Virtualised EAST-WEST PON Architecture Supporting Low-Latency communication for Mobile Functional-Split Based on Multi-Access Edge Computing

SANDIP DAS[1,*], FRANK SLYNE[1], ALEKSANDRA KASZUBOWSKA[1], AND MARCO RUFFINI[1]

[1] The University of Dublin, Trinity College, CONNECT research centre.
[*] Corresponding author: dassa@tcd.ie

Compiled July 1, 2020

Ultra-low latency end-to-end communication with high reliability is one of the most important requirements in 5G networks to support latency-critical applications. A recent approach towards this target is to deploy edge computing nodes with networking capabilities, known as Multi-access edge computing (MEC), which can greatly reduce the service end-to-end latency. However, the use of MEC nodes poses radical changes to the access network architecture. This requires to move from the classical point-to-multipoint (or point-to-point) structure, used to deliver residential broadband and Cloud-RAN services, to a mesh architecture that can fully embed the MEC nodes with all other end points (i.e., mobile cells, fixed residential and businesses, etc.).

In this paper, we propose a novel PON based Mobile Fronthaul (MFH) transport architecture based on PON virtualisation, that allows EAST-WEST communication along with traditional NORTH-SOUTH communication. The architecture enables the endpoints of a PON tree, where usually ONUs are located, to also host MEC nodes by deploying an edge OLT capable of communicating directly with adjacent ONUs, by reflecting wavelength signals from the splitter nodes. We experimentally show that signal backscattering due to the reflection at the splitter does not affect the system performance. In addition, using protocol level simulations, we show how this architecture can maintain low-latency ($\approx 100\mu s$) in varying mobile traffic conditions by offloading ONUs (i.e., where remote units of Cloud-RAN cells are located) to other edge OLTs through dynamic formation of virtual PON (vPON) slices. Furthermore, our results show how an efficient migration strategy for ONUs can be chosen depending on the traffic load, different functional split configurations, and the PON capacity.

http://dx.doi.org/10.1364/ao.XX.XXXXXX

1. INTRODUCTION

As the deployment of 5G networks picks up pace, telecommunication industries are continuously challenged by the need to support ever-increasing data traffic demand, massive connectivity and highly diverse quality of service. Some of the key network requirements in modern 5G networks [1] include high throughput, ultra-low end-to-end latency and deterministic Quality of Service (QoS). Specifically, ultra-low end-to-end latency (1-10ms) has been a critical requirement in 5G networks that support various latency-critical 5G applications such as tactile internet, logistics, mission-critical control and traffic and road safety [2].

Cloud Radio Access Networks (C-RAN), along with Functional Split processing, are regarded as the most promising 5G radio technologies that support these requirements. In the 5G New Radio (5G-NR) architecture of C-RAN, the baseband processing functions are split into three parts: Central Unit (CU), Distributed Unit (DU) and Remote Unit (RU) [3]. The CU and

DU processing functions of several cells are centralised and virtualised in either at a Central Office (CO) or at a nearby cloud edge processing site, while the cell processing part is retained at the cell site and is called RU. This architecture better facilitates Radio Access Networks (RAN) virtualisation with flexible assignment of computing resources across different entities. The distribution of baseband processing functions between CU, DU and RU is identified by 8 split points in the Long Term Evolution (LTE) baseband processing chain, where split-8 is the legacy CPRI split as in 4G/LTE, and split-1 is identified between Packet Data Convergence Protocol (PDCP) and Radio Resource Control (RRC). Interfaces based on these RAN functional split options are broadly classified in two categories: High Layer Split (HLS) and Low Layer Split (LLS) interface. In the 5G-NR architecture, the CU contains all RAN functions above the HLS interface while the DU contains all RAN functions between LLS and HLS interfaces, and the RU contains all RAN functions below the LLS interface. As the HLS interface is defined for a split option at the higher layer of the protocol stack, the fronthaul transport bandwidth and latency is relaxed for this interface. On the other hand, the LLS interface has higher transport bandwidth and stricter latency requirements.

Sharp data rate increase in 5G and stringent latency requirements in fronthaul transport makes the use of the legacy CPRI interface impractical, as the fronthaul rate is fixed and, for example, for a 100 MHz radio bandwidth with 32 antennas and 16 bits resolution per I/Q sample, the fronthaul bandwidth requirement would already exceed 150 Gbps ($\approx$ 157.3 Gbps) [3]. Moving to a higher layer split (such as split option-6, between MAC-PHY) would relax the fronthaul bandwidth requirement; however, less processing functions can be centralised. The choice of optimal 5G NR split points depends on specific deployment scenarios. 3GPP introduced the use of split option 2 (PDCP/high RLC) as the reference HLS split [4] (standardised as F1 Interface [5]), while it left the selection of LLS open across a range of different split options (option 6 for MAC/PHY split or option 7 for intra-PHY split) (TR 38.816 [6]). The Fx interface is a generic notion for these LLS configurations at ITU-T Gsupp-66 [3]; it is standardised as NGFI-I interface by IEEE [7] and Open-Fronthaul interface by the O-RAN alliance [8]. However, any LLS split option below the MAC layer (split-6), that uses the evolved Common Public Radio Interface (eCPRI) fronthaul transport scheme (the evolved CPRI standard for 5G [9]) at the Fx interface requires a very stringent fronthaul transport latency of $\approx$100 μs. In order to meet this low transport network latency requirement, the transmission distance between the RU and the first processing site of the LLS (DU) may be shortened through the deployment of limited capacity cloud processing resources, called Multi Access Edge Computing (MEC), close to the cell-sites.

