This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Zhang, Jiawei; Liu, Xiaozhu; Liang, Minghui; Yu, Hao; Ji, Yuefeng

Low Latency DWBA Scheme for Mini-Slot based 5G New Radio in a Fixed and Mobile Converged TWDM-PON

Published in:
Journal of Lightwave Technology

DOI:
[10.1109/JLT.2021.3117972](https://doi.org/10.1109/JLT.2021.3117972)

Published: 01/01/2022

Document Version
Peer reviewed version

Please cite the original version:
Zhang, J., Liu, X., Liang, M., Yu, H., & Ji, Y. (2022). Low Latency DWBA Scheme for Mini-Slot based 5G New Radio in a Fixed and Mobile Converged TWDM-PON. *Journal of Lightwave Technology*, 40(1), 3-13.
<https://doi.org/10.1109/JLT.2021.3117972>

© 2021 IEEE. This is the author's version of an article that has been published by IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Low Latency DWBA Scheme for Mini-Slot based 5G New Radio in a Fixed and Mobile Converged TWDM-PON

Jiawei Zhang, Xiaozhu Liu, Minghui Liang, Hao Yu, and Yuefeng Ji

Abstract—Cooperative dynamic bandwidth allocation (CO-DBA) enables time division multiplexed passive optical network (TDM-PON) to become a possible solution for mobile fronthaul. Meanwhile, status-report DBA (SR-DBA) still retains its advantage on bandwidth efficiency for conventional PON services. The collaboration of CO- and SR-DBA is an enabling approach for fixed-mobile convergence into a single PON infrastructure. However, a resource allocation conflict between CO- and SR-DBA is to arise when using an emerging 5G new radio interface based on the mini-slot. This conflict will cause a large delay that can not meet the time-sensitive fronthaul traffic requirement. In this study, we consider a time and wavelength multiplexed PON (TWDM-PON) for fixed and mobile services accommodation because of its capacity and flexibility. A novel DWBA scheme is proposed by means of wavelength and bandwidth preemption (WBP) mechanism. The basic rule of WBP is that a time-sensitive traffic can preempt the wavelength and bandwidth allocated to the time-tolerant traffic when a conflict happens. We compare it with an existing integrated CO- and SR-DWBA scheme without preemption under various situations. Also, we evaluate the performance of the proposed DWBA with different preemption policies.

Index Terms—Low latency, TWDM-PON based fronthaul, mini-slot based 5G NR, wavelength and bandwidth preemption.

I. INTRODUCTION

THE fifth generation (5G) of mobile network is opening new paradigms of communications, which is expect to address much broader range of applications and use cases. Compared with 4G that focuses a human-centric communication, 5G enables machine-type communication in a variety of industrial and societal fields, which requires a very stringent low latency, like 1ms for ultra-reliable and low-latency communication (URLLC). The high-quality performance of 5G needs a new air interface, called 5G new radio (NR). Both 5G NR and 4G long term evolution (LTE) use orthogonal frequency division multiplexing (OFDM) as the waveform. However, one difference between them is that LTE uses a fixed numerology of 15 kHz sub-carrier spacing (SCS), whereas NR specifies a scalable numerology with sub-carrier spacings of 15, 30, and 60 kHz below 6 GHz, and 60 and 120 kHz above 6 GHz [1].

Another difference is the decreasing number of OFDM symbols in a slot. The slot, a similar concept of transmission time interval (TTI) in LTE, is a basic frame structure and scheduling cycle in the mobile networks. In 4G LTE, there are 14 OFDM symbols in a slot or a TTI, which leads to a slot length of 1ms at 15kHz SCS. The 5G NR defines the mini-slots with lengths of 2, 4 and 7 symbols in the specification of standard. With higher SCS and less OFDM symbols, the time duration to transmit a mini slot can dramatically decrease, e.g., 36 μ s for mini-slot at 60kHz SCS with 2 OFDM symbols (see Table I), which is beneficial for ultra-low latency services.

Recently, fixed and mobile network infrastructures are showing to converge in the perspective of 5G and beyond [2]. A strong motivation of this convergence is driven by a cost reason. The radio access network (RAN) can share the existing fixed access infrastructure with mobile fronthaul/backhaul to connect functional blocks of RAN, referred as radio unit (RU), distributed unit (DU) and central unit (CU) defined by 3GPP [3]. This sharing or convergence of fixed and mobile infrastructures will allow reducing costs and time for cost-sensitive deployments of small cells. Passive optical networks (PON) have been successfully demonstrated as a cost-effective fixed/wired broadband access solution for residential users and business districts in the past few years, and now it is being evolved to accommodate emerging mobile services such as 5G small cells. Among the PON technologies, time and wavelength multiplexed passive optical network (TWDM-PON) is considered as an effective solution for fixed and mobile services convergence, in which multiplexing in both the time and wavelength domain is suitable for serving a large number of broadband clients (e.g., small cells and residential users) with bursty traffic patterns exploiting statistical multiplexing between clients.

A. Latency reduction on a TDM-PON based fronthaul

To accommodate fronthaul traffic in a TDMA subsystem (either TDM or TWDM), upstream latency is a big challenge. There are two main factors that affect fronthaul latency greatly. One of them is caused by the discovery window, which is periodically opened by an optical line terminal (OLT) for registering newly connected optical network units (ONU) and

Jiawei Zhang, Xiaozhu Liu, Minghui Liang and Yuefeng Ji are with the State Key Laboratory of Information Photonics and Optical Communications, Beijing University of Posts and Telecommunications, Beijing 100876, China.

