IPACT with Smallest Available Report First: A New DBA Algorithm for EPON

Swapnil Bhatia and Radim Bartoš
Department of Computer Science
Univ. of New Hampshire
Email: {sbhatia, rbartos}@cs.unh.edu

***Abstract*— Dynamic Bandwidth allocation in Ethernet Passive Optical Networks (EPONs) has been an area of intense research in recent years. Most of the proposed solutions offer clever methods for fair grant sizing, traffic prediction, and prioritized, differentiated services. Barring some work by Kamal et al. and some elements in the scheme proposed by Ma et al., no work has been done on exploring the order of granting (i.e., ONU sequencing) in an EPON. In this paper, we propose an unexplored heuristic for improving the performance of the IPACT scheme with respect to the most important metric: packet delay. In this heuristic, the OLT always grants that ONU which has the Smallest (Available) Reported queue length, First (SARF). Our simulations indicate that our heuristic can improve the delay performance of IPACT by 10-20% (when tested under the gated allocation policy).**

I. Introduction

A. EPON

An *Ethernet Passive Optical Network (EPON)* is a point to multipoint, bidirectional, high rate optical network for data communication. The EPON link is shared by multiple users. Each user connects to the EPON link through a device known as an *Optical Network Unit (ONU)*. Since the link is shared, link use must be centrally arbitrated. This function is performed by a single special device called the *Optical Line Terminator (OLT)*. The direction of communication from the ONUs to the OLT is known as *upstream* direction whereas the direction from the OLT to the ONUs is known as the *downstream* direction. The data rate in each direction is set to 1 Gbps by the IEEE EPON standard [1]. Overall, the link exhibits a star topology with the OLT at the root of the star and the ONUs at the leaves. The EPON link is shared by all users in the upstream direction. The OLT decides which ONU is allowed to transmit data and for how many bytes. The OLT uses a special control message called a *Gate* to grant transmission opportunities to ONUs. Appended to the data traffic, the ONU also transmits a control message containing a *Report* of the number of bytes buffered in its queue, waiting for a subsequent transmission opportunity. The IEEE standard does not specify the actual algorithm to be used for grant allocation and leaves it open for implementation by vendors.

B. Dynamic Bandwidth Allocation

An algorithm implemented in the OLT, which uses *Report* and *Gate* messages to construct a transmission schedule and convey it to the ONUs is known as a Dynamic Bandwidth Allocation (DBA) algorithm. DBA in EPONs has been an area of intense research in recent years. Many solutions have been proposed. The challenge in designing a DBA algorithm lies in developing an algorithm that is practical, simple, efficient and meets service provider requirements. A significant portion of the body of DBA research has focused on fairness in allocating grants [2][3]. The solutions proposed resemble Weighted Fair Queuing in flavor. Another track of research has focused on improving the freshness of the queue information available to the OLT by employing some variant of a traffic prediction algorithm [4][5]. A third track has focused on minimizing the idle periods on the upstream channel by clever interleaving of messaging delays with data transmissions [6][7][8]. The distinction between inter- and intra-ONU scheduling has resulted in new suggested solutions for these two portions [9][10][11].

When it comes to public subscriber access networks such as EPONs, applications such as voice and video are the main sources of revenue for operators. Voice and video are very sensitive to delay and video traffic is quite bursty in nature. For this reason, packet delay is considered to be the most important benchmark to measure the efficacy of any proposed DBA algorithm. There are at least two ways to interpret this benchmark. Much of the existing work appears to focus on bounding the inter-service delay at any ONU, i.e., the time until an ONU is serviced again is guaranteed to be bounded. All frame-based schedulers [3][7][12][13] are based on this approach. The issue of grant sizing is treated as a separate question in this approach. It usually remains unclear as to why the particular chosen cycle or frame length and the particular chosen grant sizing heuristic when taken together would yield a low delay DBA strategy. (In fact, schemes based on Packetized Generalized Processor Sharing-based approach provide fairness to all ONUs only at the cost of delaying all ONUs equally.) In our opinion, a second approach to the DBA problem could focus on minimizing the achievable per packet delay in an EPON. Instead of just bounding the inter-service (also called the “cycle”) time per ONU, the following question seems natural and interesting: What is the minimum per-packet delay that can be achieved under the IEEE EPON architecture? An attempt to provide an answer would likely involve viewing the DBA as some variant of a scheduling problem. Scheduling theory [14], with its rich set of models and results [15], can offer many insights into the basic structure of the DBA problem and in turn shed light on the limits of the performance achievable under the IEEE EPON architecture. Our present paper is not devoted to finding a complete answer to the question posed above; the answer may be difficult to arrive at. However, while such a theoretical line of research may be long and arduous, it can provide valuable ideas based on a solid theoretical foundation for novel and highly effective solutions to the DBA problem. In this paper, we follow this approach and propose one such solution.