MEC is an emerging technology for 5G, which extends the concept of edge cloud and can be used to assist C-RAN meeting the above 5G requirements. MEC brings highly efficient cloud computing and storage capabilities at the edge and can be used by a RAN to offer low-latency and high bandwidth data processing for latency-critical applications. It can also offer content caching near to end users in order to alleviate the overall network load on data transmission through caching and forwarding contents at the edge of the network. Standardisation efforts are actively ongoing within the European Telecommunications Standards Institute (ETSI) Industry Specification Group (ISG) [10] to effectively integrate MEC in the 5G networks.

Because 5G will increasingly make mobile cells more dense, there is a danger that the use of dedicated point-to-point fibre optical transport network will make the cost of cell deployment prohibitively expensive. In addition, point-to-point solutions do not offer much flexibility, when RU connectivity must be migrated between edge cloud nodes. Passive Optical Networks (PONs) are instead recognised as a low cost alternative to dedicated point-to-point fibre to provide high capacity to end users, being the optical solution of choice for fibre to the premises services. In addition, where a PON installation is available, customers can be connected in a short time to a high-capacity fibre connection. Multi-wavelength solutions such as NG-PON2 have already been developed to further increase the capacity and flexibility of PONs, where for example a wavelength channel could be used to support a small pool of mobile cells that require high-priority [11]. In addition, PON rates are further increasing, with 50 Gb/s due to be standardised soon. For this reason, PONs are widely regarded as a competitive solutions for mobile fronthaul services (e.g., considering different functional split options). Indeed, there has been increasing interest in the use of PONs as optical Mobile Fronthaul (MFH) transport media, as they can use an already deployed Optical Distribution Network (ODN) to provide fronthaul transport for RUs along with serving residential users [3]. However, achieving low latency is a major challenge for PONs in the upstream direction, because of its Time Division Multiple Access (TDMA) nature, which requires a centralised and deterministic scheduling operation from the Optical Line Terminal (OLT). A solution was proposed in [12] and recently standardised by ITU-T [13], called cooperative DBA, which adopts a mechanism where User Equipment (UE) scheduling information is passed directly from the DU to the OLT. This bypasses the report/grant mechanism of the classical Dynamic Bandwidth Allocation (DBA), thereby enabling low MFH transport latency.

Cooperative DBA optimises latency in PON upstream scheduling when transporting LTE (and future 5G) C-RAN low split signals. However, as the load from cells increases, migration of DUs might be required in order to maintain such low levels of latency. Due to statistical multiplexing of cells traffic in the PON, such migration of traffic between edge nodes is needed to even out the load, and as a consequence performance, so as to reduce application level latency.

2. PROBLEM DESCRIPTION AND RELATED WORKS

To this end, researchers in the last few years, have begun to conduct research into how edge cloud nodes may be integrated seamlessly into the optical access architecture. Authors in [14] proposed a solution where additional physical links are deployed between PONs and which interconnect ONUs directly. Through the deployment of an edge node at one of the ONU sites, a low-latency network can be created. It uses, however, an excessive number of wavelengths, typically, of the order of N^2 (where N is the number of PON trees in the network). In addition, the wavelength assignment in the proposed scheme is static. In [15], a similar solution was proposed which rely on enhancing local connectivity between ONUs by deploying a star coupler and additional fibers for each ONU; however this approach does not scale as network densification increases.

In order to overcome this bottleneck, in [16] we proposed a novel MFH architecture, based on PON virtualisation, which also enables EAST-WEST PON communication. Here the PON end-points can serve Broadband end-users and 5G RUs, but can also host edge nodes. This gives end-points the ability to carry out ultra-low latency communication that is direct and

