QoS-Aware Cooperative Bandwidth Allocation for Mobile Fronthaul Traffic in TDM-PON

Xiang Lu*, Luis Velasco, and Marc Ruiz
Optical Communications Group (GCO), Universitat Politècnica de Catalunya (UPC), Barcelona, Spain;
**e-mail: xiang.lu@upc.edu*

ABSTRACT
Emerging mobile fronthaul services introduce new challenges for traffic scheduling in time-division multiplexing passive optical networks (TDM-PON), as they impose stringent deterministic latency requirements and high reliability demands. We propose a novel QoS-aware Co-DBA scheme for TDM-PON-based fronthaul to better satisfy the various latency requirements of fronthaul traffic flows. Through simulations, we investigate the conditions under which deterministic latency guarantees can be achieved for mobile fronthaul traffic under different traffic loads.

Keywords: Time-division multiplexing passive optical network, Mobile fronthaul, Deterministic latency.

1. INTRODUCTION

Time-division multiplexing passive optical networks (TDM-PON) have been widely deployed for fixed broadband access due to their cost efficiency and high capacity. In recent years, both the line rates and split ratios of TDM-PON systems have continued to increase, and they are expected to support 100 or even 200Gbit/s per wavelength in the near future [1]. Consequently, future TDM-PON systems will extend far beyond supporting a single fiber-to-the-home (FTTH) scenario with typical per-household data rates around 1 Gbit/s [1], and to support more complex deployment scenarios, such as mobile network fronthauling [1] [2].

To further reduce deployment costs and improve bandwidth efficiency, radio access networks (RAN) has long considered TDM-PON as a promising transport solution for mobile fronthaul. In particular, since the cooperative transport interface (CTI) was incorporated into the O-RAN specifications, cooperative dynamic bandwidth allocation (Co-DBA) scheme has become a key technique for reducing latency in PON-based 4G and 5G mobile fronthaul networks, and has attracted extensive research attention [3] [4]. Uzawa et al. proposed a network slicing-based Co-DBA scheme that schedules mobile fronthaul traffic with IoT traffic while mitigating the latency impact of PON discovery windows [5]. To enable compatibility with mini-slot-based 5G new radio interface, Zhang et al. proposed a dynamic wavelength and bandwidth allocation scheme to alleviate the high latency arising from scheduling conflicts between fixed access and mobile fronthaul traffic caused by the introduction of mini-slots [6]. In addition, Ciceri et al. proposed a low latency mobile fronthaul resource allocation algorithm aimed at improving bandwidth utilization across multiple PON tenants [7].

Although these schemes can maintain the average fronthaul latency at a low level, future 6G mobile networks are expected to pursue not only low latency but also stronger latency determinism. However, the provision of deterministic delay bounds has been largely overlooked in TDM-PON based fronthaul. For instance, compared with 5G, the fronthaul latency bound in 6G is expected to be reduced by more than 10 times, reaching below 100 μs, while the reliability is expected to improve by 100 times [2]. Furthermore, fronthaul flows may exhibit various latency requirements, as the physical distance between the radio unit (RU) and the distributed unit (DU) affect the fronthaul latency requirement of each fronthaul traffic.

