Resource Allocation in Passive Optical Networks for Low-Latency Mobile Fronthauling Services

Oscar J. Ciceri, Carlos A. Astudillo, Gustavo B. Figueiredo, Zuqing Zhu, and Nelson L. S. da Fonseca

***Abstract*—Passive Optical Network (PON) technology offers a cost-effective alternative to support Beyond 5G Mobile Network Fronthauling (MFH). However, MFH dimensioning for such networks is challenging, given its high bandwidth and strict latency requirements. The Functional Split of the Radio Access Network (RAN) has been introduced to provide flexibility in resource allocation and reduce these requirements. In contrast to the conventional MFH requirement of RF-PHY splitting, the MFH traffic produced by high-layer splittings depends on the actual user traffic load. This dependency causes patterns of spatiotemporal variation in MFH traffic due to the daily movements of mobile users. This paper introduces a resource allocation mechanism that capitalizes on the spatiotemporal imbalance of mobile traffic to reduce the bandwidth required to support the RAN functional splitting over PONs. The results show that the proposed scheme offers higher bandwidth utilization, resulting in lower upstream delays compared to the baseline mechanisms.**

***Index Terms*—Low-latency, Mobile Network Fronthauling, Passive Optical Networks, Time-wavelength division multiplexing, 6G.**

I. INTRODUCTION

The 3rd Generation Partnership Project (3GPP) has introduced Radio Access Network (RAN) functional splits, (Figure 1) for reducing bandwidth demands and delays in the Mobile Fronthauling (MFH). It does so by disaggregating traditional RAN into three entities, namely, the Centralized Unit (CU), Distributed Unit (DU), and Remote Unit (RU). This disaggregation allows for a division of the protocol stack of Cloud Radio Access Networks (CRANs), enabling a larger number of functions to be located at remote sites than is traditionally done in CRANs.

Passive Optical Networks (PONs) provide an attractive technology for MFH due to the point-to-multipoint topology, which is suitable for fine-granular transport services. The 5G RAN nodes (*i.e.,* CU, DU, and RU) can be mapped to the transport elements of the PON (*i.e.,* Optical Line Terminal (OLT) and Optical Network Units (ONUs)). The possible mappings of the functional split options to the optical architecture are high-layer split (*i.e.,* CU/OLT-Midhaul-ONU/DU/RU), low-layer split (*i.e.,* CU/DU/OLT-Fronthaul-ONU/RU) and cascaded split (*i.e.,* CU/OLT-Midhaul-ONU/DU/OLT-Fronthaul-ONU/RU).

5G network deployments usually rely on single vendors with proprietary interfaces and equipment, which hinders the rapid adoption of 5G RAN functional splits. In response, O-RAN ALLIANCE has introduced the Open Radio Access Networks (O-RAN), which promotes openness and interoperability using open, low-cost off-the-shelf hardware and software for modular radio components. O-RAN allows network operators to customize their infrastructure to specific requirements and select from a variety of open components from multiple suppliers, thus avoiding the limitations of proprietary single-vendor solutions [ORAN-WG9.XPSAAS.0-R003-v04.00].

The O-RAN also defined the low-layer split as the highest benefit use case, where operators only need to place a small weatherproof RU at distributed antenna sites, while the cascaded split architecture can be considered to allow additional flexibility. The low-layer split still requires using high-capacity and low-latency fronthaul transport interfaces. In O-RAN, the elements of the 5G network are named slightly differently. The CU, DU, and RU are called O-CU, O-DU, and O-RU, respectively. However, for the sake of simplicity in this paper, these entities will be referred to without the prefix 'O-', as CU, DU, and RU respectively.

The O-RAN Forum recently collaborated with the ITU-T Q2/15 group to standardize the Cooperative Transport Interface (CTI) to address the stringent latency requirements of low-layer split. Originally proposed as the Cooperative Dynamic Bandwidth Allocation (DBA) (CO-DBA) method [1], the CTI was formalized in the standards [O-RAN.WG4.CTI-TMP.0-R003-v04.00], [ITU-T Rec.G.S71]. The CTI facilitates optical-wireless cooperative control between RAN and PON by transmitting scheduling information from the mobile scheduler (CU/DU) to the PON scheduler (DBA) in the OLT, which enables the OLT to allocate upstream bandwidth before the arrival of uplink mobile data from the RU at the ONU.

Recent PON standard 50G-EPON (IEEE 50 Gb/s Ethernet PON 802.3ca-2020) employs Time and Wavelength Division Multiple Access (TWDMA) for upstream transmissions between an ONU and an OLT, which supports up to two 25 Gbps wavelength channels with non-tunable transceivers.

Despite such a capacity, bandwidth sharing [2] is crucial for 5G and 6G networks to support the transport of explosive growth in traffic demands. PON tenants (customers) can be conventional end users (*e.g.,* residential and enterprise subscribers) as well as service providers, such as Virtual Network Operators or Mobile Network Operators (MNOs). Hereinafter, a *shared PON* scenario refers to a PON with multiple cus-

O. J. Ciceri, C. A. Astudillo and N. L. S da Fonseca are with the Institute of Computing, University of Campinas 13083-852, Brazil (emails: oscar@lrc.ic.unicamp.br, castudillo@lrc.ic.unicamp.br, nfonseca@ic.unicamp.br).

G. B. Figueiredo is with the Institute of Computing, Federal University of Bahia, Brazil (email: gustavobf@ufba.br).

Z. Zhu is with the School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230027, P. R. China (email: zqzhu@ieee.org).