II. IPACT WITH SMALLEST AVAILABLE REPORT FIRST (SARF)

The main contribution of this paper is the idea of allocating grants in an order that minimizes packet delay. To this end, we propose the use of the Smallest Available Report First heuristic and through simulations, demonstrate its efficacy in reducing packet delay.

A. IPACT

Kramer et al. proposed the simple Interleaved Polling with Adaptive Cycle Time scheme as a solution for DBA in EPONs [16][17]. One of their main contributions is the observation that an OLT need not wait for a transmission from an ONU to finish, before sending a *Gate* message to the next ONU. The IPACT scheme can therefore transmit downstream control messages to the ONUs while receiving data transmissions from other ONUs in the upstream, thus minimizing upstream underutilization due to walk time. This can be a significant problem in public access networks with high bandwidth and delay products. In addition, Kramer et al. also propose various policies for calculating the size of the grant allocated in response to an ONU's *Report* message. However, the order in which ONUs are serviced is unspecified and is assumed to be round-robin. Note that the order may not necessarily be static, since IPACT may send grants in a different order in any cycle in an attempt to minimize the idle period of the upstream channel due to walk time to a farther ONU [16].

B. IPACT+SARF

We propose a new heuristic for use with the IPACT scheme. In fact, our heuristic is quite simple and independent of IPACT, and therefore could be used with any DBA scheme.

Figure 1 shows the SARF heuristic. SARF generates and responds to two events: $SEND\_GRANT$ and $REPORT_i$. A $REPORT_i$ event occurs when a *Report* message is received from ONU i. A $SEND\_GRANT$ event is generated by the SARF algorithm in step (4). Although our heuristic may be used with many different DBA schemes, we illustrate its use with IPACT in this paper.

Initially, the OLT sends out, to all active ONUs, grants large enough to transmit a *Report* message (not shown in Figure 1). Next, whenever the OLT receives a *Report* message from ONU i, it first updates entry i in its table of current, known queue lengths and also marks ONU i as "pending". (A pending ONU is one which has transmitted a single new *Report* message, but has not yet been serviced.) Next, unlike plain IPACT, the SARF heuristic does not send out a grant to ONU i immediately in response to a report received from ONU i (except under a special condition described later). Instead, it defers grant transmission to the latest possible time. Suppose the *Report* from ONU i is received at time t. Let $S(t)$ denote the earliest time at which the upstream channel is known to become available as of time t. Then, SARF defers the grant transmission to a time t_g such that if a grant is transmitted at time t_g to ONU i, it will cause the transmission from ONU i to arrive at the OLT exactly at time $S(t)$. Thus, SARF defers grant transmission to the latest time possible, without introducing any extra idle time (i.e., it maintains the "work-conserving" property of IPACT schedules). Clearly, a grant must be transmitted to an ONU sufficiently early so as to allow for the grant message transmission delay as well as the round-trip delay to the ONU. This is accomplished in step (4) of Figure 1. However, it may so happen that when a *Report* message is received at time t, $S(t) \leq d_i$ where d_i is the round-trip time to ONU i. In this case the grant message cannot be deferred since doing so would introduce unnecessary idle time (thus violating the "work conserving" property[1]). In this special case, the grant to ONU i is sent immediately.

When a $SEND\_GRANT$ event is triggered, the SARF heuristic is used to determine the sequence in which ONUs will be served. Of all the pending ONUs, the one with the smallest queue length is selected. Using a *Gate* message, the ONU is served a transmission grant. (Notice that the heuristic is independent of the policy used for deciding the grant size.) The status of the serviced ONU is changed from pending to "served". The ONU with the next smallest queue length is served next. This process continues until all pending ONUs have been served, exactly like the plain IPACT scheme. This completes one cycle of service and the same behavior is repeated for the next cycle. This is the basic description of the SARF heuristic when used in combination with the IPACT scheme.