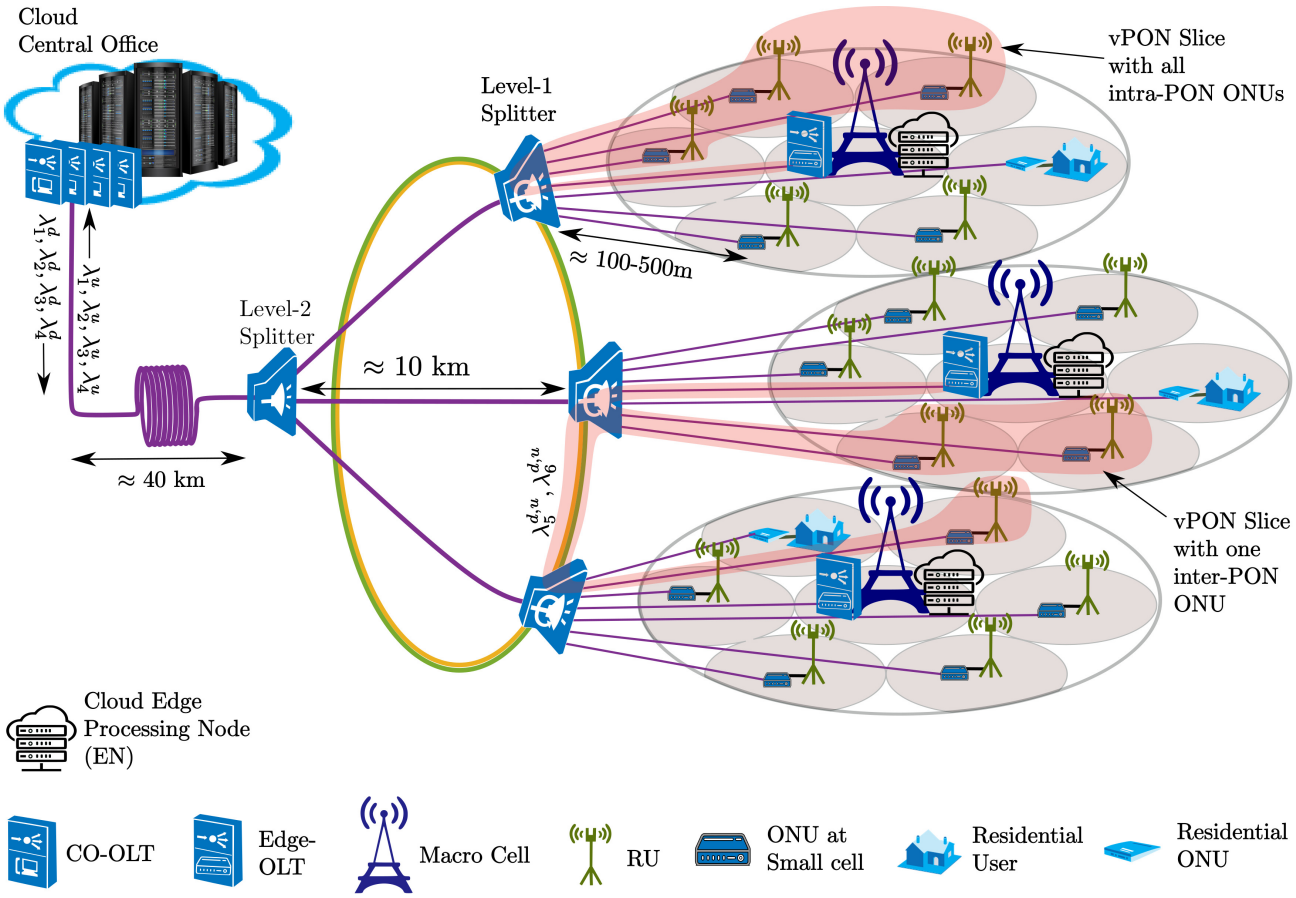

Fig. 1. System Architecture.

does not need to be routed electronically through the CO where the main OLT is located. In this way, for example, RUs can dynamically redirect their connection from DUs/CUs located at the central offices (i.e., at the source of the PON tree) towards ones located at the edge (i.e., at the leaves of the PON tree) using dynamic reconfiguration of Virtualized PON (vPON) slices. This largely improves the statistical multiplexing ability of the PON to support low-latency services. While the concept of dynamic vPON was also proposed in [17], there the authors restrict the location of edge nodes at the the splitter nodes only and consider dynamic offloading only between edge nodes and CO.

Our work introduces instead the ability to create virtual PONs across a mix of COs and edge nodes, which can be located anywhere in the PON. In addition, we propose a novel CO-assisted dynamic vPON slice formation mechanism for offloading ONUs between edge OLTs, to provide ultra-low end-to-end transport latency for MFH. This is important, because our virtualisation mechanism enables the seamless creation and management of slices to support diverse traffic patterns and requirements, thus delivering a fully integrated transport mechanism for MEC.

In this work, we extend our work in [16] in two ways. Firstly, we provide experimental proof that our envisaged mechanism of reflecting back wavelength channels does not introduce sensible impairments into the system, thus validating the architecture from a physical layer perspective. Secondly, we investigate the performance of end-to-end latency on multiple functional splits (split-8 with Variable Rate Fronthaul (VRF) and split-7.1). The results obtained using discrete event simulation show how the proposed scheme helps to determine the maximum number of ONUs in a vPON slice for a dense deployment of RUs with heterogeneous splits.

The rest of the work is organised as follows: Section 3 provides the details of our proposed PON architecture. In section 4, the vPON slice formation and EAST-WEST communication mechanism are presented. Section 5 describes the experimental setup for the validation of the proposed physical layer architecture and discusses the testbed results. Section 6 provides an overview of the simulation framework, and discuss the simulation results. Finally, section 7 concludes the study.

3. PROPOSED VPON ARCHITECTURE FOR MEC SUPPORT THROUGH EAST-WEST COMMUNICATION

Fig. 1 illustrates the system architecture of our proposed C-RAN over PON scenario. We consider a Time-Wavelength Division Multiplexing (TWDM)-PON based mobile fronthaul network, shared with residential users, as shown in Fig. 1. RUs are connected with DUs through a two-stage splitter hierarchy (although more stages can be considered). While our architecture can support multiple scenarios of edge cloud convergence, in this work we consider a popular mobile cell placement strategy, where several small cells are deployed to provide offload capability to a macro cell. Further, we consider that MEC servers with limited processing capacity are deployed at the macro-cell sites in order to process delay-sensitive traffic. The level-1 splitter connects all the RUs belonging to the coverage area of each macrocell. We refer to this as level-1 PON tree. The level-2 split-