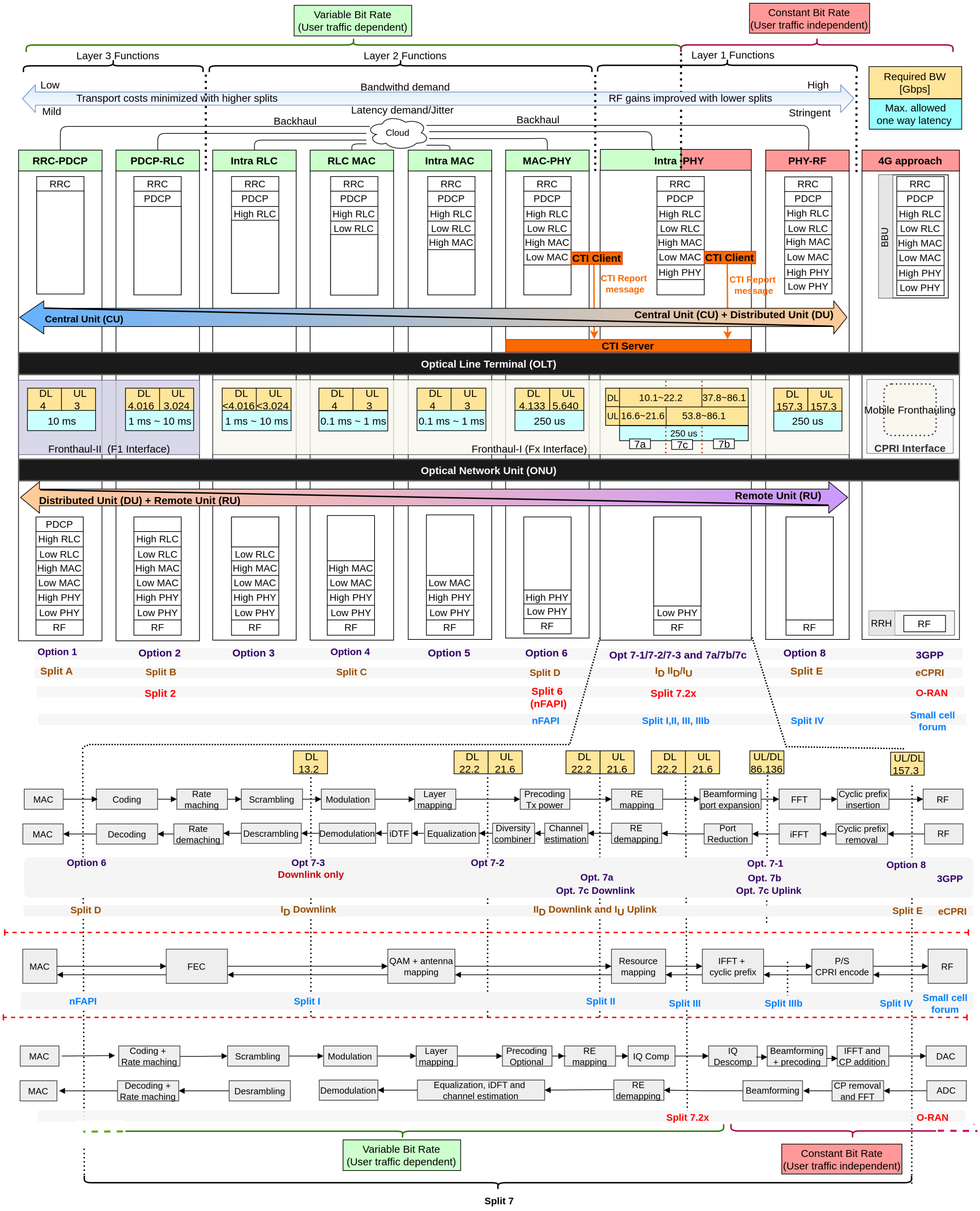


Fig. 1: RAN functional split options with delay requirements and bandwidth examples from 3GPP TR 38.801. Note that options 7a, 7b, and 7c are not equivalent to options 7-1, 7-2, and 7-3. Also, the option 7-2 standardized by the 3GPP is not equivalent to ORAN split 7.2x. Split options 1 to 7.2x are variable bit rate. Adapted from [ITU-T S.G Supp.66] and [O-RAN.WG9.XTRP-REQ-v01.00]

tomers and services, where customers owning a single ONU (*single-ONU customers*) and customers leasing/owning several ONUs (*multi-ONU customers*) may coexist.

In shared PONs, resource allocation in the upstream plays a critical role in the provisioning of MFH services. Although static resource allocation can be employed, it leads to inefficient use of the PON capacity. Variable bandwidth granting and user loads add complexity to meet the strict Quality of Service (QoS) requirements of MFH traffic. Dynamic Wavelength and Bandwidth Allocation (DWBA) algorithms are employed to provide bandwidth guarantees to the ONUs according to predefined Service Level Agreement (SLA) [3]. In the case of multi-ONU customers, the individual bandwidths of the ONUs to the same customer can be aggregated into a single SLA, known as a multi-ONU SLA model [2]. However, DBWA algorithms for multi-ONU SLA support introduce network delays to distribute the unused resource, which is inappropriate for MFH. However, dynamic resource allocation can improve statistical multiplexing gains and reduce delays.

MNOs deploy their base stations in different city areas, including residential and commercial areas. Due to its large footprint, a PON network can support the MFH traffic of base stations in distinct areas. Thus, DWBA algorithms must consider the tidal effect caused by the intrinsic mobility of users across distinct areas during the day. The tidal effect refers to the spatiotemporal variations in mobile traffic patterns due to user mobility. The tidal effect can cause an unbalanced load on the ONUs during the day. For instance, residential ONUs and commercial ONUs are usually overloaded at different times of the day.

Several algorithms employed either load prediction or information provided by the CTI to decrease the delay (Table I). However, none of them have addressed traffic spatiotemporal imbalance in *shared PONs* with *multi-ONU customers*. Consequently, employing these algorithms does not allow multi-ONU customers to capitalize on the load imbalance among their ONUs. If these algorithms were employed, bandwidth could be wasted and, consequently, the costs would increase. Considering traffic imbalance is crucial to ensure efficient resource allocation and meet the MFH service requirements.

This paper presents a DWBA algorithm for bandwidth allocation in TWDM-PON networks that supports RAN function splits. It leverages MFH's traffic imbalance to allocate resources among multi-ONU customers. The load imbalance and bandwidth waste problems in multi-ONU customers were addressed by implementing a mechanism with online scheduling and a two-cycle compensation method. The online mechanism avoids introducing additional delays, while the compensation mechanism allows a better distribution of unused resources within the ONUs of multi-ONU customers. This improvement reduces the bandwidth required to meet the delay requirements in multi-ONU customers. It differs from other approaches by allowing the distribution of excess bandwidth without requiring the arrival of all Report messages from multi-ONU customers, which reduces access delays. In addition, the two-cycle compensation method allows for additional bandwidth in each scheduling cycle, increasing the chances of providing ONU demands.

The input to our algorithm uses information obtained from the CTI about the traffic from the DU to the OLT for ONUs supporting RAN functional splits, and it employs the Status Report mechanism [3] for ONUs that do not support functional splits.

The performance of the proposed algorithm is compared to those of baseline algorithms, such as the Status-Report and Cooperative Interface approaches. Results derived by simulation show that the proposed algorithm produces high bandwidth utilization, making it possible to satisfy strict delay requirements.

The original contributions in this paper are:

- A DWBA algorithm that meets the requirements of the MFH functional splits over *shared PON* networks with *multi-ONU customers*.
- An analysis of the impact of RAN functional splits on MFH transport requirements and the provisioning issues in MFH over EPON networks.
- A methodology for obtaining the MFH traffic characteristics based on actual traffic datasets from two MNOs.
- Discussion of the open issues and challenges.

II. RAN FUNCTIONAL SPLITTING AND ITS IMPLICATIONS FOR MOBILE FRONTHUALING REQUIREMENTS

Recently, efforts have been made to decrease the MFH data rate and delays, which resulted in the definition of RAN functional splits. The Common Public Radio Interface protocol has been enhanced, resulting in a packetized protocol called eCPRI, which supports new split options. The functional split ranges from the most straightforward configuration (option 8) to the most complex one (option 1) on the RU side.

The Common Public Radio Interface (CPRI) protocol, supporting the PHY-RF option 8, involves centralized baseband processing at the CU, while the RU handles radio frequency (RF) functions. The intra-PHY variants (option 7) distribute functions differently. option 7.1 has FFT/IFFT in the RU, and option 7.2 adds pre-filtering/precoding. In contrast, option 7.3 (downlink only) locates the encoder in the RU. Option 6 (MAC-PHY) centralizes the MAC functions, leaving PHY and RF at the remote site. Option 5 (intra-MAC) places time-critical MAC functions in the RU. Option 4 (RLC-MAC) splits the RLC functions between CU and DU. Option 3 (intra-RLC) splits the RLC layer into high and low RLC. Option 2 (PDCP-RLC) locates RRC and PDCP in the CU, enhancing traffic control. Option 1 (RRC-PDCP) separates the user plane from the CU for better traffic management.