C. Zero-length queues

Although the SARF heuristic chooses the ONU with the smallest queue length for service, it treats ONUs with zero queue lengths differently. An ONU with a zero queue length has no data to transmit. Under low loads, without special exception, such ONUs will always be served first. Serving zero-length queues still requires the allocation of a grant to accommodate the next *Report* message as well as the mandatory guard band overhead. (Guard bands are small intervals of time inserted between two consecutive grants to two different ONUs to prevent any possible overlap of transmissions due to small errors in time synchronization at those ONUs.) While such data-less grants do not contribute to lowering the packet delay at the source ONUs which have no packets to send, they do increase the delay faced by the succeeding ONUs. Hence, when choosing an ONU to serve next, ONUs with zero-length queues are treated as if they have a queue length that is equal to the average queue length taken over all ONUs. Moreover, we weigh this average by the number of times an ONU reports a zero length queue, consecutively. Thus, an ONU which has

[1] We note that violating the work-conserving property may not necessarily be a bad idea [18].

TABLE I
TERMINOLOGY USED IN THE IPACT+SARF ALGORITHM.

Term	Description
SEI	Scheduling Endpoint Indicator signifying the earliest time at which a new transmission can be scheduled on the upstream
$REPORT_i$	Event indicating receipt of a *Report* message from ONU i
$SEND\_GRANT$	Event signifying the latest opportunity to make a grant decision
MAX_RTT	Largest round-trip time among all connected ONUs in seconds
$GATE\_LENGTH$	Duration of a *Gate* message in seconds

reported a zero length queue many times consecutively will likely be served at the end of the cycle.

D. SARF Rationale

The delay faced by any packet p in the EPON consists of three components: the reporting delay, the grant delay and the intra-grant delay. The reporting delay is the time between the arrival of a packet p at an ONU and the time at which it is counted and reported by the ONU to the OLT. The grant delay is the time between reception of a report by the OLT and the reception of the first bit of data associated with the report by the OLT (i.e., beginning of the actual grant). The intra-grant delay is the time packet p at ONU i waits after the beginning of a grant for ONU i for other preceding packets to be transmitted. Notice that in IPACT, the grant delay depends on the distance of the scheduling endpoint from the current time. As per the SARF heuristic, if the OLT chooses to serve the smallest request, it will increase the SEI by the smallest possible value. Thus, the SARF heuristic minimizes the grant delay faced by packets. In turn, the average cycle length may also be reduced thus leading to a reduction in the reporting delay. We observe that the main result, that for identical release times, the Shortest Processing Time rule provides the minimum completion time, is well-known in the area of scheduling theory [14]. We leverage this knowledge and successfully apply it to the DBA problem. However, we also note that this may *not* be the minimal achievable delay under the IEEE EPON architecture; we are currently investigating more sophisticated approaches to this problem.

III. RELATED WORK

We note that to our knowledge, the work by Kamal et. al [19] was the first to realize the value of deferring a grant decision in order to benefit from more information. In their Prioritized MPCP scheme, low priority grants are deferred to the latest possible time (without introducing extra idle time) allowing higher priority grants to be scheduled earlier. In this case, the order of granting is changed to enforce priority, not to minimize delay. However, Prioritized MPCP does suffer from other shortcomings such as priority inversion under certain load conditions as well as unfairness based on round-trip time to different ONUs.

Ma et al. [20] also propose a Dynamic Polling Order Arrangement (DPOA). However, for reasons not discussed in the paper, they order ONUs in descending order of the reported queue length. Their scheme (IPACT with DPOA) shows better performance than IPACT alone in the specific range of medium loads (0.5 to 0.8). The authors reason that the improvement in delay is due to better channel utilization. They also conclude that at high loads, since the channel is already heavily utilized, the polling order does not matter. However, their reasoning only holds for the conditions of their experiments (limited grants[2] and Poisson traffic), which are different from ours (gated unlimited grants and self-similar traffic).

