffic of a queue to the aggregate effective bandwidth of the traffic of the QoS queues supported by the ONU. A given ONU uses the grants not utilized by QoS queues to transmit from best effort queues.

The advantage of the DEB approach is that individual flows (or classes of flows) are provided with deterministic QoS guarantees, ensuring lossless bounded-delay service. In addition, best effort traffic flows can utilize bandwidth not needed by QoS traffic flows. One main drawback of the DEB approach is that it requires increased complexity to conduct admission control and update proportions of effective bandwidths of ongoing flows. In particular, conducting admission control and allocating grant resources may result in significant overhead for short-lived traffic flows (or classes of traffic).

Relative QoS Assurances

DBA for Multimedia — In dynamic bandwidth allocation for multimedia [9], traffic in each ONU is placed into one of three priority queues (high, medium, and low). These priorities are then used by the DBA algorithm to assign bandwidth. The sizes of the three priority queues in each ONU are reported to the OLT. Based on these, the OLT issues grants separately for each of the priorities in each of the ONUs. In particular, bandwidth is first handed out to the high-priority queues, satisfying all the requests of high-priority flows. The DBA algorithm then considers the requests from medium-priority flows. If it can satisfy all of the medium-priority requests with what is left over from high-priority requests, it does so. Otherwise, it divides the remaining bandwidth between all medium-priority flows, where the fraction of the bandwidth granted to each medium-priority flow is related to the fraction requested by each flow to the total of all medium-priority requests. Finally, if there is any leftover bandwidth after satisfying high- and medium-priority requests, it is distributed among low-priority flows in a manner identical to the case where all medium-priority flows requests cannot be fully satisfied.

Note that in the DBA for multimedia approach, bandwidth is essentially allocated using strict priority based on the requirements of each priority traffic class of the entire PON (all the ONUs connected to a single OLT). One feature of this approach is that the OLT controls the scheduling within the ONU. This comes at the expense of reporting the occupancies of the individual priority queues and issuing multiple grants to each ONU per cycle. Also, the OLT has the additional burden of deciding on the scheduling among the queues in the ONU. Note that the strict priority scheduling based on the traffic classes at the PON level may result in starvation of ONUs that have only low-priority traffic.

DBA for QoS — DBA for QoS [10] is a method of providing per-flow QoS in an EPON using differentiated services. Within each ONU, priority packet queuing and scheduling is employed per the differentiated services framework. This is similar to DBA for multimedia, but recall that in DBA for multimedia priority scheduling was performed at the PON level (all ONUs connected to a single OLT). In contrast, in DBA for QoS priority scheduling is performed at the ONU level.

Before we proceed to DBA for QoS [10], we review the IPACT extension to multiple service classes [5], which may be viewed as a precursor to DBA for QoS. In [5] a simulation study is conducted of supporting differentiated service to three classes of traffic with strict priority scheduling inside the ONU. The authors noticed an interesting phenomenon they dubbed *light-load penalty*. What they noticed was that under light loading, the lower-priority class experienced a significant average packet delay increase; the maximum packet delays for the higher priorities also exhibited similar behavior. This appears to be caused by queue reporting occurring at some time before strict priority scheduling is performed. This allows higher-priority traffic arriving after queue reporting but before the transmission grant to preempt lower-priority traffic that arrived before queue reporting. This problem seems to be exacerbated under light loading. The authors discuss two methods of dealing with light-load penalty. The first method involves scheduling the packets when the REPORT message is transmitted and placing them in a second-stage queue. This second-stage queue will be emptied out first into the timeslot provided through a grant in a GATE message. The second method involves predicting the number of high-priority packets arriving between queue reporting and the grant window so that the grant window will be large enough to accommodate the newly arriving high-priority packets. This second method inherently lowers the delay experienced by higher-priority traffic from that with the two-stage queuing approach.

In DBA for QoS [10] the authors incorporate a method similar to the two-stage queuing approach mentioned above. Specifically, in the DBA for QoS method the packet scheduler in the ONU employs priority scheduling only on the packets that arrive before some $t_{request}$, which is the time at which the REPORT message is sent to the OLT. This avoids the problem of having the ONU packet scheduler request bandwidth based on buffer occupancies at time $t_{request}$ and then actually schedule packets at time t_{grant} to fill the granted transmission window. If this mechanism is not employed, lower-priority queues can be starved more severely because higher-priority traffic arriving between $t_{request}$ and t_{grant} would tend to take away transmission capacity from the lower-priority queues. Note that this problem only arises with strict priority scheduling, which schedules lower-priority packets only when the higher-priority packet queues are empty. With weighted fair queuing, which serves the different priority queues in proportion to fixed weights, this problem would not arise.

In DBA for QoS, each ONU is assigned guaranteed bandwidth in proportion to its SLA. More specifically, let B_{total} denote the total upstream bandwidth. Let w_i denote the weighing factor for ONU i. The weighing factors are set in proportion to the SLA of ONU i such that the weighing factors of all ONUs supported by the OLT sum to one, that is, $\Sigma_i w_i = 1$. ONU i is then assigned the guaranteed bandwidth $B_i = B_{total} \cdot w_i$. Note that the sum of all guaranteed bandwidths equals the total available bandwidth. In other words, the total upstream bandwidth is divided among the ONUs in proportion to their SLAs.