Centralizing RAN functions reduces costs, complexity of RU, and energy consumption. It facilitates resource sharing and cross-cell cooperation, enabling advanced schemes such as Coordinated Multipoint (CoMP) and soft handovers. For example, options 6, 7, and 8 support CoMP functionality, while options 7-3, 6, and 5 restrict some CoMP functions due to latency issues (*e.g.,* uplink joint reception). A centralized approach requires a large MFH capacity to transport In-phase and Quadrature (I/Q) components between remote and centralized locations. In contrast, distributing time-critical functions to a remote site alleviates the delay requirements on the MFH interface compared to centralized options.

TABLE I: Literature review on resource allocation for mobile fronthauling over PONs. G: Gated; L: Limited; F: Fixed; V: Variable; U: Unlimited; S: Simulation; A: Analytical; E: Experimentation; Y: Yes; N: No; NS: Not Specified; NA: Not Applicable. The *maximum cycle length* is the maximum duration of a PON scheduling cycle.

Feature		This	[1]	[4]	[5]	[6]	[2]	[7]	[8]	[9]	[10]	[11]	[12]	[13]
Standard family	IEEE (EPON)	X	X	NS	-		X	X	-	X	-	-	-	-
	ITU (GPON)	-	-		X	X	-	-	X	-	X	X	X	X
Multi. access tech.	TDMA-PON	-	X	X	X	X	-	-	X	X	X	-	-	X
	TWDMA-PON	X	-	-	-	-	X	X	-	-	-	X	X	-
Rate per wavelength	10 Gbps	-	X	X	X	X	X	X	X	X	X	-	-	X
	25 Gbps	X	-	-	-	-	-	-	-	-	-	X	X	-
Wavelengths (λs) per OLT		2	1	1	1	1	1	U	1	1	1	2	4	1
Simultaneous TX λs per ONU		1	1	1	1	1	1	U	1	1	1	1	1	1
Splitting option		6	8	6	8	7.3	NA	8	NS	6	6	7.1	7.2	2/6
Maximum PON cycle length [ms]		0.25	0.5	V	0.03125	0.125	1	NS	NA	0.2	NA	NA	1	NA
PON range [km]		5	10-20	2	6	20	10-20	20	10	1-20	5	5-20	10	10
SLA support		Y	N	N	Y	N	Y	N	N	N	N	N	Y	Y
Grant sizing policy for MFH ONUs		L	G	G	F	L	L	G	F	F/G	G	G	L	L
Conventional customers		Y	N	N	Y	N	Y	N	N	N	N	Y	Y	Y
Perf. eval. based on real deployment		Y	N	N	Y	N	N	N	N	N	N	N	N	N
3GPP TR-38.816 traffic modeling		Y	N	N	Y	N	N	Y	N	N	N	Y	Y	N
Performance evaluation		S	E	A	E	S/E	S	S	A	S/E	E	S	S	S
Approach	Status-Reporting	Y	N	N	Y	N	Y	N	N	N	N	Y	Y	Y
	Traffic-Monitoring	N	N	N	N	Y	N	Y	N	N	N	N	N	N
	Cooperative-Interface	Y	Y	Y	N	N	N	N	Y	Y	Y	N	Y	Y
Bandwidth Sharing		Y	N	N	N	N	Y	N	N	N	N	N	N	Y
Publication year (20YY)		24	14	16	18	18	18	19	20	20	21	21	22	23

Although several split options are available, the ones chosen by the 3GPP and most research works are splits 2, 6, and 7. Split option 2 (PDCP-RLC) was selected as the high-layer split point (F1 Interface) [TS 38.470]. Moreover, the most common low-layer splits are split option 6 for MAC/PHY and split option 7 for intra-PHY [TR 38.816].

In parallel, the O-RAN Alliance narrowed down the functional splits to 2, 6, and 7.2x. O-RAN split 7.2x is based on the 3GPP option 7-2 but differs by including precoding and support for modulation compression (see Figure 1). Moreover, the fronthaul traffic of the split 7.2x is user-data dependent [O-RAN.WG9.XTRP-REQ-v01.00].

Moreover, the O-RAN ALLIANCE and the Small Cell Forum (SCF) have agreed to develop the split 6 (nFAPI) interface specification. This 5G nFAPI specification defines the functional split between the 5G MAC and PHY functions, allowing MAC and higher-layer functions to be virtualized. Split 6 requires reduced bandwidth, meaning high-quality fiber is not needed to link every DU and RU, unlike other lower-layer split options.

Split options 7-1, and 8 generate a Constant Bit Rate (CBR) traffic and introduce high redundancy in transmitted I/Q signals. The radio interface I/Q is a sample and quantized radio, producing CBR traffic that scales with cell-site bandwidth and antennas. In contrast, other split options produce a variable bit rate traffic dependent on the user data plane traffic, which is bursty than options, 7-1, and 8.

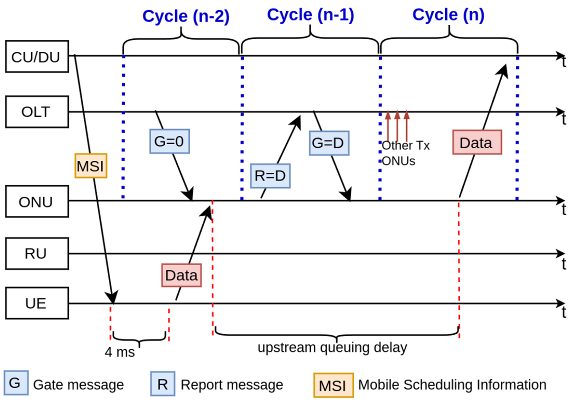
III. APPROACHES AND RELATED ISSUES IN THE SUPPORT OF MOBILE FRONTHAULING OVER EPON NETWORKS

Infrastructure service Providers (InPs) can increase revenues by offering various services over the same PON infrastructure. Nevertheless, the unique MFH requirements make QoS provisioning challenging, especially in scenarios with coexisting MFH and conventional PON services (*e.g.,* residential and enterprises). Various Resource Allocation (RA) algorithms for EPON networks have been proposed to provide guaranteed bandwidth and low latency in MFH (See Table I). The main issues of RA algorithms supporting MFH in PONs are reviewed next.

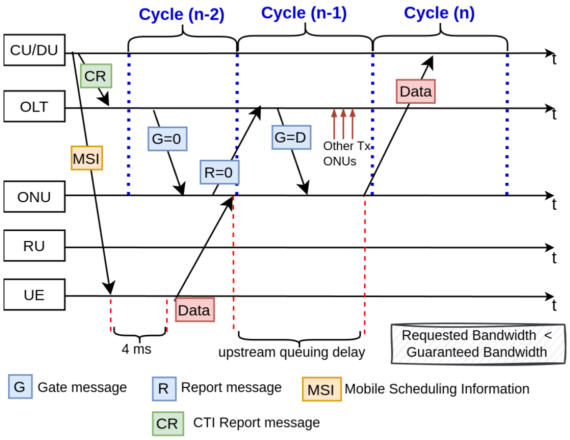
RA algorithms can be classified as Static Bandwidth Allocation (SBA) or Dynamic Bandwidth Allocation (DBA). The former allocates a fixed transmission window for each ONU, independently of the ONU load, guaranteeing deterministic delays. However, bandwidth can be wasted, increasing costs, especially when dealing with split options with variable rates. The latter allocates transmission windows per cycle, depending on the offered load, the delay requirement, and the available bandwidth. They increase statistical multiplexing gain in scenarios with unbalanced loads but introduce challenges for managing the available bandwidth in scenarios with low-latency requirements. Moreover, DBA schemes use Gate and Report messages to coordinate upstream transmissions between the ONUs and the OLT.