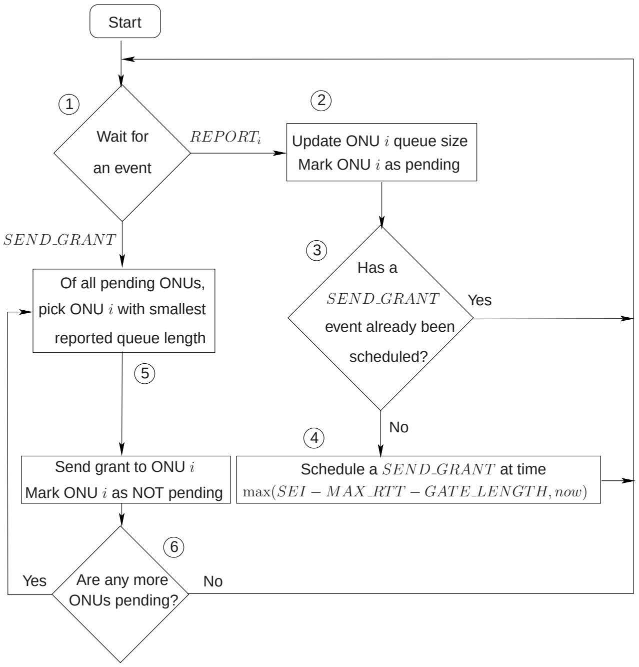

Fig. 1. The SARF algorithm.

We also note that the idea of differentiating between small and large requests (queue lengths) is also present in the work of Assi et al. [3]. However, the motivation for this in their scheme is slightly different (i.e., to maximize the overlap of walk time with transmissions in progress so as to minimize channel idling). We generalize this idea of differentiation based on request size in the sense that in the SARF heuristic, the size of the request (queue length) determines the exact position of an ONU in the polling sequence.

IV. SIMULATION RESULTS

A. Simulation Details

We implemented the SARF heuristic with the IPACT scheme under gated-allocation policy, i.e., the OLT always allocates a grant exactly equal to the reported queue length with a fixed additional amount to accommodate the next *Report* message. In our simulation, we distributed the total load λ randomly across the $N \in \{4, 8, 16, 32\}$ ONUs. We achieved this using the following method: Given a total load $0 \leq \lambda \leq 1$, first pick uniformly at random, $N-1$ non-decreasing real numbers $0 \leq r_i \leq \lambda$, $i \in \{1, \cdots, N-1\}$. Then, assign

[2] Their paper [20] does not provide details about the grant policy used in their simulation experiments.

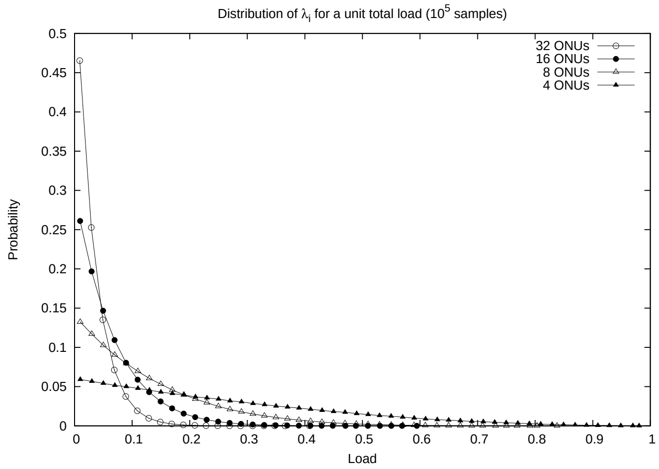

Fig. 2. The marginal probability of an ONU load as measured from simulations.