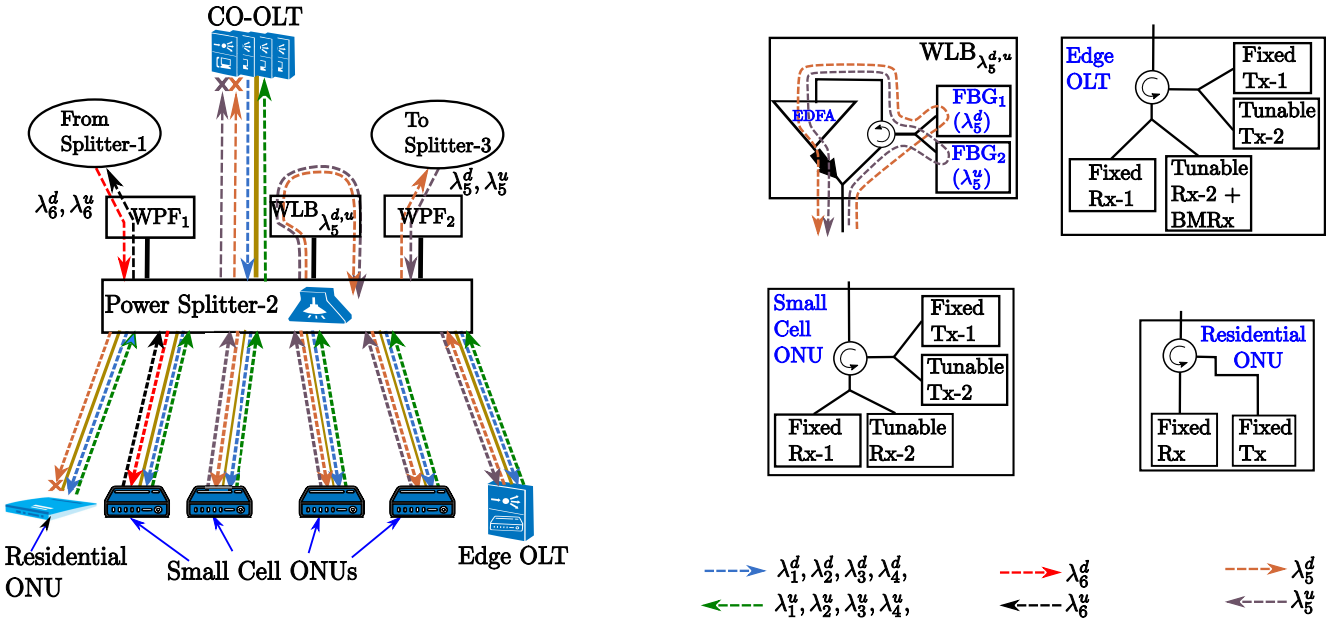

Fig. 2. Architecture of the level-1 splitter.

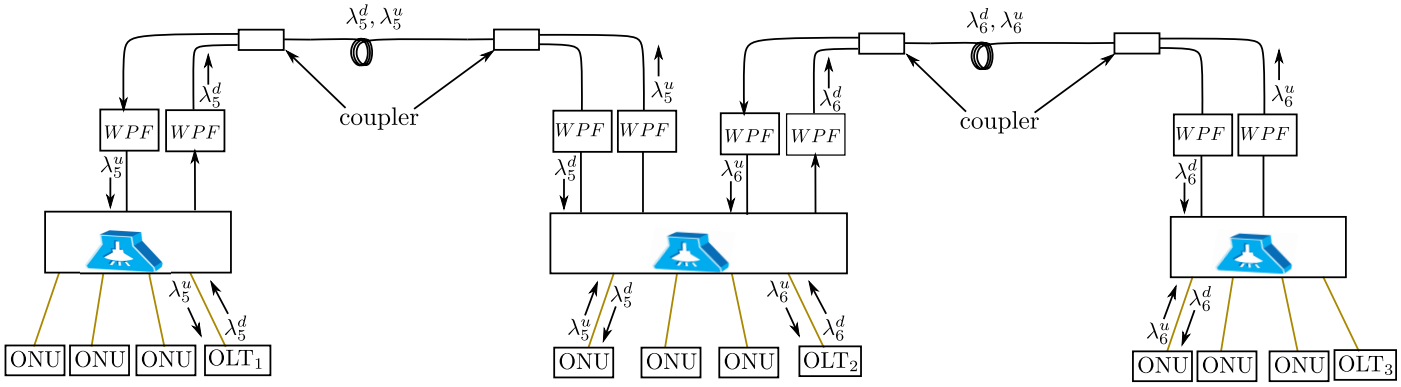

Fig. 3. Architecture of inter-splitter communication using the proposed EAST-WEST communication over PON.

ter interconnects the level-1 splitters to the CO. However, unlike [17], we propose an interconnection between level-1 splitters to establish communication between level-1 PON branches. It is important to emphasise that this interconnection can be implemented either through direct cable routes between level-1 splitters or else, if existing ducts are not available, as an overlay over the existing level-1 to level-2 splitters' fibre routes (the difference in performance is shown later in Fig. 8), *which does not require any additional fiber ducting* (although it increases the latency by a fixed and deterministic amount of time, due to the additional propagation distance).

Fig. 2 presents the architecture of the proposed level-1 splitter. Each level-1 splitter uses three additional blocks, namely Wavelength Loop Back (WLB)$_{\lambda_i^{d,u}}$, Wavelength Pass Filter (WPF)$_1$ and WPF$_2$, where $\lambda_i^{d,u}$ is the operating wavelength pair (λ_i^d for downstream and λ_i^u for upstream) of the edge OLT. Each block connects to the upper side of the splitter, which can have as many ports as there are on the lower side (splitter are inherently symmetrical in this sense). As shown in the right hand side of Fig. 2, WLB$_{\lambda_i^{d,u}}$ makes use of two reconfigurable Fibre Bragg Gratings (FBGs) connected through a coupler, a circulator and, where required (depending on the splitter loss) an Erbium-doped or semiconductor optical amplifier, reflecting back selected wavelengths towards the edge. Therefore, if the operating wavelength of the edge OLT is λ_5^d (for downstream) and λ_5^u (for upstream), WLB$_{\lambda_5^{d,u}}$ (as in Fig. 2), reflects back λ_5^d and λ_5^u towards the edge OLT (notice that λ_5^d and λ_5^u are different physical wavelengths and the subscript $_5$ indicates that they are both associated to the same vPON slice). This enables the edge OLT to connect to the ONUs of its own level-1 PON tree (the vPON is shown as the red shaded area in Fig. 1). WPF$_2$ lets λ_5^d and λ_5^u pass through to connect to the upper side of the splitter of the adjacent level-1 PON tree, enabling the edge OLT to connect to the ONUs of its neighbouring level-1 PON tree (also shown as red shaded area). Similarly, WPF$_2$ lets the operating wavelength of the edge-OLT of the adjacent level-1 PON tree (λ_6^d and λ_6^u in this case) pass through, enabling the ONUs of the current Level-1 PON tree to send and receive upstream and downstream traffic to/from the neighbouring edge-OLT. This process of inter-splitter communication is illustrated in Fig. 3. Two wavelength pass filters are used to filter wavelength channels in both directions towards the adjacent splitters. An additional coupler (not shown in the

figure) can be inserted between the WPFs and the splitter in order to save on the number of splitter ports. As mentioned above, from a physical perspective, the fibre linking level-1 splitters can be routed directly between them if a fiber duct is available. It should also be noted that while a ring structure enables direct connectivity also between the two furthest splitters, this is not a necessity for the system to operate. In addition, if an existing fiber duct is not available along the direct path between splitters, the fibre route can re-use the existing ducts, by going forward and back through the routes linking the two Level-1 splitters, respectively, to the Level-2 splitter above. We refer this second longer route as logical ring architecture in the subsequent sections.