A. The Bandwidth Request Problem

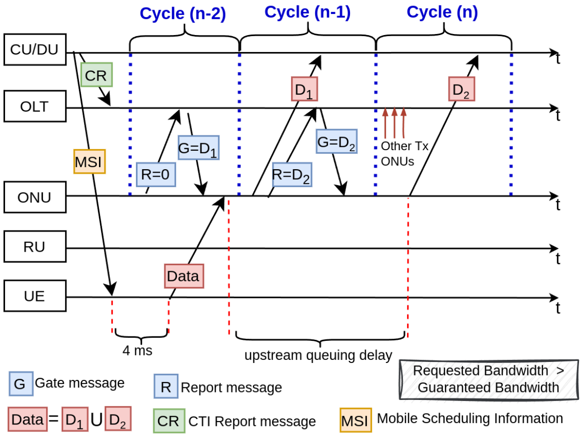
The Status Reporting scheme uses explicit ONU buffer occupancy reports for bandwidth distribution. Bandwidth is requested in Report messages sent from the ONUs to the OLT, whereas information on bandwidth availability is sent in Gate messages from the OLT to the ONU [3]. However, this adds queueing delays in packet transmission, which can harm MFH traffic. *The Bandwidth Request problem*, one of the most critical issues in RA, is produced by the Status Reporting



(a) Bandwidth Request Problem



(b) Maximum Cycle Length Problem



(c) Grant Sizing Policy Problem

Fig. 2: Resource allocation issues for low-latency MFH over PONs

scheme (Fig. 2a). The OLT must wait for a request message before granting bandwidth, which implies that the upstream delay will be at least one scheduling cycle in duration. This delay can be as long as one millisecond, which is much longer than the delay required for split option six and above.

One approach for addressing this problem, called Traffic-Monitoring or Traffic Prediction, involves estimations of upcoming MFH traffic to avoid waiting for a report message, which can reduce latency to acceptable levels ([6], [7]). This approach uses traffic prediction based on report messages to forecast traffic arrivals shortly and allocate bandwidth without the OLT to receive requests from ONUs. However, traffic forecasting is inaccurate and adds complexity and additional processing time.

Cooperation between the CU and the OLT was first proposed in [1] and has been widely adopted as a critical technique for low-latency MFH ([4], [8]–[10], [12], [13]), which was recently standardized in ITU-T Rec. Series G Supplement 71 (G.Sup.coDBA) and O-RAN Cooperative Transport Interface Transport Control Plane Specification (O-RAN.WG4.CTI-TCP.0-v02.00). This cooperation allows the OLT to obtain accurate information about upcoming traffic by exploiting Mobile Scheduling Information, which is then used to inform mobile users about resource allocation 4 ms before the uplink transmission. In this process, the CU/DU makes scheduling decisions and informs each User Equipment (UE) about the allocated bandwidth. The CU/DU determines the corresponding fronthaul traffic load per RU based on the scheduling allocations to the associated UEs. Then, the CU/DU informs the OLT about the traffic load per RU for the given slot via specific signaling messages (*i.e.,* Cooperative Transport Interface (CTI) Report messages) that contain the traffic identification and traffic volume in the given time interval, as shown in Fig. 2b. Low-latency upstream transmission on the MFH with the TDM-PON can be achieved by distributing the bandwidth to the ONUs based on the transformed UE uplink grant information.

B. The Maximum Cycle Length Problem

The upstream delay depends on the cycle duration because each ONU usually transmits only once per cycle. Even when traffic monitoring or cooperative schemes are employed, mobile traffic can arrive at the ONU just after an ONU transmission in that cycle. These frames remain in the ONU queue at least until the next transmission cycle, as shown in Fig. 2b. Hence, the upstream delay can be as long as the maximum cycle length (the maximum duration of a PON scheduling cycle) when the network is overloaded.

Moreover, the polling overhead (bandwidth waste) depends on the number of guard periods per cycle and cycles per second. As the maximum cycle length increases, the overhead decreases. Thus, there is a trade-off between overhead and delay, which must be carefully addressed. PON systems that adhere to ITU-T [5], [6], and IEEE [9] standards have a fixed grant cycle of 125 μs and a maximum grant cycle of 200 μs, respectively, which meet the stringent latency requirement of low-layer split options. For example, Bidkar et al. [5] proposed

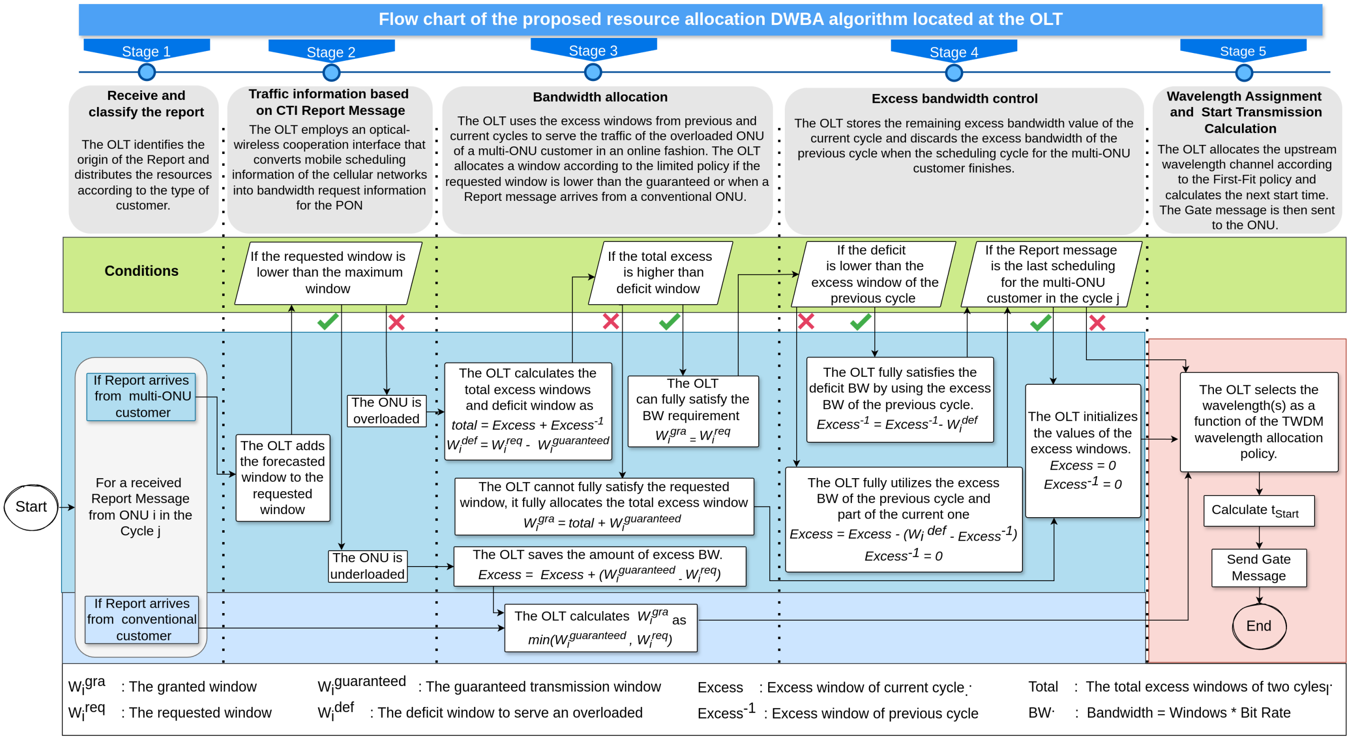


Fig. 3: Flow chart of the proposed resource allocation algorithm

a technique that allocates multiple transmission opportunities per cycle to reduce the upstream delay of CPRI traffic over GPON networks. However, multiple grant allocations or very short scheduling cycles ($< 250\ \mu s$) increase the number of guard periods and bandwidth overhead.