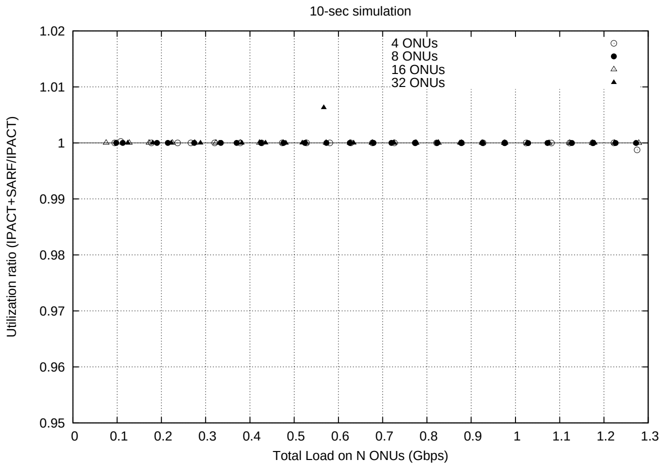

Fig. 4. The utilization ratio of IPACT+SARF/IPACT.

the value $\lambda_i = r_i - r_{i-1}$ as the load for ONU i, with $r_0 = 0$ and $r_N = \lambda$. In this way, we assigned a random and hence nonuniform load to the N ONUs in each experiment. Generated in the manner described above, the random load λ_i on any ONU follows a Beta distribution $\beta(1, N)$ [21]. Figure 2 shows an example of the probability distribution of individual ONU loads when total load $\lambda = 1$. As illustrated, the chance of generating a load λ_i at ONU i is highly skewed in favor of lower loads. The traffic workloads for simulations were generated using a self-similar model with a measured Hurst parameter of approximately 0.8 [22]. Each simulation was allowed to run for 10 seconds of simulation time. With self-similar traffic, simulations should be run for a much longer duration. However, self-similar trace generation and the simulations themselves are both computationally intensive tasks. Therefore, we report preliminary results with shorter simulations. Longer, more rigorous simulations are currently in progress. At least 40 simulation runs were conducted for each 0.05-length interval of load. (The exact number is difficult to fix due to the inherent error in generating the exact load with a self-similar traffic generator for a relatively short trace.) The guard band was set to 2 μs.

B. Results

Figure 3 shows the relative performance of IPACT with and without the proposed SARF heuristic. We plot the actual total load on the x-axis. On the y-axis, we plot the relative reduction in the average per-packet delay. This is calculated as

$$\Delta = \frac{\delta_{IPACT} - \delta_{SARF+IPACT}}{\delta_{IPACT}}, \tag{1}$$

where δ_{IPACT} is the delay of the plain IPACT scheme and $\delta_{IPACT+SARF}$ is the delay of the IPACT scheme combined with the SARF heuristic. Note that in each experiment, the traffic trace used as workload to evaluate the IPACT scheme was the same as the trace used to evaluate the IPACT+SARF scheme (i.e., IPACT was run with and without the SARF heuristic on the same trace). The "scatterplot" of points shows all actual measurements of the reduction in the delay, i.e., it shows 40 points per 0.05-long interval of load, each being a result of a 10-second long simulation. Each of these points is plotted at the actual load generated by the trace. For example, if the intended target load was l and the actual load generated by the trace generator was $l + \epsilon$, then, the relative reduction in delay measured (say r) is plotted as $(l + \epsilon, r)$ and not (l, r). Here, ϵ can be considered as the error in generating the required load. Since ϵ is unpredictable, there is no simple way of plotting an averaged curve that captures the main trend. Therefore, we "binned" the delay reduction measurements as follows. We selected a bin size b ($b = 0.05$ in case of Figure 3) and created bins (intervals) B_k of the form $[b{\cdot}k, b{\cdot}(k{+}1)), k \geq 0$. Any measurement $(l + \epsilon, r)$ was dropped into the bin B_k where $k = \lfloor (l + \epsilon)/b \rfloor$. All measurements in each bin were averaged and the resulting average difference r_k was plotted as $(\hat{l}, r_k)$ where $\hat{l}$ is the average of all the loads in bin B_k. Clearly, our proposed SARF heuristic shows significant improvement over plain IPACT across most load values. For very low loads, it is likely that service sequence will not make a difference since the channel may be underutilized. Other cases where the SARF heuristic performs worse than plain IPACT may be explained by the observation that the SARF heuristic may result in many more, but smaller grants resulting in a somewhat larger overhead in the form of guard bands and *Report* messages. Figure 4 shows that the relative channel utilization of the IPACT+SARF scheme is very close to that of the original IPACT scheme.