Hao Yu is with the Aalto University, Espoo, Finland. (corresponding author e-mail: zjw@bupt.edu.cn).

ranging between OLT and ONU. This window will affect the transmission of fronthaul ONU and increase the upstream latency. It could be reduced to about 5 μ s in [4] through a novel intra-channel activation, called whisper ranging, to avoid the interruption of low-latency services. In addition, a recent study proposed to assign the discovery window by using network slicing [5], which is to accommodate fronthaul traffic and discovery window through different slices. Another biggest latency consumer is the procedure of dynamic bandwidth allocation (DBA), which is focused in this paper. Status-report DBA (SR-DBA) is a widely used and bandwidth-effective DBA scheme for conventional fixed access services, e.g., fiber to the home (FTTH). SR-DBA relies on the “report and grant” mechanism, in which each optical network unit (ONU) reports its buffer status to the optical line terminal (OLT), and OLT assigns the upstream bandwidth according to the received ONUs’ reports. Due to the report-grant mechanism, SR-DBA will bring into a millisecond delay, which is unable to meet the latency requirement of fronthaul. There are many related works focused on shortening the length of legacy SR-DBA cycle. *Tashiro et al.* proposed and demonstrated a novel DBA scheme called cooperative DBA (CO-DBA) [6], and it is soon be standardized by ITU-T [7, 8]. Besides the ITU-T, the O-RAN alliance is working on the next generation of RAN equipment, which specifies the interface between CU/DU and transport network (e.g., PON), called cooperative transport interface (CTI) [9]. This is a big progress to solid this technology to practical application from the perspective of wireless network. CO-DBA depends on the cooperation between CU/DU and OLT. CU/DU informs OLT, through a dedicated signaling (e.g., *mobile-info* message in [10]), about the upcoming traffic load of a RU that connects to an ONU. Then OLT allocates fronthaul bandwidth to the ONUs according to the *mobile-info* message without the need of report from ONU. As a result, the latency caused by the “report and grant” mechanism can be significantly reduced. Based on the CO-DBA, some research works made further efforts on the latency reduction. For example, *Ou et al.* [11] proposed to divide a wireless TTI into multiple transport packets that are transmitted between ONU and OLT. It reduced the waiting time and sending time of fronthaul traffic. In addition, *Sato et al.* proposed a cooperative resource control system based on CO-DBA [12], and demonstrated the feasibility of TDM-PON based fronthaul. Also, *Uzawa et al.* improve the practicality of CO-DBA by considering the actual data arrival period of fronthaul traffic

[13]. In [14], *Daisuke et al.* proposed a two-stage optimization of uplink forwarding order by considering the fronthaul characteristics to improve the bandwidth utilization. Besides the CO-DBA, another work to eliminate the uplink DBA process is called mobile-PON [15] proposed by *Zhou et al.* The idea of mobile-PON is to divide the uplink of PON into time slots or transport blocks (TB), where each TB carries a unique wireless resource block (RB). Based on it, two TB-RB mapping schemes (i.e., static and dynamic) were proposed. The static one uses a fixed TB-RB mapping (like fix bandwidth allocation), so that the latency is low but bandwidth efficiency is also low. The dynamic approach can achieve both low latency and high bandwidth efficiency, but it requires the collaboration between CU/DU and OLT, similar to CO-DBA. The above works considered the TDM-PON to carry mobile traffic only, which is not practical and cost-efficient in real network scenarios. For this reason, *Nomura et al.* proposed a DBA scheme that combines CO- and SR-DBA for fixed and mobile services accommodation into a single PON infrastructure [16]. They experimentally demonstrated the two DBAs worked together to achieve low-latency and high bandwidth efficiency.

B. Motivation and our contribution

Compared with SR-DBA, CO-DBA can be seen as a “pre-know and grant” mechanism, for which OLT can calculate an upcoming fronthaul bandwidth in advance according to the *mobile-info* message. Let T_p denote the time interval between when OLT receives the *mobile-info* message and when an upcoming fronthaul data (indicated by the *mobile-info* message) arrives at an ONU. The OLT performs CO-DBA and sends grant messages during the period of T_p . In a combined CO- and SR-DBA scenario, T_p is an important factor that should be longer than a length of SR-DBA cycle to guarantee the fronthaul bandwidth allocation without being affected by FTTH traffic. On one hand, a SR-DBA cycle is typical 1ms or 1.5ms in the most of literatures, by considering of the balance between time-tolerant FTTH services and high bandwidth efficiency of PON [6], [15]–[17]. On the other hand, T_p depends on the wireless data plane latency (further discussed in Section II). In a slot/TTI based 4G LTE, T_p is about 4ms [18], which is larger than a typical SR-DBA cycle. However, T_p will dramatically decrease to a small value in a mini-slot based 5G NR (smaller than a legacy SR-DBA cycle), and this will cause a resource allocation conflict between CO-DBA and SR-DBA. Fig. 1

Fig. 1. An example of bandwidth allocation conflict between CO- and SR-DBA for mini-slots.

Fig. 2. The compositions of time duration of T_p .

shows an example of the conflict between the two DBAs. The transmission of mini-slots $m-n$ (red rectangles) will conflict with FTTH services (yellow rectangles). This is because when the OLT receives the *mobile-info* message for mini-slot m at $t_{info,m}$, the uplink bandwidth has already been allocated by SR-DBA at t_{i-1} . Thus, there is no room for mini-slots $m-n$ when they arrive at ONUs at a later time. Therefore, the conflict mini-slots $m-n$ have to wait until the end of SR-DBA cycle i , and this introduces a large delay which might not meet the latency requirement. Note that in Fig. 1, the two OLT lines represent the same OLT physical entity, which are two receiving ends for fronthaul traffic and FTTH traffic respectively.

All the previous works [6], [10]-[16] did not consider the impact of this conflict, because their models were based on the slot/TTI model, in which the T_p (4ms) is larger than a typical SR-DBA cycle. As analyzed above, OLT has enough time to allocate bandwidth for fronthaul ONUs in advance, and no conflict happens. To the best of our knowledge, we are the first to study the mini-slot based 5G NR over a TDMA based PON. To solve the latency problem, we propose a novel dynamic wavelength and bandwidth allocation (DWBA) scheme by means of a wavelength and bandwidth preemption (WBP) mechanism. The basic rule of WBP is that a time-sensitivity fronthaul traffic can preempt the wavelength and bandwidth allocated to the time-tolerant traffic when a conflict happens. We compare it with an existing integrated CO- and SR-DWBA scheme [16] without preemption in many aspects.

The rest of this paper is organized as follows. In Section II, we analyze the latency model of T_p for mini-slot based 5G NR, which is an important part to our motivation. In Section III, the procedure of the proposed DWBA and WBP policies are detailed. In Section IV, simulation results are presented to compare the performance of our DWBA with an existing approach under various situations. Section V concludes the paper.