Table I indicates that the importance of the maximum cycle length has received little attention in most previous investigations (*e.g.,* [1], [4], [7]).

C. The Grant Sizing Problem

As seen in Table I, various resource allocation schemes to support MFH over PONs employ a Gated policy in which the OLT allocates a transmission window equal to the forecast/requested one (*e.g.,* [1], [4], [7], [10], [11]). However, this policy does not allow QoS provisioning (*e.g.,* guaranteed bandwidth and delay) for several types of services co-existing on the same PON infrastructure, as in the current shared PONs. This problem occurs when customers overuse the total available bandwidth, thus increasing the cycle duration if there is no traffic shaping in place.

InPs usually limit to guaranteed bandwidth allocation for PON customers according to a pre-defined SLA. In such a policy, the OLT grants a transmission window equal to the minimum between the requested window and the maximum allowed transmission window (Fig. 2c). However, customers who have an offered load greater than their guaranteed bandwidth (overloaded customers) usually need several scheduling cycles to complete the sending of packets in the ONU queue. The uplink queuing delays of these overloaded customers depend on the number of cycles required for the OLT to provide the required bandwidth. Thus, the Traffic-Monitoring and Cooperative schemes still rely on fluctuation in traffic and guaranteed bandwidth at the ONU level. When there is an overload condition, ONUs may require several PON cycles to send a packet in the buffer to the OLT, which increases the overall delay of MFH services, as shown in Fig. 2c.

A bandwidth-sharing mechanism was proposed in [2] to address the bandwidth starvation problem of overloaded ONUs in backhauling scenarios with multi-ONU customers. This mechanism guarantees that the bandwidth of ONUs belonging to the same customer can be shared so that unused bandwidth from underloaded ONUs can be assigned to overloaded ones per cycle, thus reducing the number of scheduling cycles needed to serve an overloaded customer. However, calculating the unused bandwidth of an underloaded ONU requires that the OLT wait for the arrival of all Report messages from the ONUs belonging to the same customer before sending Gate messages to the overloaded ONUs, increasing latency.

In summary, despite MFH traffic prediction or cooperation between Mobile and PON, devices can reduce the latency introduced by *bandwidth request problems*, *grant sizing policy problems*, and *maximum cycle length problems*. However, the effect of such problems can still generate delays longer than those required by MFH services.

IV. DWBA Algorithm for EPON-based Mobile Fronthauling

To address the problems mentioned in the previous section, we introduce a novel DWBA algorithm to provide high-throughput and low-latency for 5G mobile fronthauling services, called *Resource Allocation for Low-Latency Mobile Fronthauling* (*RALM*) algorithm. Our proposal deals with multiple wavelengths and employs an adaptive polling cycle for dynamic resource allocation to meet the IEEE 50 Gb/s EPON standard requirement while meeting PON customer SLA demands. Although typically bandwidth and wavelength are allocated separately and on different time scales, in a 50 Gb/s EPON architecture (IEEE 802.3ca-2020), the 50G ONUs can operate on two wavelengths per cycle by using two non-tunable transceivers, which prevents additional wavelength switching delay due to wavelength tunability.

The proposed mechanism adopts the widely-used Cooperative Transport Interface (CTI) approach proposed in [1] and standardized on ITU-T Rec. Series G Supplement 71 (G.Sup.coDBA) to tackle *the problem of bandwidth requests*. A *bandwidth-sharing* mechanism is also employed to ameliorate *the problem of grant sizing* and exploit spatial-temporal characteristics of MFH traffic. The waiting time associated with bandwidth-sharing techniques can be avoided by employing an excess bandwidth compensation approach. Our proposal aims to allow excess bandwidth from previous and current cycles to serve the upcoming traffic of overloaded ONUs of a multi-ONU customer in an online fashion. This procedure allows for immediate service (on-the-fly) for bandwidth requests for MFH traffic.

The OLT grants a maximum transmission window per scheduling cycle for the ONUs (i.e., Limited Policy) belonging to a single-ONU customer to provide such guarantees. On the other hand, a multi-ONU customer establishes an SLA for its group of ONUs (multi-ONU SLA) [1]. Such an SLA defines a guaranteed bandwidth for each ONU, and its group of ONUs. Then, the OLT employs the unused bandwidth of the underloaded ONUs to grant a bandwidth greater than the guaranteed one for the overloaded ONUs, which improves the resource distribution for multi-ONU customers.

Moreover, the maximum cycle length is chosen not to exceed the latency requirement, thus addressing the *maximum cycle length problem*. Our approach selects wavelengths as a function of the Time and Wavelength Division Multiplexing (TWDM) allocation policy, which is, to the best of our knowledge, the first solution to simultaneously address the Bandwidth Requests, Maximum Cycle Length, and Grant Sizing problems in QoS provisioning for MFH in *shared PONs*.

Fig. 3 summarizes the proposed scheme, which resides in the OLT. When a Report message arrives from a conventional ONU, the OLT calculates the transmission window according to the limited policy (Stage 3). If the Report message comes from a multi-ONU customer, the OLT updates the requested window using the optical-wireless cooperation procedure (Stage 2). If the requested window is smaller than the guaranteed windows, the ONU is fully served and the value of the unused bandwidth is stored (Stage 3). Otherwise, the OLT grants additional bandwidth to an overloaded ONU by utilizing the excess bandwidth from the previous and current cycles (Stage 4). Moreover, when the scheduling cycle ends for the multi-ONU customer, the OLT saves the remaining excess bandwidth value of the current cycle and discards the excess bandwidth of the previous one. Moreover, the wavelength is assigned by the OLT on the upstream channel according to the First-Fit policy and calculates the next start time. In this policy, the OLT grants the first available wavelength. Finally, the Gate message is sent to the ONU (Stage 5).

V. PERFORMANCE EVALUATION

The performance of the proposed DWBA scheme was evaluated using the Ethernet PON (EPON) simulator (EPON-Sim), previously validated in [2]. The EPON-Sim simulator was extended to support TDWMA-PON and MFH services. The performance of the proposal was evaluated for various bandwidth allocation schemes and the tidal MFH traffic in shared PON networks. Four different baseline approaches were employed: 1) Status Reporting (*i.e.,* First-Fit with SR [3]); 2) Cooperative Interface (*i.e.,* First-Fit with CTI [1]); 3) Bandwidth Sharing (*i.e.,* MOS-IPACT with SR [2]); and 4) Combined Cooperative Interface with Bandwidth Sharing (MOS-IPACT with CTI). This makes it possible to evaluate the effect of the proposed scheme with various Dynamic Wavelength and Bandwidth Allocation (DWBA) schemes for mobile fronthauling in PON networks.