V. Fairness

As discussed in Sec. I-B, the issue of fairness in bandwidth allocation to ONUs in an EPON has received much attention in EPON DBA research [2][3][11]. In the proposed SARF algorithm, the choice of the next ONU to serve is guided solely by the reported queue size. Thus, an ONU with a smaller reported queue size may always be served before an ONU with a larger queue. This may be unfair according to some definitions of fairness [23]. Figure 5 illustrates this unfairness manifesting in the form of increased variance in the average of the average packet delay faced by the N ONUs

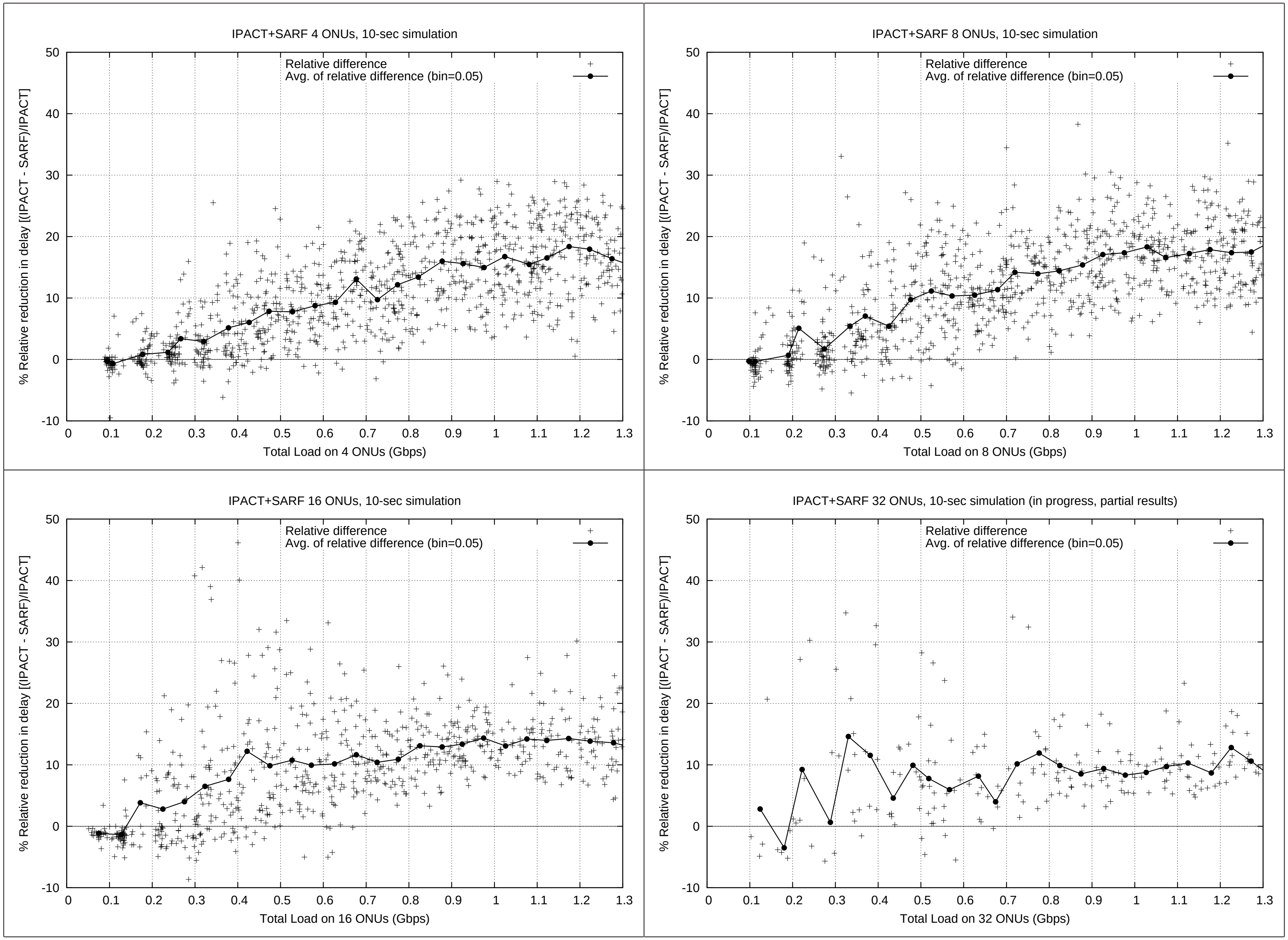

Fig. 3. Comparison of mean packet delay of IPACT against that of IPACT combined with the proposed SARF heuristic.