II. RETHINKING THE LATENCY MODEL OF T_p IN THE MINI-SLOT BASED 5G NR

The time duration of T_p is a key factor for applying CO-DBA in a T(W)DM-PON system. In this Section, we give the latency model of T_p , and analyze how it impacts on the CO-DBA in a mini-slot based 5G NR. As shown in Figs. 1 and 2, T_p is the time duration from when OLT receives the *mobile-info* message to when fronthaul traffic is arrived at an ONU. It is mainly determined by the wireless user plane latency. *Scheduling-info* is a wireless signaling to inform a user entity (UE) about the granted uplink bandwidth of the air interface. It is delivered via the physical downlink control channel (PDCCH) between CU/DU and UEs. Typically, there are several UEs are connected to an RU, and several RUs are connected to a CU/DU. A *scheduling-info* message contains granted bandwidth for all the UEs in the same CU/DU. Suppose the mobile system and PON system are time synchronized, and *scheduling-info* message and *mobile-info* message are sent out by CU/DU simultaneously. Typically, CU/DU and OLT, as well as RU and ONU, are geographically co-located. Thus, the propagation

delay between them could be ignored. We also suppose UEs are very close (e.g., tens or hundreds of meters) to RUs for small cell, so the propagation delay over the air could be also ignored. In addition, a strong RU processing ability is assumed, that the processing delay of RU can be ignored comparing with UE. With the above facts, T_p can be expressed as in (1), and its compositions are illustrated in Fig. 2 as well.

$$T_p = T_{PON\_prop} + T_{OLT\_fa} + T_{UE\_rx} + T_{UE\_tx} + T_{UE\_fa} + T_{UL\_duration} \quad (1)$$

T_{PON_prop} is the PON propagation delay that depends on the length of fiber span. Suppose the maximal distance between OLT and fronthaul ONUs is 10km, thus the maximal one-way latency of T_{PON_prop} is 50us (i.e., 5us/km). T_{OLT_fa} is the frame alignment delay of OLT, which is the time required after being ready to transmit a frame until transmission actually starts. There are two widely used ways to present frame alignment delay, which are average frame alignment time and worst frame alignment time. The average frame alignment time assumes that the data arrives at any time of a PON frame, thus T_{OLT_fa} is 0.5 times the length of a PON frame in this case. The worst frame alignment time assumes that every data has to wait a whole frame (worst case) to be transmitted at the next frame, thus T_{OLT_fa} is the length of a PON frame. Considering a GPON system (i.e., frame length is 125us), the T_{OLT_fa} is 62.5us for the average case and 125us for the worst case. T_{UE_rx} and T_{UE_tx} are the processing delay in the UE. T_{UE_rx} is the time interval between UE receives the *scheduling-info* message and decodes it. This latency depends on the UE processing capability and SCS, which is typically below 100us according to the T_{UE_rx} . The calculation model of T_{UE_rx} and T_{UE_tx} is detailed in the Appendix Section. T_{UE_fa} is the frame alignment delay of UE, as similar to T_{OLT_fa} . In the average case, T_{UE_fa} is 0.5 times the length of a mini-slot, while it equals to the length of a mini-slot in the worst case. $T_{UL\_duration}$ is the time for transmitting a basic wireless frame, which equals to the length of a mini-slot. In a given PON system, T_p is relevant to the length of a mini-slot. When the length of a mini-slot decreases, the value of T_p will decrease as well. If T_p is smaller than a SR-DBA cycle, the

TABLE I [1], [21, 22]
KEY PARAMETERS OF MINI-SLOTS WITH DIFFERENT NUMBER OF OFDM SYMBOLS AND DIFFERENT SCS SCHEMES

SCS	Parameters	N_{sym}		
		7 symbols	4 symbols	2 symbols
15 kHz	Symbol duration (us)	66.67		
	Mini-slot duration (us)	500	286	143
	Max. Num. of RBs	270		
	D_{FH} (kbits)	1451.52	829.44	414.72
30 kHz	Symbol duration (us)	33.33		
	Mini-slot duration (us)	250	143	71
	Max. Num. of RBs	133		
	D_{FH} (kbits)	715.008	408.576	204.288
60 kHz	Symbol duration (us)	16.67		
	Mini-slot duration (us)	125	71	36
	Max. Num. of RBs	65		
	D_{FH} (kbits)	349.44	199.68	99.84

$T_{convert}$, T_{CO-DBA_proc} and T_{OLT_fa} respectively. The frequency of sending *scheduling-info* message and *mobile-info* message is $\frac{1}{T_{min-slot}}$, where $T_{min-slot}$ is the time duration of a mini-slot. Generally (without a specific design), if OLT receives a *mobile-info* message, then it triggers CO-DBA process. However, as $T_{min-slot}$ decreases, the frequency of CO-DBA process is growing up. To improve the CO-DBA efficiency, OLT should be allowed to collect a batch of mobile-info messages in a short time (i.e., $T_{collect}$ in Fig. 2), and grants bandwidth for several mini-slots together. $T_{collect}$ can be obtained based on the lengths of T_p , $T_{convert}$, T_{CO-DBA_proc} and T_{OLT_fa} .

Fig. 3 (a) illustrates an example in which an OLT grants three mini-slots together after receiving three *mobile-info* messages during a $T_{collect}$. In addition, for each ONU, it periodically receives fronthaul data from its corresponding RU in every $T_{min-slot}$ cycle. According to the frequency multiplexing in wireless networks, some RUs may share the whole frequency resources, as shown in Fig. 3(b). These RUs will send fronthaul data to the ONUs within a same $T_{min-slot}$ cycle. During a $T_{min-slot}$ cycle, the DU will receive all the fronthaul data from multiplied RUs for processing. For consecutive $T_{min-slot}$ cycles, the DU will process them in order. The volume of fronthaul data varies with RU's load under a lower functional split option, e.g., option 7-2.

III. DWBA PROCEDURE AND WBP POLICIES

The proposed DWBA aims to address the delay caused by the conflict between CO- and SR-DWBA in a TWDM-PON system, which is enabled by the WBP policy. In this Section,

Fig. 4. Illustration of possible cases in a 25G PON: (a)/(b) conflict/preemption in one ONU's grant period; (c)/(d) conflict/preemption across two ONUs' grant period.

the procedure of our DWBA will be detailed, and different WBP policies are proposed for different scenarios and purposes.