This architecture has many advantages compared to other solutions from the literature. Firstly, unlike [14], each edge OLT can communicate with the ONUs of its one-adjacent level-1 PON tree by using only one pair of wavelengths. In addition, by using one more pair of wavelengths (λ_7^d and λ_7^u), each edge OLT can connect to two of its adjacent level-1 PON trees. This also greatly simplifies the wavelength assignment and lightpath allocation for the dynamic vPON formation. Secondly, unlike [15], it does not require the deployment of additional fibres to realize intra-PON communication within the same Level-2 splitter, as the signal is looped back through the same splitter. It should be emphasized here that looping back the signal through the splitter using the proposed WLB action has the potential to introduce backscattering, as part of the signal is reflected back towards the source. However, we experimentally show (see our results provided in the experimental evaluation section) that this has a negligible effect on the system performance.

The OLT located at the central office (the CO-OLT) can employ a one-channel XGS-PON or a TWDM PON (e.g., NG-PON2) with four (or more) pair of wavelengths ($\lambda_1^i, \ldots, \lambda_4^i \mid i \in d, u$) for upstream and downstream. $\lambda_1^{d,u}$ is dedicated for exchanging control information such as wavelength reconfiguration and vPON slice information in the MFH with all small cell Optical Networking Units (ONUs) and edge OLTs. The surplus bandwidth of $\lambda_1^{d,u}$ is shared with the users for data transmission along with $\lambda_2^i, \ldots, \lambda_4^i \mid i \in d, u$, which are used for data transmission only. In order to dynamically connect to the edge OLT and CO-OLT, each small cell ONU (which can be expected to be more expensive than residential ONUs) employs one fixed (i.e., to reduce cost) and one tunable transceiver, so that a control channel to the CO is always available. Thus, the service disruption period of an ONU is significantly reduced when the virtual association is dynamically switched from one OLT to another. The OLT hosted at the MEC node also incorporates a similar pair of transceivers where the fixed transceiver is dedicated for exchanging the control channel information with CO, and the tunable transceiver is dedicated for providing the datapath for the dynamic vPON slices. At this point, it is important to emphasise that the higher cost ONUs and OLTs, with multiple wavelength channels, are only required for the small cells and MEC end points. All other residential ONUs can adopt traditional, single wavelength XGS-PON units, thus enabling the use of low-cost end points where required. We should point out that in point-to-point fibre deployments, similar low-latency performance could in principle be achieved by deploying dedicated mesh routes to each edge-OLTs and/or CO for each small-cell ONUs. However, the network would be much more complex with additional fibre routes and transceivers, making the cost of deploying such networks prohibitive. Our proposed architecture can achieve low-latency in standard PON deployment scenarios, while keeping the fibre deployment cost at a minimum, and only using higher-cost ONUs (i.e., with respect to XGS-PON units) at the small cell sites. The vPON slice allocation is carried out at the CO and communicated to the edge OLTs through Physical Layer Operation and Maintenance (PLOAM) messages from the OLTs located at the CO.

4. VPON SLICE ALLOCATION AND EAST-WEST COMMUNICATION

We consider the topology of a converged access/metro architecture [18, 19], where the main CO is located 50km away from the edge. Of this, 40 km are used by the main feeder fibre, 10 km by the distance between level-1 and level-2 splitters, while the distance from the last splitter to the edge is up to 500 meters. Although the proposed system can support different distance distributions, this is an example of a popular converged access/metro architecture [19], currently under standardisation. In our proposed architecture, each wavelength channel follows the XGS-PON specification. The small cells implement C-RAN with LLS split, as described in [20], where each RU is served by an ONU and the OLT, DU, and CU is either at the edge (MEC) or CO. The mobile core network functions are hosted at the CO regardless of the placement of the CU/DU. We consider eCPRI traffic over the fronthaul interface between RU and DU. More in details, we consider two types of split. One is a split-8, operating over an intelligent adaptive VRF scheme [21], whose experimental operation was demonstrated in [22],[23], which makes the line rate proportional the cell load. The other is a split-7.1 [20]. Both splits use a variable transport rate, that is proportional to the actual traffic at the cell.

The vPON slice allocation is carried out at the CO and communicated to the edge OLTs through PLOAM messages from the OLT located at the CO. Once they power up, the ONUs and the edge OLTs tune to the wavelength corresponding to the control channel of the CO (for instance, λ_1) and then complete the standard XG-PON ranging process.

After the ranging process is completed, the CO generates the required vPON slices, providing information on the edge OLT, ONUs associated to the vPON slice and the wavelength/s for the slice. The vPON slice information is sent to the edge OLTs through the PLOAM messages in the same control channel wavelength (λ_1). In the same downstream PHY frame, a wavelength tuning command is sent to the member ONUs of the corresponding vPON slice through PLOAM messages. In this way, the OLT and member ONUs corresponding to the particular vPON slice configure the wavelength channel simultaneously.