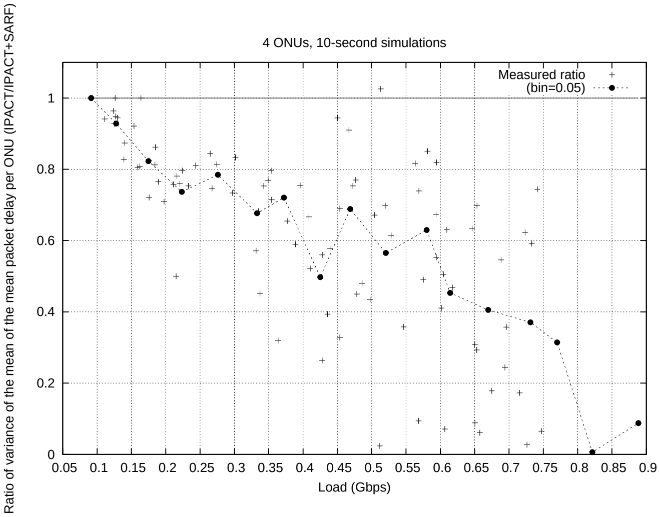

Fig. 5. The ratio (IPACT/IPACT+SARF) of variance of the average (taken over all ONUs) of the average packet delay (taken over all packets per ONU) measured by a total of 50 10-second long simulations (preliminary results for 4 ONUs only).

under the SARF heuristic as compared to IPACT. To this criticism, we offer the following two responses. First, our approach is in stark contrast to other DBA approaches which seek to include fairness as one of the main requirements of the DBA algorithm. In our approach in this paper, this is not our main concern. Instead, we focus on developing a DBA algorithm which, *by design*, attempts to minimize the overall average packet delay in the EPON. If that entails treating ONUs unfairly, then we allow the algorithm to make this intelligent decision. We believe that this is the main novelty of our approach to DBA algorithm design. Second, an element of fairness can be incorporated into our algorithm in a simple way. Instead of choosing the smallest report, one could easily choose the weighted smallest report and use the weights to provide fairness to ONUs.

VI. Conclusion and Future Work

The main contribution of this paper is the idea of exploiting the order in which grants are allocated to ONUs to minimize per packet delay. We proposed a new heuristic for the DBA problem which allocates grants using the Shortest Available Report First strategy. Our heuristic is independent of the actual DBA scheme and may be used in other DBA algorithms. We demonstrated its effectiveness using IPACT under the gated allocation policy. Our heuristic improves the delay performance of IPACT by about 10-20%. The SARF heuristic shows considerable promise and may be extended and improved. This work is currently in progress. An idea similar to SARF may be useful in discovering the optimal DBA algorithm possible under the IEEE EPON architecture. We are currently exploring these ideas.

Acknowledgments

We express our sincere gratitude to Cisco Systems for supporting our work under the Cisco University Research program. We thank Dr. G. Kramer for making the self-similar synthetic traffic generator publicly available.

References

[1] *IEEE Standard for Local and Metropolitan Area Networks, Part 3: CSMA/CD Access Methods and Physical Layer Specifications (802.3ah)*, IEEE Standard, 2005.

[2] X. Bai, A. Shami, and C. Assi, “Statistical Bandwidth Multiplexing in Ethernet Passive Optical Networks,” in *IEEE Global Telecommunications Conference (GLOBECOM)*, St. Louis, MO, USA, Nov. 2005.

[3] C. M. Assi, Y. Ye, S. Dixit, and M. A. Ali, “Dynamic Bandwidth Allocation for Quality-of-Service Over Ethernet PONs,” *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 9, pp. 1467–1477, Nov. 2003.

[4] 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, pp. 1001–1002, June 2003.

[5] Y. Luo and N. Ansari, “Limited Sharing with Traffic Prediction for Dynamic Bandwidth Allocation and QoS Provisioning over Ethernet Passive Optical Networks,” *Journal of Optical Networking*, vol. 4, no. 9, pp. 561–572, 2005.