A. Mobile Traffic Modelling

A large data set containing data from two MNO cellular networks in Ireland [14] was used to capture the impact of relevant aspects such as topology, spatial traffic demands, demographics, and MNO information on MFH performance. This dataset provides base station location, operator, base station to user cluster associations, area type of user clusters, and the cumulative distribution function (CDF) of the demands of users served at the peak hour for each type of area.

Each BS was classified as commercial, residential, or rural according to the most representative type of user served. The peak offered traffic load was generated by using Monte Carlo simulations. First, we simulate the maximum traffic load for each user served using the data demand CDF for the corresponding type of area. Then, the user traffic demands corresponding to each base station are aggregated.

An InP with an EPON system and a radius of 5 km covering the north of Dublin was simulated. In that region, one MNO owns 44 BSs, of which 40% serve residential areas; the rest serve commercial areas. Moreover, about 15% of the base stations (six) were assumed to employ PON as their MFH. The BS traffic distribution was assumed to vary as a function of BS location and period [15], so the tidal effect could be fully captured. Thus, three commercial and three residential base stations were selected within the region (shown in Fig. 4) to maximize the Jain fairness index of the offered load at peak hours in the selected region. The BSs shared their guaranteed bandwidths when bandwidth-sharing-based schemes were used.

In this way, the BS offered loads for each scenario were obtained, as well as the distances between the OLT and the MFH ONUs used in the simulations. The load values obtained were scaled by a factor $c = 3$ to obtain compatible values with the 3GPP 38.801 release. The scaled offered loads obtained were (25.51, 30.09, and 21.46) Mbps and (27.45, 23.36, and 30.00) Mbps for residential and commercial BSs, respectively.

B. Simulation Model and Setup

A 50G TWDMA-EPON tree topology network with 32 ONUs was simulated. Each ONU can transmit on a single 25 Gbps wavelength allocated dynamically. There is an MNO renting part of the PON from InP to support the MFH serving six BSs, as described in Section V-A. Each MFH ONU is connected to its corresponding RU through a local 100Gbps Ethernet interface.

Option 6 is assumed for all the BSs as this is the most demanding one with the largest bandwidth and lowest latency requirements for the upstream variable-rate split options (see Fig. 1). The BS peak loads obtained in Section V-A and the same assumptions proposed in 3GPP TR 38.801v14 were used to generate the peak offered load of the kth ONU/RU (P_k). The P_k values obtained were (4170, 4445, and 3927) Mbps and (4287, 4041, and 4440) Mbps for residential and commercial ONU/RUs, respectively.

The guaranteed bandwidth B_k of the MFH ONUs was varied from $0.8 \cdot P_k$ to $1.2 \cdot P_k$. For clarity, hereafter, P_k has been omitted from the values B_k. The variation in guaranteed bandwidth indirectly reflects the impact of the variation in traffic load on network delay. The rest of the ONUs in the PON was assumed to be conventional in testing a coexisting scenario with support for different PON services. Each conventional ONU had a guaranteed bandwidth equal to the remaining available bandwidth in the PON divided by the number of conventional ONUs.

Moreover, the mean offered load of conventional ONUs was 85 % of their guaranteed bandwidth to simulate high loads. In practice, additional network resources are allocated to the system when it achieves around 85 % of its capacity [15].

The load generated by ONU/RU follows a Poisson distribution with a mean value equal to the offered load for the corresponding scenario, while that for the traffic of conventional ONUs follows a self-similar traffic model [2]. Moreover, the RU was assumed to generate bursts of Ethernet frames (MFH data) every 250 μs. The maximum cycle length was set to 250 μs, the propagation delay was considered to be 5 μs/km, and the guard period between transmissions from different ONUs in the PON was 0.624 μs.

Two scenarios lasting 60s each were simulated: 18:00 (6:00 pm) and 23:00 (11:00 pm). They were replicated ten times. The BSs in commercial areas reached their peak load at 18:00, while those in residential areas reached only 38.1% of their maximum load. At 23:00 hours, the BS in the residential areas reached their peak load, while those in commercial areas reached only 8.1% of their maximum load [15]. These intervals highlighted the load imbalance in ONU/RUs at different times and locations in the city.

Since MFH ONUs serving BSs during off-peak traffic hours experience lower delay values than those serving BSs in peak traffic hours. The analysis in the next section focuses on MFH ONUs in the peak traffic hour for each time scenario.

C. Simulation Results and Discussions

The simulation results show that MFH traffic does not experience packet loss per ONU/RU (figures not shown in this paper) because the aggregated guaranteed bandwidth is lower than the aggregated offered load, which is sufficient to serve such traffic. Thus, we analyze the guaranteed bandwidth per ONU/RU and the corresponding MFH delay. The MFH delay is the time elapsed between the arrival of the frame at the ONU and at the OLT.

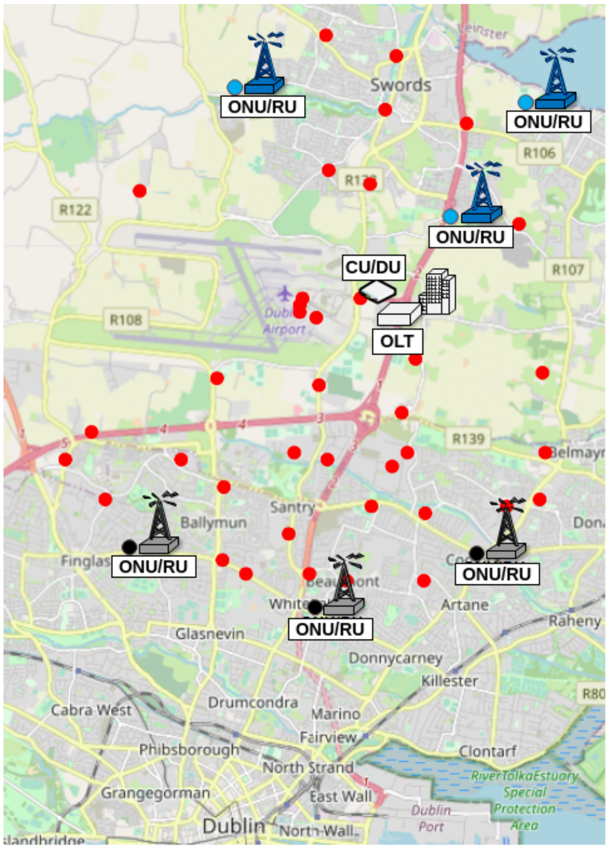


Fig. 4: Geographical location of base stations in a 5 km-radius region in Dublin, Ireland. Points indicate base station locations. Blue Points and black points indicate urban base stations (located in commercial areas) and suburban base stations (located in residential areas) used in the simulations, respectively.