Our DWBA allows the conflict mini-slots (high priority) to preempt the wavelength and bandwidth resources allocated to FTTH services (low priority). Fig. 4 (a) and (c) illustrate two possible conflict cases in a TWDM-PON system with 25Gbps upstream data rate. Fig. 4 (b) and (d) show our DBA procedures to deal with the two conflict cases. Let's take Fig. 4 (a) as an example, where conflict and pre-emption happens in one ONU's grant time. We have three ONUs connected to an OLT, one of them is a fronthaul ONU, and the others are FTTH ONUs. Suppose all the ONUs use the same wavelength (λ_1) for upstream transmission. The OLT grants the transmission time t_1 and t_2 for the bursts of FTTH ONUs (B_1 and B_2) through SR-DBA. After that, OLT receives the *mobile-info* message for the incoming fronthaul burst, and this will cause a conflict with FTTH burst B_1 according to the previous analysis. With pre-emption, our DBA scheme allows OLT allocate resource for fronthaul burst preferentially, and reallocate the bandwidth and wavelength for the conflict FTTH burst which has been previously assigned through SR-DBA. As shown Fig. 4(b), t_{pre} denotes the preemption time of pre-emptive traffic (i.e., fronthaul). The expected earliest time of t_{pre} should be equal to the arrival time of fronthaul traffic $t_{arrival}$. Due to the preemption, the original FTTH burst will be divided into three parts (or bursts) for reallocation. The first burst (450 kbits) and second burst (1050 kbits) are unaffected parts (B_{unafct}), which means that they can be transmitted during the previous granted time slot for the original FTTH burst. The third burst (580 kbits) is the part that is squeezed out of original burst by the preemption. Without affecting other FTTH ONUs' DBA result,

we would like to arrange this part (i.e., the purple rectangle) at the end of SR-DWBA cycle for reallocation, rather than extending size of the original allocated burst of the pre-empted ONU. We call this part "delay burst" (B_{delay}), which will increase the latency of FTTH traffic. We denote $t_{realloc}$ as reallocated transmission time. Both t_{pre} and $t_{realloc}$ will be obtained by WBP policy and sent out by OLT via grant messages to the pre-emptive ONU and the preemptable ONU respectively. After receiving the grant messages, the fronthaul ONU starts to transmit its traffic at t_{pre} then the FTTH ONU will divide its data into blocks and transmit them at different $t_{realloc}$. Fig. 4 (c) and (d) show the conflict and pre-emption across two FTTH ONUs, and the preemption will delay two FTTH bursts.

The WBP policy is to obtain two kinds of system time (i.e., t_{pre} and $t_{realloc}$), among which the pre-emption time determines the reallocated transmission time. Based on the above conflict and pre-emption cases, we propose three straightforward but representative policies for different scenarios. In **Policy1 (P1)**, fronthaul traffic can pre-empt FTTH traffic immediately when it arrives at an ONU, so the pre-emption time equals to the arrival time of fronthaul traffic. P1 is more suitable for a scenario that consists of mobile services with ultra-low latency requirement and FTTH services with best effort. In **Policy2 (P2)**, the pre-emption across two FTTH ONUs is not allowed, because it affects the average FTTH latency by delay two FTTH ONUs' bursts. When a fronthaul traffic arrives at the border of two FTTH ONUs, it has to wait until the transmission of the first ONU is finished, thus may introduce a very short fronthaul delay. In **Policy3 (P3)**, the minimal length of reallocated burst is constrained, because a smaller burst will reduce bandwidth efficiency (i.e. the ratio of payload/(payload+overhead)). The overhead includes the guard band, preamble and delimiter, as well as the Forward Error Correction (FEC) which brings about 13% redundancy by using the (248,216) Reed Solomon code [19]. Here, we define B_{min} as the lower bound of a reallocated burst. Therefore, when a FTTH burst is shorter than B_{min} , it cannot be pre-empted by a fronthaul ONU (unpreempted burst).

In summary, P1 is the most aggressive preemption policy, which can achieve the lowest latency performance among them. However, it introduces more "delay bursts" of FTTH traffic, and brings more overheads to lower the bandwidth efficiency. P2 limits the pre-emption across two FTTH ONUs that reduce the number of delay bursts and lower the complexity of preemption. P3 is a policy that balance the tradeoff between latency and bandwidth efficiency through limiting the size of a reallocated burst.

For the three policies, we can extend them for supporting multiple wavelengths, and choose the best wavelength by considering of pre-emption time, latency influence on FTTH traffic, and bandwidth efficiency. The general procedure of our DWBA based on WBP policy is described in Algorithm 1. When the OLT receives a *mobile-info* message, it then calculates the fronthaul bandwidth of a mini-slot that will arrive

---

Algorithm 1: General procedure of our DWBA with WBP

- 1: **WHEN** a *mobile-info* message for an upcoming mini-slot is received.
  - 2: Calculate the fronthaul bandwidth of this mini-slot.
  - 3: **FOR** each wavelength i in a TWDM-PON system
  - 4: Check the priority of the ongoing traffic of wavelength i when conflict happens.
  - 5: **IF** the priority is high (preemption unable), wait for the next low priority traffic and obtain the preemption time t_{pre}^i according to the rule of WBP policy.
  - 6: **ELSE** obtain the preemption time t_{pre}^i of fronthaul traffic according to the rule of WBP.
  - 7: **END IF**
  - 8: Obtain the reallocated transmission time $t_{realloc}^i$.
  - 9: **END FOR**
  - 10: Choose the earliest preemption time t_{pre}^i among the wavelengths.
  - 11: **IF** more than one wavelength has the same t_{pre}^i ,
  - 12: Choose the wavelength which has the minimal number of preemption.
  - 13: **IF** more than one wavelength has the same number of preemptions, then choose the first available wavelength.
  - 14: **END IF**
  - 15: **END IF**
-

later. Then, we obtain the earliest preemptable time t_{pre}^i for wavelength i (line 4-7) according to the above-mentioned WBP policies. Based on the t_{pre}^i , we can determine the reallocated transmission time $t_{realloc}^i$. Among all the candidate wavelengths, we choose the wavelength that has the earliest t_{pre}^i . If more than one wavelength has the same preemption time, we choose the wavelength which has the minimal number of preemption. Otherwise, we choose the first available wavelength among the all (First Fit). Note that, if there is no conflict happens, we allocate fronthaul bandwidth based on CO-DWBA.