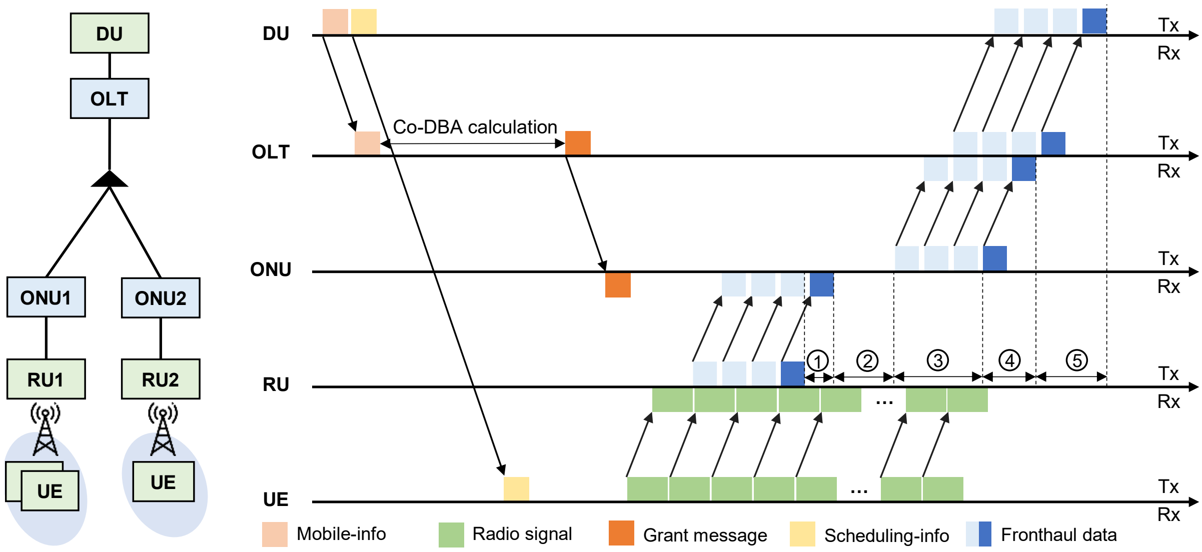
To address the limitation, this paper proposes a new QoS-aware Co-DBA scheme for the TDM-PON-based fronthaul. The proposed scheme classifies services according to their diverse latency requirements and allocates bandwidth accordingly. Simulation results exhibit that the proposed scheme can more effectively satisfy the latency requirements of different fronthaul traffic flows, thereby improving service reliability.

2. TDM-PON-BASED FRONTHAUL AND COOPERATIVE DBA

Figure 1 illustrates a general TDM-PON-based fronthaul architecture and the corresponding Co-DBA procedure. Based on the upstream transmission requests from the user equipment (UE), DU first generates two types of messages related to the upcoming mobile traffic. The scheduling-info message is delivered to the optical line terminal (OLT), while the mobile-info message is sent to the UE. In this way, the OLT can start assigning upstream bandwidth for the optical network units (ONUs) in advance. After receiving the message, the UEs uploads the traffic to the RUs within the assigned mobile slots. The traffic is then processed by the RUs and forwarded to the ONUs. Meanwhile, the ONUs have already received grant messages from the OLT, enabling them to send the fronthaul traffic in time without further bandwidth negotiation with the OLT. As a result, the latency of fronthaul traffic in PON can be significantly reduced compared with that under conventional scheduling schemes.

Under the Co-DBA mechanism, the upstream latency of a fronthaul packet in such a system can be expressed as

979-8-3195-4420-9/26/$31.00 ©2026 IEEE.



(a) (b)

Fig. 1 (a) TDM-PON based fronthaul architecture and (b) Co-DBA procedure.

$$T_{FH} = T_{RU-ONU} + T_{align} + T_{queue} + T_{trans} + T_{OLT-DU}. \qquad (1)$$

where (1) $T_{RU\text{-}ONU}$ denotes the transmission and propagation delay from the RU to the ONU; (2) T_{align} denotes the PON alignment time caused by the offset between the packet arrival time at the ONU and the start of its granted time slot; (3) T_{queue} denotes the queueing delay experienced by the packet inside the ONU before being transmitted in the allocated time slot; (4) T_{trans} denotes the transmission delay between ONU and OLT; and (5) $T_{OLT\text{-}DU}$ denotes the transmission and propagation delay from the OLT to the DU. Since the RU and the ONU are typically deployed close to each other, $T_{RU\text{-}ONU}$ can be neglected. Similarly, the delay from the OLT to the DU, $T_{OLT\text{-}DU}$, is also negligible. Therefore, the upstream latency is mainly determined by the delay components inside the PON, i.e., T_{align}, T_{queue}, and T_{trans}. Among them, T_{trans} is unavoidable, as it includes propagation and transmission time. Consequently, the key to guaranteeing fronthaul latency lies in reducing T_{align} and T_{queue}.