A. DBA procedure and Dynamic vPON slicing

The allocation of upstream bandwidth in a vPON slice is done by the corresponding OLT independently of other slices, once the vPON slice is configured from the CO. The DBA process in each vPON slice works as follows. Following the cooperative DBA concept [13], each OLT receives scheduling information of the UEs 4 ms prior to the transmission of data corresponding to the particular Transmit Time Interval (TTI). In this work, we consider the case where each UE connected with a RU is scheduled with one Resource Group (RG) (which is equal to two Physical Resource Block (PRB) in LTE). The eCPRI packet size corresponding to the particular TTI can then be obtained from Table 1, considering that the bandwidth adaptation scheme is applied according to the number of UEs linked to a given RU

increases. The OLT aims to schedule the entire eCPRI payload output from the RU-ONUs for the corresponding TTI within its duration. Therefore, considering the DBA cycle of 125 μs, the OLT is required to schedule the entire eCPRI payload over 8 grant cycles. The allocation algorithm for upstream bandwidth in a vPON slice, described below, follows a similar approach of three stage bandwidth allocation of XGS-PON.

Stage-1: Fixed bandwidth assignment: In this first stage, a fixed amount of upstream bandwidth (R_F^i) is allocated to each ONU (ONU$_i$) regardless of its traffic demand.

Stage-2: Guaranteed bandwidth assignment: After scheduling the fixed bandwidth assignment (R_F^i), the OLT carries out the guaranteed bandwidth assignment (R_G^i) by allocating bandwidth to each ONU until either their respective provisioned level (defined as assured bandwidth, R_A^i) is reached or their traffic demand (R_L^i) is satisfied, i.e., $R_G^i = \min\{R_F^i + R_A^i\,;\, \max\{R_F^i; R_L^i\}\}$. We define $R_A^i = (C - \sum_i R_F^i)/N_{\text{ONU}}^{sl}$, where C denotes the fronthaul capacity and N_{ONU}^{sl} denote the number of ONUs in the vPON slice.

Stage-3: Non-assured bandwidth assignment: The surplus bandwidth is calculated and distributed in a non-assured form ($S_{NA} = C - \sum_i R_G^i$), among the eligible ONUs whose traffic demands were not satisfied in assignment stage-2. The OLT allocates non-assured bandwidth components to eligible ONUs until either all the ONUs reach their saturation level (whichever is smaller between their maximum provisioned bandwidth (R_M^i) and the offered load (R_L^i), i.e., $\min\{R_M^i; R_L^i\}\}$) or the surplus bandwidth pool (S_{NA}) is exhausted

If the DBA cannot schedule the entire eCPRI payload within 8 grant cycles, (for example if the aggregated upstream bandwidth is higher than the available bandwidth), the leftover segment of the eCPRI packet is queued at the ONU side, which will be scheduled for transmission in the successive grant cycles, along with the eCPRI data for the next TTI. This, leads to increased latency in the successive fronthaul packets.

Since the fronthaul rate varies depending on the actual load at the RU, statistical multiplexing could be exploited by oversubscribing the number of ONUs per edge OLT in a vPON slice. As the load per ONU increases however, the fronthaul transport latency also increases. When the latency reaches a pre-established threshold, the CO re-configures the vPON slices to dynamically offload some ONUs to a nearby edge OLT, keep the latency below threshold.

5. EXPERIMENTAL EVALUATION OF THE PROPOSED ARCHITECTURE

Fig. 4 illustrates the experimental setup for the proof-of-concept of the proposed architecture. In this setup, we use a Xilinx VCU-108 board to generate *burst mode traffic* at 10.3125 Gbps rate to emulate the upstream ONU traffic. The original signal (S_i^λ) is generated by a wavelength-tunable SFP+ module which is programmed to transmit the optical signal at λ = 1546 nm. The optical signal propagates through 15 km single-mode fibre (where back scattering from the reflected signal will also occur) which is then connected to one of the two downside ports of a 2x4 splitter. On the other side of the splitter, one port is connected to an FBG centered at 1546 nm wavelength. The reflection port of the FBG is fed to the Erbium-Doped Fibre Amplifier (EDFA) for amplification, which is then looped back to the second upside port of the splitter (signal S_a^λ, refers to the signal after amplification), so that the signal is reflected back towards the access side of the PON. We can see that the S_a^λ is reflected also towards the 15km fiber, where it will generate backscattering (S_b^λ), which will also be amplified by the EDFA, after going through the FBG, although delayed with respect of the original signal (S_i^λ). Finally, the output of one of the unconnected downside ports (which is the sum of S_a^λ plus the back scattered signal S_b^λ) is detected using a photodiode, the output of which is then terminated on a real time scope, which operates burst-mode reception, to measure the BER performance. The power falling on the detector is controlled using a variable optical attenuator in order to measure the receiver performance against different received optical powers. The performance of the system is tested by measuring the BER as a function of the received optical power.

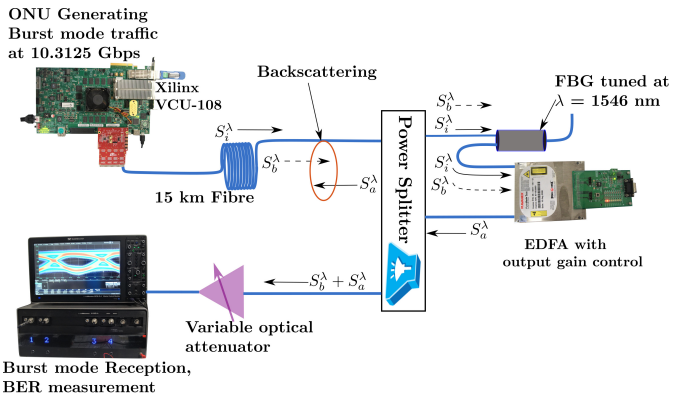

Fig. 4. Expermental setup for the proof-concept of the proposed architecture.