IV. ILLUSTRATIVE NUMERICAL EXAMPLES

In this Section, we compare the proposed DWBA with an existing integrated CO- and SR-DBA scheme without preemption [16]. In Ref. [16], they used slot/TTI based model, while this paper considers a mini-slot based model. In addition, we evaluate the performance of different WBP policies under different situations.

A. Simulation setup

We consider a TWDM-PON system with four wavelengths, and each of them supports 25Gbps. There are 32 ONUs in a TWDM-PON, and the distance between ONU and OLT follows a Gaussian distribution (μ, σ) , where the mean is $\mu = 10\text{km}$ and the variance is $\sigma = 2\text{km}$. The propagation delay over fiber is 5us/km. The guard band between the two adjacent bursts is 1 us, and the (248,216) Reed Solomon code is used in the simulation that introduces about 13% overheads. For the P3, the minimal preemptable burst length (B_{min}) is set to 9us for the Figs. 5-9, but it varies in Fig. 10 to verify its impact on P3. Let γ be the proportion of fronthaul ONUs to the total ONUs, which is a variable parameter in the simulation. The overall traffic load, denoted by ρ , is distributed to fronthaul ONUs and FTTH ONUs according to γ , in which fronthaul ONUs produce $\rho\gamma$ traffic load and the FTTH ONUs take the rest, i.e., $\rho(1 - \gamma)$. The value of ρ is normalized in the simulation, and the maximum value is 1. Fronthaul traffic is a periodic flow, and the cycle is the time duration of a mini-slot $T_{min-slot}$. Fronthaul data rate depends on the functional split option. Here, we adopt option 7-2 of 3GPP for uplink [20], where we split between channel estimation and resource element de-mapping. The fronthaul data rate depends on the actual traffic load. For a mini-slot in the uplink, the fronthaul data volume (D_{FH}) can be calculated by (2), according to the specification of Small Cell Forum [21].

$$D_{FH} = N_{ant} \times N_{iq} \times N_q \times N_{sc} \times N_{sym} \times N_{rb} \quad (2)$$

where N_{ant} is the number of antennas, which is 2 in the simulation. N_{iq} denotes the number of in- and quadrature-phase parts (=2), and N_q is the number of quantization bits of the in- and quadrature-phase parts, which are typically 16 bits. N_{sc} is the number of sub-carriers per RB, which is 12 for both 4G and 5G systems. In this study, we consider an FDD mobile system, and discuss three sub-carrier spaces, which are

TABLE II
SIMULATION PARAMETERS

Parameter		Value
PON system	# of wavelengths	4
	Capacity of wavelength	25Gbps
	Fiber propagation delay	5 us/km
	# of ONUs	32
	Proportion of Fronthaul ONU	γ
Mobile service	Normalized network load	ρ
	SCS	15, 30, 60 kHz
	# of OFDM symbols in a mini-slot	2, 4, 7
	Split option	Option 7-2
	Mobile system bandwidth	50 MHz
Fixed service	# of allocated RB	ρN_{RB}
	Packet size	[64-1518] bytes
	Packet arrival time	Poisson
	SR-DWBA cycle	1ms

15kHz, 30kHz, and 60kHz. At a given wireless frequency width, the maximum number of RBs will be changed for different SCS schemes. Table I shows the maximum number of RBs for different SCS schemes under 50MHz frequency width [22]. Note that, the maximum number of RBs is independent with the number of OFDM symbols. N_{sym} denotes the number of OFDM symbols in a mini-slot. N_{rb} is the number of occupied RBs. It varies with traffic load that equals to ρN_{RB} , where N_{RB} represents the total number of RB in a mini-slot. Table I shows the D_{FH} under different SCS schemes and OFDM symbols. The SR-DWBA cycle is set to be 1ms (the same as simulation cycle), during which the FTTH ONUs collect packets whose sizes are randomly distributed in the range of [64, 1518] bytes. The traffic load is distributed among the FTTH ONUs following a Gaussian distribution, and it will be discussed in the simulation result (see Fig. 8). For each FTTH ONU, the packet arrival time follows a Poisson distribution. We let λ denote the average packet arrival rate, which equals to FTTH traffic load divided by the packet length. Therefore, the average packet interval time can be expressed as $t = -\frac{1}{\lambda} \ln U$, where U is randomly distributed in [0,1] and λ is Gaussian distribution with mean $\rho(1 - \gamma)$ and variable variance divided by the packet length. Some key parameters of simulation environment setup are detailed in Table II. Simulation are experimented over 10^5 simulation cycles, and the results are shown below.

B. Numerical results and analysis

Emerging applications in 5G and beyond (e.g., industrial communications) require a stringent latency bound (a.k.a., deterministic latency), like 99.999% of the packets should be transmitted at a very low latency. To evaluate the performance of deterministic low latency of the proposed DWBA, we take the upper bound latency (worst case) as a main metric, rather than average latency that evaluates the conventional services.

1) Latency distribution

Latency distribution is an import factor to reflect the upper and lower bounds of latency. Fig. 5 shows the latency distribution of different DWBA policies under 15kHz SCS with 2, 4, and 7 OFDM symbols. Each bar in the figure corresponds to a 10us time slot on the x-axis, and y-axis shows

Fig. 5. Frequency of latency under half and full load with different OFDM numerologies: (a) 2 OFDM symbols; (b) 4 OFDM symbols; (c) 7 OFDM symbols.

the frequency of frames that appear in each time slot. We consider both half loaded (red bar) and full loaded (blue bar) scenarios, and the proportion of fronthaul ONUs is $\gamma = 0.5$. As expected, our DWBA (with WBP policies) have a lower upper bound of latency than integrated CO- and SR-DWBA without preemption. This is because our DWBA introduces the preemption mechanism that allows the conflict fronthaul traffic

Fig. 6 Upper bound latency vs. traffic load: (a) 2 OFDM symbols with different SCS schemes; (b) 30KHz SCS with different OFDM numerologies.

to preempt the wavelength and bandwidth of FTTH traffic. Thus, when a fronthaul traffic arrives at an ONU, it can be transmitted immediately or in a very short waiting time. Another observation is that our DWBA schemes have a relevant shorter tail than the scheme without preemption. Most frames distribute in a small area of latency, which shows good performance on the jitter.