The proposed scheme achieved the required delay value (<250μs) for both scenarios with guaranteed bandwidth per ONU/RU greater than or equal to 105% of the average peak hour load value (Fig. 5). The MFH ONUs in the 18-hour scenario require 10% more guaranteed bandwidth than those in the 24-hour scenario due to the differences in the off-peak BS traffic loads of the two scenarios. Since the bandwidth leased must satisfy the worst-case scenario, the MNO needs 1.05 guaranteed bandwidth per ONU/RU. The other schemes fail to produce satisfactory MFH delay values for the split option considered for the guaranteed bandwidth values tested.

Our proposal not only meets the MFH delay requirements, but also significantly reduces the required guaranteed bandwidth per ONU/RU compared to other schemes. This leads to increased bandwidth utilization and, importantly, to a cost reduction for MNOs renting bandwidth from the InP

The schemes employing the Cooperative Interface produce lower delay values than those that do not employ the Cooperative Interface. Moreover, bandwidth-sharing-based schemes give 99.99th percentile delay values lower than those without this technique. Furthermore, methods employing an online policy with the cooperative interface (i.e., First-Fit with CTI and RALM) generate 99th percentile delays lower than those using an offline policy. Our proposal combines all of these features to meet the stringent MFH requirements of a network that provides QoS guarantees for all supported services.

The evaluated schemes do not decrease the 99.99th percentile of the delay value after a specific increase in the contracted bandwidth. There is a fundamental limit in the performance of resource allocation mechanisms that cannot be surpassed by simply increasing the bandwidth. Even with the CTI interface and appropriate maximum cycle length settings, online disciplines, and bandwidth sharing mechanisms, the *grant sizing problem* must be properly addressed to meet MFH low latency requirements.

VI. Future Research Directions

This section presents future research directions for applying PON technology in 6G fronthauling networks.

A. MFH Traffic Forecasting

Most current approaches for low latency MFH services employ the cooperation between the Mobile and the PON network (*i.e.,* CTI approach). However, for split options 1 to 5, the CTI approach cannot be implemented because the mobile scheduling information (*i.e.,* LOW MAC) is located on the ONU side instead of the OLT side, where it is needed (see Figure 1). Alternatively, the ONU could communicate with the OLT to provide the necessary CTI information for bandwidth allocation before the arrival of uplink mobile data. However, this process increases complexity and delay while requiring additional bandwidth due to additional messages sent, making it feasible only for Enhanced Mobile Broadband (eMBB) and massive machine-type communications (mMTC) services.

Another option would be to perform traffic forecasting at the OLT without employing cooperation between the Mobile and the PON networks. There is a trade-off between latency overhead and prediction accuracy. Such a trade-off is especially relevant when involving Ultra-Reliable and Low-Latency Communications (URLLC) since it may require a one-way access delay as low as $100\,\mu s$. Machine learning algorithms can be used for traffic prediction with high accuracy and in acceptable time frames to cope with this trade-off.

B. MFH-Aware PON Dimensioning and Planning

The design of PONs does not consider the requirements of MFH traffic. Principles in PON design must be defined to maximize network utilization while minimizing the guaranteed bandwidth of a group of ONUs, while supporting these demands under various delay constraints. For example, groups of ONU/RUs could be employed to optimally exploit the concept of bandwidth sharing. Another potential approach is to design an MFH network to reduce the number of wavelengths and the OLT equipment required.

C. MFH Traffic Management

Traffic management in the ONU is vital in supporting low-latency MFH traffic over PONs, especially in meeting variable-rate split options under dynamic resource allocation schemes that employ the Gated bandwidth allocation policy. Traffic management mechanisms are essential to guarantee deterministic delay bounds. For example, the value of the parameters of the traffic shaping mechanisms should be optimal to meet the delay requirement of each split option. This tuning must be well understood to achieve compliance with the definition of split options.

VII. Conclusion

This paper has introduced a novel resource allocation mechanism for supporting 5G mobile network fronthauling in EPONs. The main RA issues and approaches to support low-latency MFH in these networks are discussed. Our proposal includes bandwidth sharing for multi-ONU customers, the use of the CTI interface, and maximum cycle length tailored to MFH traffic requirements. Simulation results show that our proposal provides lower delay values than do existing schemes under realistic traffic scenarios. Our proposal increases network utilization and statistical multiplexing gain for MNOs employing PON-based MFH services, leading to lower MFH costs than those of existing approaches. Moreover, with our proposal, InPs can offer attractive business models.

Acknowledgment

This work was partially supported by CNPq grant # 313057/2020-6, CNPq INCT grant # 405940/2022-0, and the Sao Paulo Research Foundation (FAPESP) for the grant # 2023/00673-7.

References

[1] T. Tashiro *et al.*, “A novel DBA scheme for TDM-PON based mobile fronthaul,” in *OFC 2014*, 2014, pp. 1–3.

[2] O. J. Ciceri, C. A. Astudillo, and N. L. da Fonseca, “Dynamic bandwidth allocation with multi-ONU customer support for ethernet passive optical networks,” in *2018 IEEE Symposium on Computers and Communications (ISCC)*. IEEE, 2018, pp. 00 230–00 235.

[3] S. B. Hussain *et al.*, “Low-latency dynamic wavelength and bandwidth allocation algorithm for NG-EPON,” *Journal of Optical Communications and Networking*, vol. 9, no. 12, pp. 1108–1115, 2017.

[4] H. Ou *et al.*, “Passive optical network range applicable to cost-effective mobile fronthaul,” in *2016 IEEE International Conference on Communications (ICC)*, 2016, pp. 1–6.

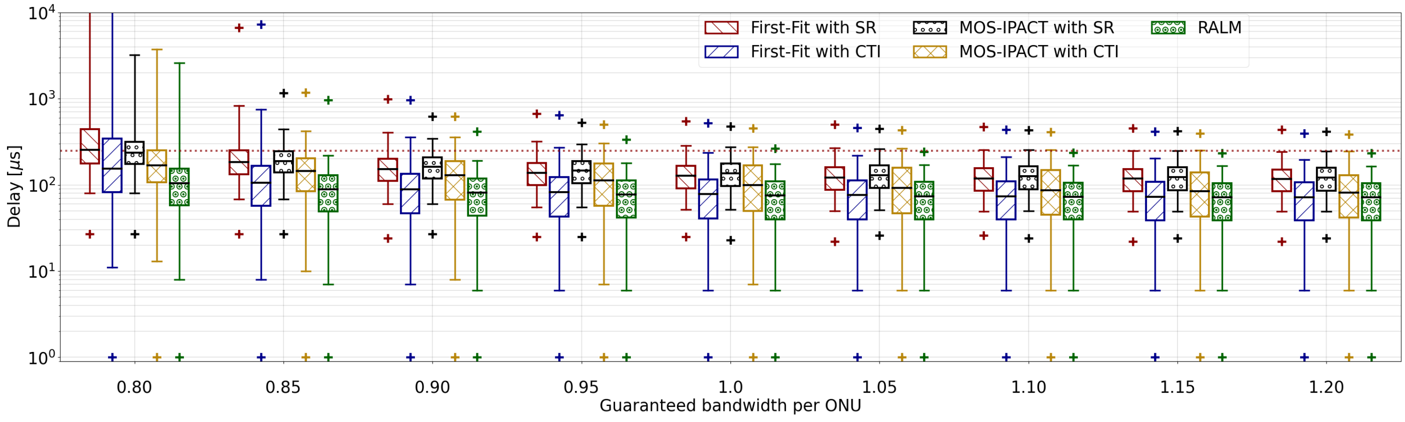
[5] S. Bidkar, J. Galaro, and T. Pfeiffer, “First demonstration of an ultra-low-latency fronthaul transport over a commercial TDM-PON platform,” in *2018 Optical Fiber Communications Conference and Exposition (OFC)*. IEEE, 2018, pp. 1–3.