3. PROPOSED SCHEME

The proposed QoS-aware Co-DBA scheme operates in four main steps: (1) The OLT collects mobile-info messages from the DU; (2) Based on the messages, upstream packets that are expected to arrive at the ONU are grouped into buckets to facilitate scheduling; (3) Upstream bandwidth is allocated to ONU based on the constructed buckets; (4) Remaining bandwidth is allocated to non-fronthaul traffic. The pseudo-code for the proposed QoS-aware Co-DBA algorithm is shown in **Algorithm 1**.

In the first step, the OLT collects upstream traffic information from the mobile-info messages, which specifies the expected arrival times, packet sizes and maximum delay requirement of fronthaul traffic within upcoming mobile slots. Different FH traffic flows will be queued separately at the buffer of the corresponding ONU, thus the OLT can allocate bandwidth on a per-flow basis.

In the second step, packets belonging to the same flow f are grouped and represented as buckets. For each flow, the OLT checks the expected arrival time of the packets in sequence and assigns consecutive packets to the same bucket as long as the inter-arrival time between any two adjacent packets does not exceed a predetermined threshold. All packets of the flow are associated with corresponding buckets. In this way, the scheduling complexity can be significantly reduced. For each bucket b, its arrival time a_b, size s_b, and maximum tolerable latency d_b are recorded, where a_b equals the arrival time of its last packet, and s_b denotes the total size of all packets contained in the bucket.

In the third step, the OLT schedules fronthaul traffic at the bucket level. Assume that U denotes the set of all buckets required scheduling. To begin with, the bucket with the earliest arrival time is selected as the first bucket to be served, and the OLT assigns time slot to it based on the available bandwidth R in current polling cycle. The start time of the allocated time slot θ_b will be as close as possible to the arrival time a_b. After allocating bandwidth for the first bucket, the OLT selects the next bucket from U according to a novel online bucket selection rule, which allows buckets with different delay requirements to be scheduled in a flexible manner. Specifically, after the assignment for each bucket, the OLT records the completion time t of the scheduled bucket and identifies the subset of buckets from U whose arrival times are no later than time t. This bucket set is denoted as $C(t)$. If such buckets exist, the next bucket is chosen from $C(t)$ as the one with the earliest deadline among them; otherwise, the bucket with the earliest arrival time is selected from the remaining bucket set U. This rule prioritizes already available buckets with tighter delay constraints while avoiding the selection of buckets that have not yet arrived.

The selected bucket is then allocated bandwidth, removed from U, and the procedure repeats until all buckets in U are scheduled or the available upstream bandwidth in the current cycle is exhausted.

Algorithm 1 QoS-aware Cooperative DBA Algorithm

Input: $B_f=\{b: (a_b, s_b, d_b)\}$, F, T^{guard}, T^{header}, C, GT^{min}
Output: $G_f=\{[\theta_b, \varphi_b]\}$

$U \leftarrow \{b \mid b \in B_f, f \in F\}$
$G_f \leftarrow \Phi$
$bc \leftarrow \mathrm{argmin}_{\{b \in U\}}(a_b)$
while $U \neq \Phi \wedge R \neq \Phi$ **do**
  Allocate time slot $[\theta_i, \varphi_i]$ for bucket i from R
  $G_f \leftarrow G_f \cup [\theta_i, \varphi_i]$
  $t \leftarrow \varphi_i$
  $U \leftarrow U \setminus \{i\}$
  $C(t) \leftarrow \{b \in U: a_b \leq t\}$
  if $C(t) \neq \Phi$ **then**
    $i \leftarrow \mathrm{argmin}_{\{b \in C(t)\}}(a_b + d_b)$
  else
    $i \leftarrow \mathrm{argmin}_{\{b \in U\}}(a_b)$
  Update R
return G_f

In the fourth step, the OLT calculates the remaining bandwidth and then evenly assigns this residual bandwidth among the non-fronthaul traffic of all ONUs by filling the gaps between the already allocated time slots. This process continues until all non-fronthaul traffic has been scheduled or no residual bandwidth remains. These assignments are then recorded in the grant messages and sent to ONU.