[6] J.-S. Kim, H.-J. Yeon, and J. Lee, “HUHG: High Utilization and Hybrid Granting Algorithm for EPON,” in *International Conference on Communications (ICC)*, 2006.

[7] X. Bai, A. Shami, N. Ghani, and C. Assi, “A Hybrid Granting Algorithm for QoS Support in Ethernet Passive Optical Networks,” in *International Conference on Communications (ICC)*, May 2005, pp. 1869–1873.

[8] A. Shami, X. Bai, C. Assi, and N. Ghani, “Jitter Performance in Ethernet Passive Optical Networks,” *IEEE/OSA Journal of Lightwave Technology*, vol. 23, no. 4, pp. 1745–1753, Apr. 2005.

[9] N. Ghani, A. Shami, C. Assi, and M. Y. A. Raja, “Quality of Service in Ethernet Passive Optical Networks,” in *IEEE Sarnoff Symposium*, 2004.

[10] ——, “Intra-ONU Bandwidth Scheduling in Ethernet Passive Optical Networks,” *IEEE Communications Letters*, vol. 8, no. 11, pp. 683–685, Nov. 2004.

[11] G. Kramer, A. Banerjee, N. K. Singhal, B. Mukherjee, S. Dixit, and Y. Ye, “Fair Queuing with Service Envelopes (FQSE): A Cousin-Fair Hierarchical Scheduler for Subscriber Access Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 22, no. 8, pp. 1497–1513, Oct. 2004.

[12] M. Ma, Y. Zhu, and T. H. Cheng, “A Bandwidth Guaranteed Polling MAC Protocol for Ethernet Passive Optical Networks,” in *IEEE INFOCOM*, 2003.

[13] S. il Choi and J. doo Huh, “Dynamic Bandwidth Allocation Algorithm for Multimedia Services over Ethernet PONs,” *ETRI Journal*, vol. 24, no. 6, pp. 465–468, Dec. 2002.

[14] R. G. Parker, *Deterministic Scheduling Theory*. Chapman and Hall, 1995.

[15] K. Pruhs, E. Torng, and J. Sgall, “Online Scheduling,” in *Handbook of Scheduling: Algorithms, Models and Performance Analysis*, J. Y. T. Leung, Ed. CRC Press, 2004, ch. 15.

[16] G. Kramer, B. Mukherjee, and G. Pesavento, “Interleaved Polling with Adaptive Cycle time (IPACT): A Dynamic Bandwidth Distribution Scheme in an Optical Access Network,” *Photonic Network Communications*, vol. 4, no. 1, pp. 89–107, Jan. 2002.

[17] G. Kramer, B. Mukherjee, S. Dixit, Y. Ye, and R. Hirth, “Supporting Differentiated Classes of Service in Ethernet Passive Optical Networks,” *Journal of Optical Networking*, vol. 1, no. 8, pp. 280–298, Aug. 2002.

[18] J. A. Hoogeveen and A. P. A. Vestjens, “Optimal On-line Algorithms for Single-Machine Scheduling,” in *Proc. of the 5th Conference on Integer and Combinatorial Optimization*, 1996.

[19] A. Kamal and B. Blietz, “A Priority Mechanism for the IEEE 802.3ah EPON,” in *International Conference on Communications (ICC)*, May 2005.

[20] M. Ma, L. Liu, and T. H. Cheng, “Adaptive Scheduling for Differentiated Services in an Ethernet Passive Optical Network,” *Journal of Optical Networking*, vol. 4, no. 10, pp. 661–670, Oct. 2005.

[21] H. A. David and H. N. Nagaraja, *Order Statistics*. John Wiley, 2003, ch. 2, pp. 13–14.

[22] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, “On the Self-Similar Nature of Ethernet Traffic,” *IEEE/ACM Trans. Networking*, vol. 2, no. 1, pp. 1–15, Feb. 1994.

[23] X. Bai, A. Shami, and C. Assi, “On the Fairness of Dynamic Bandwidth Allocation Schemes in Ethernet Passive Optical Networks,” *Elsevier Journal of Computer Communications*, vol. 29, no. 11, July 2006.