We measure the Bit Error Rate (BER) against the received optical power for three configurations:

1. back-to-back (B2B) where no fibre and no splitter is inserted in the path, acting as benchmark.
2. Configuration-1, where the splitter is not connected so that there is no backscattering generated. Here the signal travels through the fibre, is reflected at the FBG, amplified at the EDFA and fed into the receiver through the variable optical attenuator.
3. Configuration-2, where the WLB action is reproduced by introducing the splitter loopback mechanism, which generates backscattering as the signal backpropagates in the fibre.

The variation in the path loss due to the removal of the fibre and the splitter (in B2B and configuration-1) is compensated using fixed optical attenuators of values 2 and 13 dB, respectively. As a result, the input power to the EDFA for all cases is kept constant.

Fig. 5 and 6 show the eye diagram of the configuration-1 and configuration-2 of the experiment. The BER performance measured through the eye diagram (2.8×10^{-14} for configuration-1 and 9.1×10^{-14} for configuration-2) shows that although backscattering does introduce a distortion in the received signal, it has minimal effect on the system performance. This is more evident from the BER performance shown in Fig. 7, where we compare these two configurations against the benchmark B2B

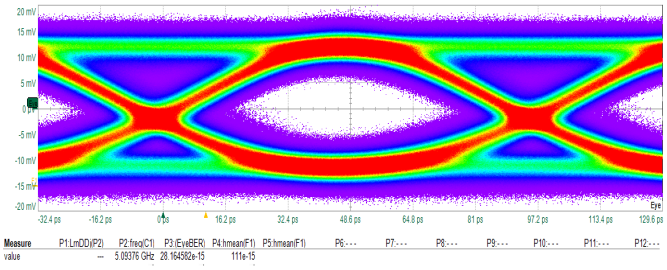

Fig. 5. Experiment Configuration-1: with 15km fiber, EDFA, FBG no splitter loopback. Fixed loss of 13dB in the path to account for the two-way splitter loss

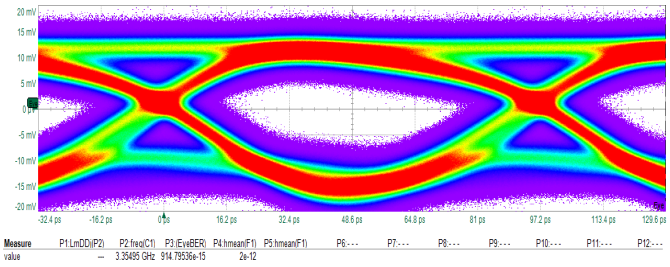

Fig. 6. Experiment Configuration-2: with 15km fiber, EDFA, FBG and splitter loopback. Configuration of the WLB action

configuration (shown by the yellow curve). The result shows a penalty of 2dB (at BER range of 10^{-10}) for configuration-1 with respect to the benchmark, which results from the signal broadening due to the fibre dispersion. The introduction of the WLB in the configuration-2 on the other hand incurs only an additional penalty of 0.3 dB over the configuration-1. These results prove that the backscattering introduced by the splitter loopback has negligible effect on the performance, while most impairments are simply due to the optical propagation through fibre.

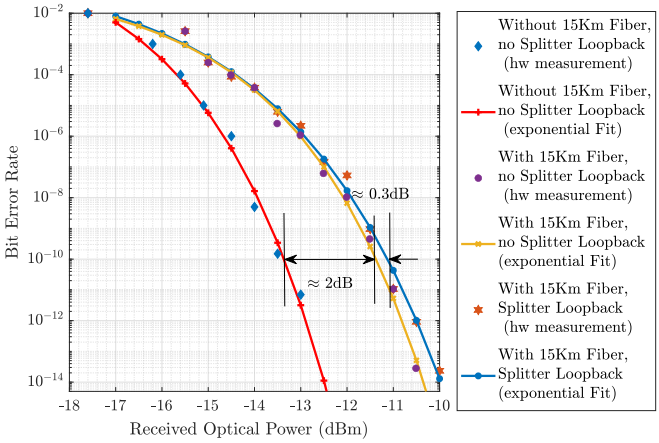

Fig. 7. BER performance against the received optical power at the OLT with burst mode receiver

6. PERFORMANCE EVALUATION OF THE PROPOSED ARCHITECTURE THROUGH DISCRETE EVENT SIMULATION

A. Simulation Overview

We simulate the protocol-level performance of the proposed architecture with the OMNET++ discrete event simulation. The topology described in section 4 is created using OMNET's network descriptor. The simulation framework closely follows the ITU-T XG-PON specification [24], where the communication protocol over the PON follows the XGTC layer specification. We consider LLS split-8, and split 7.1 for the simulation. Traditional CPRI is a fixed rate traffic. However, we have experimentally demonstrated a VRF scheme which provides variable rate over CPRI split by dynamically adjusting the radio bandwidth (and thus the LTE sampling rate) depending on the cell load with the help of an SDN controller [22]. For split option 7.1, the processing of FFT/IFFT and removal of unused subcarriers is carried out at the RU. Therefore, by dynamically adjusting cell bandwidth depending on the cell load, as previously described, the fronthaul rate for split option 7.1 also becomes variable. We use [25] to derive the fronthaul rates for the VRF split-8 (the cell bandwidth varies from 1.4 to 20MHz) and split-7.1. We then further extend this procedure to derive the variable fronthaul rates for the 5G-NR scenario. The equations to derive the fronthaul rates for the CPRI split-8 (R_{CPRI}) and split-7.1 ($R_{7.1}$) are given in (1), (2) respectively.