For every transmission grant cycle, each ONU requests bandwidth corresponding to its total backlog. If the requested bandwidth is smaller than the guaranteed bandwidth, the difference (i.e., excess bandwidth) is pooled together with the excess bandwidth from all other lightly loaded ONUs (ONUs whose requested bandwidth is less than their guaranteed bandwidth). This pooled excess bandwidth is then distributed to each of the highly loaded ONUs (ONUs whose requested

Method	Bandwidth requests/grants	Dynamic bandwidth allocation
IPACT [4]	Single buffer reporting size at instant of grant (used for next grant)	Variable cycle length, statistical multiplexing
Control theoretic extension of IPACT [6]	Single buffer reporting an estimate of buffer size at time of grant	N/A
Bandwidth guaranteed polling [7]	ONU reports intended utilization of grant	Fixed cycle length and fixed size grant units. Reservation-based with statistical multiplexing of unreserved as well as unused portion of a reservation
DEB-GPS [8]	Deterministic effective bandwidth	Generalized processor sharing
DBA for multimedia [9]	Three priority buffers reported	Fixed bound on cycle length, priority scheduling at the PON level
DBA for QoS [10]	Three priority buffers reported, one-step prediction of traffic arrivals	Fixed bound on cycle length, limited allocation with statistical multiplexing of unused portion of guaranteed bandwidth

TABLE 1. An overview of main features of the bandwidth request/grant mechanisms and DBA algorithms in surveyed DBA schemes.

bandwidth is larger than their guaranteed bandwidth) in a manner that weighs the excess assigned in proportion to the size of their request. Note that this proportional scheduling approach is in contrast to the strict priority scheduling of DBA for multimedia, which does not allocate any bandwidth to lower-priority traffic classes until the bandwidth demands of all higher-priority traffic classes are met.

We note that DBA for QoS allows the option of sending individual priority queue occupancies to the OLT via REPORT messages (a REPORT message supports reporting queue sizes of up to 13 queues) and having the OLT generate transmission windows for each individual priority queue (the GATE message supports sending up to four transmission grants). This option puts the priority scheduling that would otherwise be handled bv the ONU under the control of the OLT.

DBA for QoS [10] also considers the option of reporting queue size using an estimator for the occupancy of the high-priority queue. The estimator makes a one-step prediction of the traffic arriving to the high-priority queue between the time of the report and the time of the grant. In particular, the amount of traffic arriving to the high priority queue between report and grant in a cycle $n - 1$ is used to estimate the arrival in cycle n. The ONU then reports in cycle n the actual backlog at the time of request plus the estimated new arrivals until the time of the grant.

The simulations reported in [10] compare the average and maximum delays for the proposed DBA for QoS scheme for the service classes best effort, assured forwarding, and expedited forwarding with the delays achieved with a static bandwidth allocation to the individual ONUs. It was found that the proposed DBA for QoS scheme achieves significantly smaller delays, especially at high loads. This is primarily due to the statistical multiplexing between the different ONU permitted by DBA for QoS. It was also found that the proposed DBA for QoS scheme is quite effective in differentiating the delays for the different service classes, with the highest-priority expedited forwarding class achieving the smallest delays. The simulations in [10] also considered the average utilization of the upstream bandwidth and found that the proposed DBA for QoS schemes achieve around 90 percent utilization compared to around 50 percent with static bandwidth allocation.

Conclusion

We have provided a comprehensive survey of the schemes developed to date for allocating the bandwidth in the multipoint-to-point (i.e., ONUs to OLT) part of Ethernet passive optical networks. In Table 1 we give an overview of the surveyed schemes for dynamic bandwidth allocation.

We have found that the developed DBA schemes address two key challenges:

- Accommodating traffic fluctuations
- Providing QoS to the traffic

The first challenge has been addressed by estimating the amount of traffic arriving between two successive transmission grants to an ONU using a proportional control in the IPACT enhancement [6] and a one-step prediction in DBA for QoS [10]. One interesting avenue for future work is to develop and evaluate more sophisticated prediction mechanisms for the DBA, such as mechanisms employing more complex control mechanisms or forecasting methods.

The second challenge has been addressed by a wide variety of QoS mechanisms that have generally been adapted from the QoS mechanisms developed for the Internet. In particular, the approaches proposed for providing QoS to the upstream EPON traffic fall into the two main categories of absolute assurances (following the integrated services paradigm in the Internet) and relative assurances (following the differentiated services paradigm in the Internet), as illustrated in Fig. 3. The absolute assurance mechanisms are designed to provide lossless service with deterministic delay bounds. On the other hand, the relative assurance mechanisms differentiate among QoS levels provided to the different traffic classes; that is, higher-priority traffic classes are provided with better QoS relative to the lower-priority classes. An important avenue for future work appears to be the development of QoS mechanisms that provide absolute *statistical* QoS assurances, such as ensuring that a specific delay bound is violated (or traffic lost) with a minuscule prespecified probability. Such statistical QoS assurances appear to be especially well suited for multimedia streaming traffic, which requires strict timing constraints due to the periodic playout deadlines of the multimedia content. On the other hand, multimedia applications typically tolerate if a minuscule amount of the data does not arrive in time as the resultant very minor degradations (e.g., in the video or audio) go typically unnoticed.