Among the WBP policies, we find that P1 has a better latency performance than P2, especially for the high loaded network. This is because P1 allows the pre-emption across two FTTH ONUs while P2 does not. In addition, we observe that P2 is better than P3 in Fig. 5(a), and worse than P3 in Fig. 5(b) and (c). This is because when the number of OFDM symbols growing, the size of a mini-slot grows as well. It increases the probability of cross-ONU preemption. However, this phenomenon likely occurs in a large-sized mini-slot scenario.

2) Upper bound of latency (worst case) vs. traffic load.

Fig. 6 shows the upper bound latency under variable traffic load. The proportion of fronthaul ONUs is $\gamma = 0.5$ in this case. In Fig. 6(a), we fix the number of OFDM symbols (2 symbol per mini-slot) and vary the SCS. On the other hand, the SCS is fixed at 30KHz, and we change the

Fig. 7 Average FTTH latency vs. traffic load: (a) 2 OFDM symbols with different SCS schemes; (b) 30kHz with different OFDM numerologies.

Fig. 8 Upper bound latency vs. proportion of fronthaul ONUs: (a) 2 OFDM symbols with different SCS; (b) 30kHz with different OFDM numerologies.

number of OFDM symbols in Fig. 6(b). With the growing traffic load, the upper bound of latency increases, among which P1 have a best performance to achieve the lowest bound of latency. Even in a full loaded condition ($\rho = 1.0$), the worst latency (upper bound) is still below 250us, a common latency restriction for fronthaul interface at option 7-2 for time-tolerant services [20]. For time-sensitive services, such auto-driving, fronthaul latency is more stringent that is determined by the service requirement, which is lower than 100us in [23]. If we further reduce the latency limitation to 100us, we see that our policies still keep good performance under a moderate traffic load ($\rho = 0.8$).

Another observation is that the latency of P3 is quickly growing up when traffic load $\rho > 0.5$. The reason is mainly caused by the unpreemptable bursts of FTTH, whose lengths are less than B_{min} . As traffic load growing, the number of unpreempted bursts grow as well, which increases the fronthaul latency. As shown in Fig. 6, the increased fronthaul latency of P3 is significant when the mini-slot size is small (see 60kHz SCS situation in Fig. 6(a) and 2 OFDM symbols situation in Fig. 6(b)). We also find that the largest latency reduction by P1 is achieved at these two situation, which are 81.0% and 77.6%

respectively. It is obtained that the proposed DWBA is more suitable for the mini-slot with a smaller size.

3) Average FTTH latency vs. traffic load

With the same simulation environment of Fig. 6, Fig. 7 shows the average FTTH latency. We see that our DWBA schemes cause higher FTTH latency comparing with the scheme without preemption. This is because the WBP lags the transmission of FTTH traffic. The reduction of fronthaul latency is achieved through sacrificing the performance of FTTH traffic. The maximal latency gap appears on a smaller mini-slot situation (e.g., larger SCS when the number of OFDM is fixed, or fewer OFDM symbols when the SCS is fixed). We also observe that P2 has a lower FTTH latency than P1, especially for a larger mini-slot. By reducing fronthaul latency as the first optimization goal, we find that P2 achieves a good balance between fronthaul and FTTH latency compared with other schemes.

4) Upper bound of latency vs. proportion of fronthaul ONU

How many fronthaul ONUs can be deployed to achieve a low upper bound of latency is an important issue for network operator at the planning stage. To this end, we set up the simulation to fix the traffic load is $\rho = 1.0$, and vary the

Fig. 9 Upper bound latency vs. load deviation-mean ratio of FTTH ONU.

proportion of fronthaul ONU in the TWDM-PON system. The results are shown in Fig. 8. As the proportion of fronthaul ONUs (γ) growing up, the black curve (scheme without preemption) is going down because there is less probability of conflict between CO- and SR-DWBA. Compared with black curve, we observe that a good bounded latency performance can be achieved through the proposed DWBA scheme. This is because most of fronthaul traffic can benefit from WBP, which are transmitted immediately when they arrive at ONUs. However, as the ratio of fronthaul ONUs growing, the successful rate of preemption of P3 will decrease, because more and more reallocated bursts are emerged, which are too short to be preempted. That is why the curves of P3 growing up when γ increases to 0.7.

Another observation can be found that the curves of P1 and P2 go steady when γ is below 0.5, which shows an useful tip for planning the number of fronthaul ONUs and FTTH ONUs in a single TWDM-PON system. We also find that, P1 and P2 are growing up quickly for the mini-slots with a smaller size (e.g., 60kHz SCS in Fig. 8(a) and 2 OFDM symbols in Fig. 8(b)). This is because a large number of fronthaul bursts are generated when the size of mini-slot is small, and they will compete the resources with each other.

5) Upper bound of latency vs. FTTH load distribution

For the above simulation setup, we suppose the traffic load is uniformly distributed among the FTTH ONUs. In a real network scenario, traffic load may not equal among the ONUs. Here, we consider another load distribution, Gaussian distribution, to evaluate its impact on WBP policies. Fig. 9 shows the upper bound of latency under different load deviation-mean ratios of FTTH ONU. We set SCS to be 30kHz, and increase the number of OFDM symbols. The proportion of fronthaul ONU is $\gamma = 0.5$, and traffic load is $\rho = 1.0$. We find that a large deviation-mean ratio affects P3 seriously, especially for a small mini-slot (e.g., Fig. 9(a)), while the effects on P1 and P2 are slight. This is because a heavy loaded FTTH ONU will be preempted many times, which results in more small bursts that cannot be preempted again according to

Fig. 10 PON US efficiency vs. B_{min} : (a) 2 OFDM symbols with different

the rule of P3. It increases the bounded fronthaul latency. However, for a larger mini-slot (e.g., Fig. 9(c)), the effects on P3 is slight.