4. SIMULATION RESULTS

We evaluate the performance of the proposed QoS-aware Co-DBA scheme using a home-made MATLAB simulator. The simulated system is a 50G PON consisting of 16 ONUs. The distance between each ONU and the OLT is assumed to follow a Gaussian distribution with a mean of 10 km and a variance of 4 km. Among the ONUs, 12 of them are dedicated to mobile fronthaul traffic, while the remaining ONUs carry fixed-network traffic. Each fronthaul ONU serves one RU, and each RU generates one fronthaul flow. The latency constraint of each fronthaul flow is set to 100 µs. Split option 7-2 is considered, with a subcarrier spacing of 30 kHz. For background traffic, conventional FTTH traffic is considered and scheduled in a status-reporting manner. Each FTTH traffic flow is assumed to follow a Poisson process. Packets are modelled as Ethernet frames, with sizes uniformly distributed between 64 and 1518 bytes. As a benchmark, a multi-burst scheduling scheme is adopted for comparison. This scheme divides each upstream PON frame into multiple short subframes, so that every ONU has an opportunity to transmit one burst in each subframe. In this way, the PON alignment time can be significantly reduced [8].

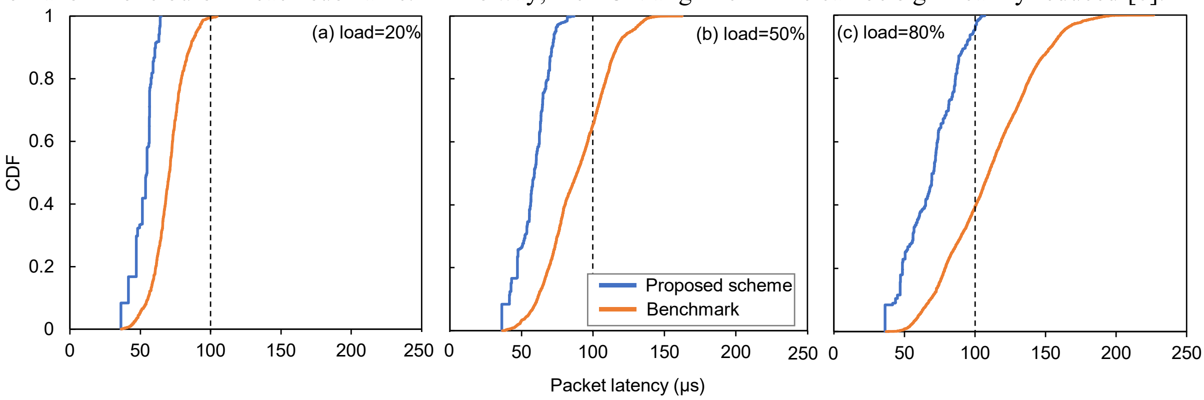


Fig. 2 Cumulative distribution functions of fronthaul packet latency under (a) 20%, (b) 50%, and (c) 80% traffic loads.

Figures 2(a), (b), and (c) show the cumulative distribution functions (CDF) of the packet latency for all fronthaul flows under three traffic loads (20%, 50%, and 80%), respectively. Under a light load of 20%, both the proposed scheme and the benchmark satisfy the latency bound for all packets. However, the proposed scheme still achieves lower average and maximum latency than the benchmark, because the allocated time slots are aligned more closely with the arrival time of each bucket, thereby making T_{align} nearly negligible. Although the benchmark increases the burst transmission frequency significantly, it does not eliminate T_{align} completely. When the load increases to

50%, the benchmark can no longer satisfy the latency requirement for all packets. In addition to T_{align}, another important reason is that the benchmark could not adapt bandwidth allocation according to the latency requirements of different traffic flows. When the load further increases to 80%, the proposed scheme is still able to keep 95% of packets within the latency bound, whereas the benchmark achieves this for only 38% of packets.