Yet another exciting avenue for future work is the development of modified EPON architectures and corresponding DBA algorithms. In [11], for instance, an additional splitter is employed to reflect the transmission of a given ONU back to all ONUs (including the sending ONU) as well as forward the signal to the OLT. This architecture allows each ONUs to keep track of the transmissions by the other ONUs and allows for the development of distributed DBA algorithms. Finally, in the future the currently costly WDM equipment may become affordable for EPONs and may open up a new design space for DBA algorithms.

ACKNOWLEDGMENT

We are grateful to the three anonymous reviewers, whose thoughtful feedback helped us to greatly improve the value of this article.

REFERENCES

[1] G. Kramer, B. Mukherjee, and A. Maislos, "Chapter 8: Ethernet Passive Optical Networks," S. Dixit, Ed., *IP over WDM: Building the Next Generation Optical Internet*, Wiley, 2003.

[2] 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.

[3] K. S. Kim, "On the Evolution of PON-based FTTH Solutions," *Info. Sci.*, vol. 149, no. 1–3, Jan. 2003, pp. 21–30.

[4] 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.

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

[6] H.-J. Byun, J.-M. Nho, and J.-T. Lim, "Dynamic Bandwidth Allocation Algorithm in Ethernet Passive Optical Networks." *Elect. Lett.*, vol. 39, no. 13, June 2003, pp. 1001–02.

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

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

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

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

[11] C. H. Foh *et al.*, "Full-RCMA: A High Utilization EPON," *Proc. OFC*, vol. 1, Mar. 2003, Atlanta, GA, pp. 282–84.

BIOGRAPHIES

MICHAEL MCGARRY (michael.mcgarry@asu.edu) received a B.S. degree in computer engineering from Polytechnic University, Brooklyn, New York, in 1997. He is currently pursuing an M.S.E.E. degree at Arizona State University, Tempe. His research interests are in the areas of Internet QoS and MAC protocol design for optical networks.

MARTIN MAIER (martin.maier@cttc.es) was educated at the Technical University Berlin, Germany, and received Dipl.-Ing. and Dr.-Ing. degrees (both with distinctions) in 1998 and 2003, respectively. Currently, he is a research associate at CTTC, Barcelona, Spain, and focuses on evolutionary WDM upgrades of optical access and metro networks. He is the author of the book *Metropolitan Area WDM Networks — An AWG Based Approach* (Kluwer, 2003).

MARTIN REISSLEIN (reisslein@asu.edu) is an assistant professor in the Department of Electrical Engineering at Arizona State University, Tempe. He received his Ph.D. in systems engineering from the University of Pennsylvania in 1998. 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.263 encoded video, at http://trace.eas.asu.edu.

STANDARDS REPORT

/*from page S7*

LONG AND EXTENDED REACH INTEROPERABILITY

The OIF PLL Working Group has begun a project to ensure multivendor interoperability for optical span lengths in excess of 80 km. Currently available specifications do not address a methodology to guarantee interoperability for reaches with residual dispersion of 1600 ps/nm or higher. Similarly, the ITU currently does not define optical performance for span lengths over 80 km. OIF is working to define methods to help equipment manufacturers ensure interoperability between multiple suppliers' products at these extended reaches. In addition, OIF is coordinating with ITU to ensure that OIF activities align and are nonoverlapping with ITU work.

SUMMARY

This article provided an overview of the organization of the OIF and its technical working groups. Details were provided on both the completed work efforts of the PLL Working Group and current work in progress. The detailed specifications for the completed work of the Optical Internetworking Forum are publicly available and may be found at www.oiforum.com/public/impagreements.html

Activities currently in progress in the electrical domain include the CEI electrical specification and the CEI Protocol project. In the optical domain , two new tunable laser specifications, ITLA-MSA and SFFITLA, a unified 10 Gb Ethernet-SONET OC-192c optical specification, and a long/extended reach interoperability document are in development.

The Optical Internetworking Forum is an open international forum. Collaboration and participation of interested parties in the ongoing work efforts is welcomed.

BIOGRAPHY

MIKE LERER (mike@mike-lerer.com) is an independent consultant. Representing Xilinx, he currently holds positions as Chairman of the OIF PLL Working Group, and Chairman of the Network Processing Forum Hardware Working Group. He is active throughout the industry in the development of advanced interface specifications and protocols (IEEE 802.3, RapidIO). Previously, he held the position of technology facilitator with Avici Systems. He is a founder of and was an executive with both Pixelworks Inc. and Small System Design Inc. He is a graduate of Massachusetts Institute of Technology with degrees in both computer science and management.