[6] D. Hisano *et al.*, “Predictive bandwidth allocation scheme with traffic pattern and fluctuation tracking for TDM-PON-based mobile fronthaul,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2508–2517, 2018.

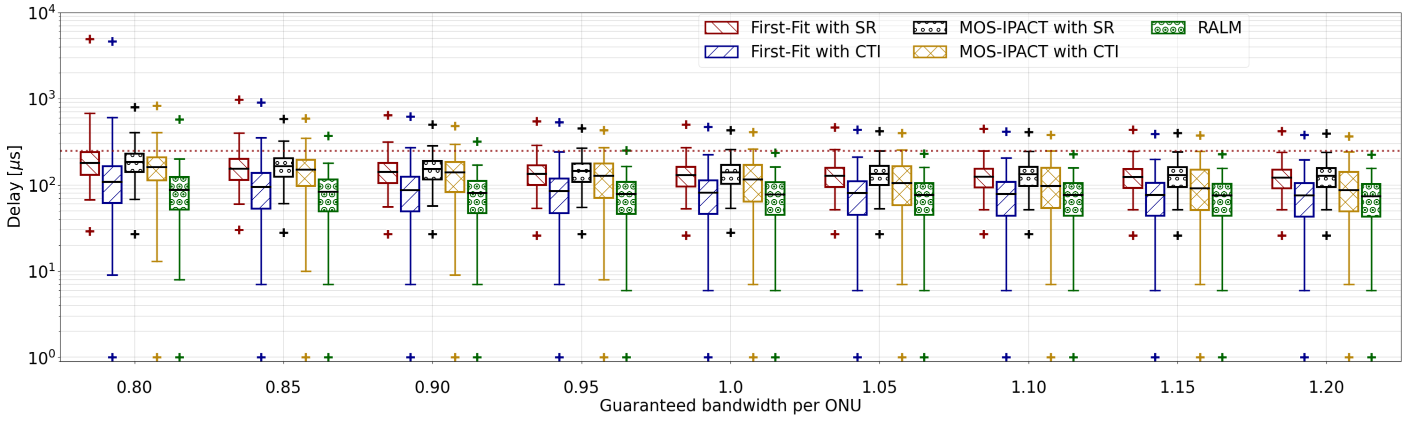
[7] M. R. Araujo *et al.*, “A novel prediction-based DWBA algorithm for NG-EPON based C-RAN fronthaul,” in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–6.

[8] S. Bidkar, R. Bonk, and T. Pfeiffer, “Low-latency TDM-PON for 5G xhaul,” in *2020 22nd International Conference on Transparent Optical Networks (ICTON)*, 2020, pp. 1–4.

[9] H. Uzawa *et al.*, “Dynamic bandwidth allocation scheme for network-slicing-based TDM-PON toward the beyond-5G era,” *IEEE/OSA Journal of Optical Communications and Networking*, vol. 12, no. 2, pp. A135–A143, 2020.



(a) Commercial area ONU/RUs at 18h



(b) Residential area ONU/RUs at 23h

Fig. 5: Delay performance of MFH ONUs for different resource allocation schemes with RAN functional split option 6; Each boxplot shows 1st, 25th (Q1), 50th (Q2), 75th (Q3), and 99th percentile of delay values; the lowest and the 99.99th percentile delay values are shown as lower and upper outlier points, respectively.

[10] L. Larsen *et al.*, “Experimental demonstration of RAN functional split over virtual PON transport network,” *CoRR*, vol. abs/2104.04645, 2021. [Online]. Available: https://arxiv.org/abs/2104.04645

[11] A. Zaouga *et al.*, “Self-adjusting DBA algorithm for next generation PONs (NG-PONs) to support 5G fronthaul and data services,” *Journal of Lightwave Technology*, vol. 39, no. 7, pp. 1913–1924, 2021.

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

[13] S. O. Edeagu *et al.*, “A hybrid dynamic bandwidth allocation scheme operating with IACG and cooperative DBA for converged fronthaul networks,” *Optical Fiber Technology*, vol. 75, p. 103199, 2023.

[14] P. Di Francesco, F. Malandrino, and L. A. DaSilva, “Assembling and using a cellular dataset for mobile network analysis and planning,” *IEEE Transactions on Big Data*, vol. 4, no. 4, pp. 614–620, 2017.

[15] A. Checko *et al.*, “Cloud RAN for mobile networks — A technology overview,” *IEEE Communications Surveys Tutorials*, vol. 17, no. 1, pp. 405–426, 2015.

BIOGRAPHIES

Oscar J. Ciceri obtained his degree in Electronics and Telecommunications Engineering at Universidad del Cauca (UNICAUCA), Colombia, in 2015, and his M.S. degree in Computer Science at the University of Campinas (UNICAMP), Brazil, in 2019. He is a Ph.D. student and researcher in the Network Computing Laboratory (LRC) at that university. His research interests include Passive Optical Networks (PONs), 5G and 5G beyond networks, Machine Learning, and virtualization. His current work is on the low-latency PON technologies for 5G and 5G beyond mobile optical access systems.

Carlos A. Astudillo obtained his Ph.D. degree in Computer Science from the University of Campinas (UNICAMP), Brazil, in 2022. He is an assistant professor at the Institute of Computing (IC), UNICAMP. He received the best paper award at the IEEE ISCC and the TAOS SAC-ICC, as well as top honors in thesis and dissertation contests in Brazil and Latin America. His research interests include network protocols and resource management in 5G/6G networks.

Gustavo Bittencourt Figueiredo [M’14, SM’18] received his B.Sc. degree in Computer Science from Salvador University (2001), and the M.Sc. (2003) and Ph.D. (2009) degrees in Computer Science from the University of Campinas. Since 2010, he has been affiliated with the Department of Computer Science at the Federal University of Bahia, Bahia - Brazil, where he is currently an Associate Professor. His main research interest includes Optical Networks, 5G Networks, Networking Modeling and Optimization, and the Design of algorithms for networking.

Zuqing Zhu [M'07, SM'12, F'23] received his Ph.D. degree from the Department of Electrical and Computer Engineering, University of California, Davis, in 2007. He is currently a Full Professor at the University of Science and Technology of China (USTC). He has published more than 360 papers in referred journals and conferences, and has received the Best Paper Awards from ICC 2013, GLOBECOM 2013, ICC 2015, and ONDM 2018. He is an IEEE Communications Society Distinguished Lecturer (2018-2021), and a Senior Member of Optica.

Nelson L. S. da Fonseca [M'88, SM'01] obtained his Ph.D. degree from the University of Southern California in 1994. He is Full Professor at the Institute of Computing, State University of Campinas (UNICAMP). He published 450+ papers and supervised 80+ graduate students. He is Senior Editor IEEE Systems Journal. He is also an Associate Editor for Peer-to-Peer Networking and Applications, Computer Networks, and Optical Switching and Networking. He is a past EiC of the IEEE Communications Surveys and Tutorials. He served as the IEEE ComSoc VP Conferences, VP Publications, VP Technical and Educational Activities, and VP Member Relations.