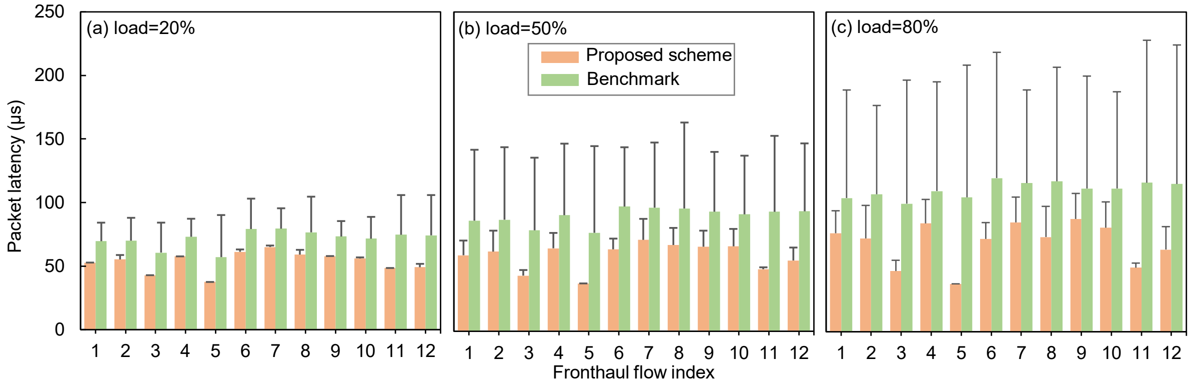


Fig. 3 Average and maximum latency of each fronthaul flow under (a) 20%, (b) 50%, and (c) 80% traffic loads.

Figures 3(a), (b), and (c) show the average and maximum latency of each fronthaul flow under the same three traffic loads. The proposed scheme consistently achieves significantly lower maximum latency than the benchmark under all load conditions. Moreover, for each fronthaul flow, the difference between the average and maximum latency remains low under the proposed scheme. This is because the bandwidth allocation process explicitly takes into account the differences in physical distance among fronthaul flows, thereby enabling better control of their latency variation. By contrast, the benchmark exhibits much larger jitter. This is due not only to the fact that T_{align} cannot be effectively eliminated, but also to the lack of differentiation in scheduling fronthaul traffic according to the actual latency requirements of different flows.

5. CONCLUSION

In this paper, we propose a novel QoS-aware Co-DBA scheme to guarantee fronthaul latency by considering the various latency requirements of fronthaul traffic. The proposed scheme can reduce the dominant delay components inside the TDM-PON and improve latency determinism for mobile fronthaul traffic. Simulation results show that, under different traffic loads, the proposed scheme outperforms existing schemes in terms of both average and maximum latency performance, especially under high-load conditions.

ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Commission MSCA-DN NESTOR (G.A.101119983) and the Horizon Europe ALLEGRO (G.A. 101092766), the project PID2024-157824OB-I00 funded by MICIU/ AEI/10.13039/501100011033 /FEDER, UE and from ICREA.

REFERENCES

[1] G. Rizzelli *et al.*, "High-speed PON solutions: recent evolutions and expected trends [Invited Tutorial]," J. Opt. Commun. Netw. vol. 18, A134-A149 (2026).

[2] A. Fayad *et al.*, "Toward 6G Optical Fronthaul: A Survey on Enabling Technologies and Research Perspectives," IEEE Commun. Surveys Tuts., vol. 27, no. 1, pp. 629-666, 2025.

[3] T. Tashiro *et al.*, "A novel DBA scheme for TDM-PON based mobile fronthaul," in *Optical Fiber Communication Conference* 2014, paper Tu3F.3.

[4] "5G Wireless fronthaul requirements in a passive optical network context, " ITU-T G. Sup66, Rev., Jul. 2019.

[5] H. Uzawa *et al.*, "Dynamic bandwidth allocation scheme for network slicing-based TDM-PON toward the beyond-5G era," J. Opt. Commun. Netw., vol. 12, no. 2, pp. A135–A143, 2020.

[6] J. Zhang *et al.*, "Low latency DWBA scheme for mini-slot based 5G new radio in a fixed and mobile converged TWDM-PON," J. Lightwave Technol., vol. 40, no. 1, pp. 3–13, 2022.

[7] O. J. Ciceri *et al.*, "Resource Allocation in Passive Optical Networks for Low-Latency Mobile Fronthauling Services," IEEE Netw., vol. 39, no. 1, pp. 278-288, 2025.

[8] Y. Weng *et al.*, "Model and Analysis of the Maximum Upstream Latency of a Deterministic Industrial PON System Applying Frame-Based Dense Burst Allocation Method," J. Lightwave Technol., vol. 42, pp. 3210-3220, 2024.