6) PON upstream bandwidth efficiency vs. B_{min}

The WBP policies will generate a large number of small bursts (i.e., reallocated burst of FTTH ONU), which impacts the PON upstream bandwidth efficiency significantly. Therefore, P3 is designed to balance the latency and the bandwidth efficiency by designing the minimal length of a burst (i.e., B_{min}). How small a burst is can achieve good bandwidth efficiency should be discussed in this work. To this end, we set up the simulation to fix the traffic load $\rho = 1.0$ and the proportion of fronthaul ONU $\gamma = 0.5$, then vary the value of B_{min} . The results are shown in Fig. 10. We find that P1 has the lowest bandwidth efficiency followed by P2, who both have aggressive preemption mechanism that may cause very small bursts of FTTH ONUs. These policies can achieve a lower fronthaul latency (see Fig. 6) at the expense of bandwidth efficiency. With the increasing of B_{min} , the curve of P3 is first going up, and then becoming steady after B_{min} is reaching some points. The value of these points depend on different mini-slot structures, i.e., SCS schemes and number of OFDM symbols.

Another observation can be found that, the smaller the mini-slot is, the lower the bandwidth efficiency is. This is because a

small mini-slot based fronthaul will bring the same-sized “delay burst” of FTTH, which results in a low bandwidth efficiency. In addition, a higher SCS scheme will increase the frequency of preemption, thus also cause low bandwidth efficiency.

V. CONCLUSIONS

For a multiple-service TWDM-PON system that converges mobile fronthaul and traditional PON services, we proposed a novel DWBA, to solve the conflict between CO- and SR-DWBA, which is caused by the mini-slot based 5G NR. A key part of the proposed DWBA is the wavelength and bandwidth preemption mechanism, which can provide low bounded fronthaul latency. We described the whole procedure of our DWBA and proposed three WBP policies for different situations. To evaluate the performance of the DWBA, we compared it with an existing integrated CO- and SR-DWBA scheme without WBP, and it was demonstrated that a significant latency reduction was achieved by using a hard WBP strategy (P1). Furthermore, we analyzed the proposed DWBA from other aspects, such as average FTTH latency, proportion of fronthaul ONU, FTTH traffic load distribution and upstream bandwidth efficiency. In a summary, we understand that the low fronthaul latency was achieved through sacrificing the latency performance of FTTH traffic and the bandwidth efficiency. The low bounded latency cannot be guaranteed for all the situations all the time. It depends on how hard strategy we use, and how do we balance the improvement of fronthaul and the degradation of FTTH traffic.

Our future work will focus on the issues that are induced by our DWBA with WBP. For the T(W)DM-PON system, it brings a signaling complexity in OLT for operating preemption. The frequency of sending grant messages should be well considered, i.e., how to design the length of $T_{collect}$. For the algorithm, the benefit of the proposed DWBA for fronthaul traffic is to sacrifice the latency performance of FTTH traffic and bandwidth efficiency, thus a compromise between FTTH traffic and fronthaul traffic should be discussed in the future. Also, this paper treats all the fronthaul traffics with the same priority, which could be further improved by differentiating the fronthaul traffics in a scene-defined network.

APPENDIX

A. Computation model of $t_{UE_{rx}}$ and $t_{UE_{tx}}$

The computation model of $t_{UE_{rx}}$ and $t_{UE_{tx}}$ is complicated and varies with different conditions. Based on the standard specification [24]–[27] and the scenario of this work, we construct the model under the following assumptions: 1) we adopt UE processing capability 2 (high UE processing ability); 2) and resource mapping type B for URLLC services as given in clause 7.4.1.1 of [24]; 3) exclude the latency of hybrid automatic repeat request (HARQ).

We obtain the computation model formula as $t_{UE_{rx}} + t_{UE_{tx}} = \frac{1}{2}(T_{proc,1} + T_{proc,2})$ according to [25]. $T_{proc,1}$ and $T_{proc,2}$ represent the UE processing procedure time for physical

downlink shared channel (PDSCH), and UE preparation procedure time for physical uplink shared channel (PUSCH) respectively. They are expressed as the following formulas:

$$T_{proc,1} = 2192 \times (N_1 + d_{1,1}) \cdot \kappa \cdot 2^{-\mu} \cdot T_c, \quad (3)$$

where N_1 is the decoding time based on μ of Table III. μ is zero if SCS is 15kHz, and when SCS is 30kHz, μ is one, and so on. The DM-RS in Table III represents the demodulation reference signals, and usually occupies 1 or 2 symbols. $d_{1,1}$ represents the overlapping symbols of the scheduling PDCCH and the scheduled PDSCH. In case of mini-slot (2, 4 and 7 symbols), $d_{1,1}$ is zero. The value of κ is 64 [26]. T_c is the basic time unit of 5G NR system and its value is 0.509ns [26].

$$T_{proc,2} = \max \{2192 \times (N_2 + d_{2,1}) \cdot \kappa \cdot 2^{-\mu} \cdot T_c, d_{2,2}\}, \quad (4)$$

where κ , μ and T_c are the same as the description above. N_2 is preparation time based on μ of Table IV. $d_{2,1}$ is zero if the first symbol of the PUSCH allocation consists of DM-RS only, otherwise $d_{2,1}$ is one. In this work, we let $d_{2,1} = 0$ according to [24]; $d_{2,2}$ equals to the switching time as defined in the clause 8.6.2 of [27]. If UE traffic load fluctuates widely that need to trigger a switch of BWP (bandwidth part), otherwise $d_{2,2}$ is zero.

Suppose we have a mini-slot with two OFDM symbols and the SCS is 60kHz. $d_{1,1}$, $d_{2,1}$ and $d_{2,2}$ are set to be zero, then we can obtain $(T_{UE_{rx}} + T_{UE_{tx}})$ is 178.5us. In addition, we consider the averaged frame alignment delay for uplink and downlink transmission, and PON distance is 10km. According to Eq. (1), we can calculate T_p is 345us. It is smaller than a typical SR-DBA cycle (1ms or 1.5ms). Therefore, the resource allocation conflict between CO-and SR-DBA will happen.

REFERENCES

- [1] J. Sachs, *et al.*, “5G radio network design for ultra-reliable low-latency communication,” *IEEE Netw.*, vol. 32, no. 2, pp. 24-21, Apr. 2018.
- [2] S. Gosselin, *et al.*, “Fixed and mobile convergence: Which role for optical networks?” *IEEE/OSA J. Opt. Commun., Netw.*, vol. 7, no. 11, pp. 1075-1083, 2015.
- [3] “Technical Specification Group Radio Access Network; NG-RAN; Architecture description,” 3GPP TS 38.401, V16.2.0, Rel.16, Jul. 2020. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3219>.
- [4] R. Bonk *et al.*, “Demonstration of ONU Activation for In-Service TDM-PON Allowing Uninterrupted Low-Latency Transport Links,” in *Proc. Opt. Fiber Commun. Conf.*, Mar. 2019, pp. 1-3.
- [5] H. Uzawa, *et al.*, “Dynamic bandwidth allocation scheme for network-slicing-based TDM-PON toward the beyond-5G era,” *IEEE/OSA J. Opt. Commun. Netw.*, vol. 12, no. 2, pp. A135-A143, 2020.
- [6] T. Tashiro, *et al.*, “A novel DBA scheme for TDM-PON based mobile fronthaul,” in *Proc. Opt. Fiber Commun. Conf.*, Mar. 2014, pp. 1-3.
- [7] “5G Wireless Fronthaul Requirements in a Passive Optical Network Context,” ITU-T G. Sup66, Rev. 1, July 2019. [Online]. Available: <https://www.itu.int/rec/TREC-G.Sup66/en>.
- [8] J.S. Wey, *et al.*, “5G wireless transport in a PON context: An overview,” *IEEE Commun. Standards Magazine*, vol. 4, no. 1, pp. 50-56, 2020.
- [9] “Cooperative Transport Interface Transport Control Plane Specification,” O-RAN Alliance, O-RAN.WG4.CTI-TCP.0-v01.00, April 2020. [Online]. Available: <https://static1.squarespace.com/static/5ad774cc74940d7115044b0/t/5eba08ea2f250476655438f/1589289114697/O-RAN.WG4.CTI-TCP.0-v01.00.pdf>.

- [10] H. Nomura, *et al.*, "First demonstration of optical-mobile cooperation interface for mobile fronthaul with TDM-PON," *IEICE Commun. Express*, vol. 6, no.6, pp. 375-380, 2017.
- [11] H. Ou, *et al.*, "Passive optical network range applicable to cost-effective mobile fronthaul," in *Proc. IEEE Int. Conf. Commun.*, May, 2016, pp. 1-6.
- [12] H. Sato, *et al.*, "Reducing Uplink Transmission Latency for Applying TDM-PON to Mobile Fronthaul," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 2018, pp. 1-3.
- [13] H. Uzawa, *et al.*, "Practical mobile-DBA scheme considering data arrival period for 5G mobile fronthaul with TDM-PON," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 2019, pp. 1-3.
- [14] H. Daisuke *et al.*, "Two-stage optimization of uplink forwarding order with cooperative DBA to accommodate a TDM-PON-based fronthaul link," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 12, no. 5, pp. 109-119, 2020.
- [15] S. Zhou *et al.*, "Low-latency high-efficiency mobile fronthaul with TDM-PON (mobile-PON)," *IEEE/OSA J. Opt. Commun. Netw.*, vol. 10, no.1, pp. A20-A26, 2018.
- [16] H. Nomura, *et al.*, "Novel DBA scheme integrated with SR-and CO-DBA for multi-service accommodation toward 5G beyond," in *Proc. Eur. Conf. Opt. Commun.*, Sep. 2019, pp. 1-3.
- [17] I.S. Hwang, *et al.*, "Polling cycle time analysis for waited-based DBA in GPONs," in *Proc. IEEE Int. Multi-Conf. of Engineers and Computer Scientists*, vol. 2, 2013.
- [18] A. Kurian, "Latency analysis and reduction in a 4G network," 2018, [online]. Available: <http://resolver.tudelft.nl/uuid:c1badd8d-a384-49a1-b958-a0c1e499c539>.
- [19] S. Bidkar, *et al.*, "Low-Latency TDM-PON for 5G Xhaul," in *Proc. Inter. Conf. on Transparent Opt. Netw.*, Jul. 2020, pp. 1-4.
- [20] "Study on new radio access technology: Radio access architecture and interfaces," 3GPP TR 38.801, V14.0.0, Rel.14, Mar. 2017. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3056>.
- [21] "Small cell virtualization: Functional splits and use cases," Small Cell Forum, SCF159.05.1.01, V 5.1, Jun. 2015. [Online]. Available: https://scf.io/en/documents/159_-_Small_Cell_Virtualization_Functional_Splits_and_Use_Cases.php.
- [22] "NR; Base Station (BS) radio transmission and reception," 3GPP TS 38.104, v. 16.4.0, Rel. 16, Jul. 2020. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3202>.
- [23] A. Checko, *et al.*, "IEEE 1914 NGFI (xhaul): efficient and scalable fronthaul transport for 5G," in *Proc. IEEE Workshop on next generation Backhaul/Fronthaul Netw.*, [Online]. Available: https://www.mti-mobile.com/wp-content/uploads/2017/09/IEEE1914_Checko_BackNets2017v1.2.pdf.
- [24] "Study on Self-Evaluation Towards IMT-2020 Submission," 3GPP TR 37.910, v.16.1.0, Rel. 16, Oct. 2019. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3190>.
- [25] "NR; Physical layer procedures for data," 3GPP TS 38.214, v. 16.2.0, Rel. 16, Jul. 2020. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3216>.
- [26] "NR; Physical channels and modulation," 3GPP TS 38.211, v. 16.2.0, Rel. 16, Jul. 2020. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3213>.
- [27] "Requirements for support of radio resource management," 3GPP TS 38.133, v. 16.4.0, Rel. 16, Jul. 2020. [Online]. Available: <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3204#>.