$$R_{CPRI} = 2N_{ant}R_s N_{res,CPRI} N_{ovhd} N_{8B10B} \tag{1}$$

$$R_{7.1} = 2N_{MIMO_L}\left(N_{res,traffic}\frac{N_{scrr}}{T_{OFDMsymb}} + N_{bins}N_{res,PRACH}\frac{1}{T_{PRACH}}\right) \tag{2}$$

In the equations above, N_{ant} is the number of antennas and N_{MIMO_L} is the number MIMO layers at the RU. R_s is the LTE sampling rate, $N_{res,CPRI}$ is the resolution in terms of number of bits, N_{ovhd} and N_{8B10B} are the CPRI specific overheads for control and line coding. N_{scrr} is the number of usable subcarriers per multi-carrier OFDM symbol, which scales linearly with the bandwidth. $T_{OFDMsymb}$ is the duration of a single multi-carrier symbol including the cyclic prefix (we assume the normal cyclic prefix mode here). $N_{res,traffic}$ is the resolution bits for the traffic channel and $N_{res,PRACH}$ is the resolution bits for the Physical Random Access Channel (PRACH). T_{PRACH} is the periodicity of PRACH (which is 10 ms) and N_{bins} is the number of bins per PRACH allocation. Table 1 lists the fronthaul rates derived form the equations, using the parameter values provided in Table 2. The fronthaul rate for split-8 goes from 153 to 2,457 Mb/s, while the split-7.1 from 43 to 675 Mb/s. We use (1), (2) along with the parameters listed in the Table 2 to extend the computation of the fronthaul rates to the 5G scenario. Table 3 provides the fronthaul rates used in the simulation corresponding to 5G-NR scenario where the sample bandwidth configuration and parameters are collected from 3GPP recommendation for 5G-NR [26]. These values consider 2 antennas and 2 MIMO layers per RU. For a 100MHz bandwidth, 32 antennas and 8 MIMO layers per RU, with 16 bits of resolution per I/Q, the split-8 reaches 157.6 Gbps while split-7.1 reaches a fronthaul rate of 22.06 Gbps.

We consider a Poisson distributed end user traffic (i.e., measured at the UE) with exponential inter-arrival time. The mapping of user traffic to fronthaul traffic follows the same process

LTE BW Config	Sampling Rate	N_{scrr}	eCPRI rate (Mbps)	
			Split-8	Split-7.1
1.4	1.92	72	153.6	43.694
3	3.84	180	307.2	104.2
5	7.68	300	614.4	171.43
10	15.36	600	1228.8	339.50
15	23.04	900	1843.2	507.58
20	30.72	1200	2457.6	675.65

Table 1. eCPRI rates corresponding to split-8 with VRF and split-7.1. $N_{ant} = N_{MIMO_L} = 2$, $T_{OFDMsymb} = 71.4\mu s$

$N_{res,CPRI}$	N_{ovhd}	N_{8B10B}	$N_{res,traffic}$	$N_{res,PRACH}$	N_{bins}
16	16/15	10/8	10	10	839

Table 2. Parameters for calculating the CPRI and eCPRI rates for LTE and NR

described in [21] and summarized as follows. Let us consider the users arrive at the RU following a Poisson distribution with intensity γ (arrivals/unit time) and submit a connection request, which we refer to as the call request, following the terminology used in queuing theory. Upon arrival, if accepted in the system, a service session is initiated and one RG resource (which is equal to two PRBs for LTE) is allocated (we refer to this as server) which is occupied for the duration of the accepted call (we refer to this as the holding time). Therefore the number of servers (κ) can be determined by the number of RGs in the highest bandwidth configuration of the cell. Following this, we can calculate the maximum number of users that can be served for each bandwidth configuration of the RU from Table 1. In order to explain how the variable rate fronthaul system operates, let us consider a scenario where an RU is serving already the maximum number of users for the current cell bandwidth configuration. If a new user arrives, which cannot be handled within the current bandwidth, the SDN controller triggers a request to increase the cell bandwidth. As a consequence, the fronthaul rate (for both CPRI and split 7.1) also increases to support the higher bandwidth configuration (as listed in Table 1). Similarly,

5G-NR BW Config	Sampling Rate	N_{scrr}	eCPRI rate (Mbps)	
			Split-8	Split-7.1
20	30.72	612	2457.6	689.104
30	46.08	936	3254.4	1052.14
50	61.44	1596	4915.2	1791.68
70	92.16	2268	7372.8	2544.66
100	122.88	3276	9830.4	3674.13

Table 3. Fronthaul rates corresponding to split-8 (CPRI) with VRF and split-7.1 (eCPRI) for 5G-NR scenario. $N_{ant} = N_{MIMO_L} = 2$, $T_{OFDMsymb} = 71.4\mu s$, Subcarrier Spacing (SCS) = 30KHz

when a call departs and the number of remaining users can be supported by the next lower bandwidth configuration, both wireless spectrum and fronthaul rate are decreased accordingly. If the average holding time of the call is τ time units (or the call departure rate is $\mu = \frac{1}{\tau}$ calls/unit time), then the traffic load (Erlang) is given by $\rho = \gamma/\mu$. From the RUs perspective, the system maintains a steady state if $\frac{\gamma}{\kappa\mu} < 1$. As the RU requires some local processing time for the LLS functional split processing and the encapsulation of the eCPRI traffic, we model this through a uniform distribution with an upper limit of 125 μs. We measure the latency as the time between the packet arrival at the RU corresponding to a particular TTI, and its reception at the DU.

B. Results

Fig. 8 shows a latency reduction of over 10 times between RU and DU, obtained by edge vPON slicing w.r.t. the use of OLTs located at the CO. The figure also shows the difference in latency when the fibre routes are overlaid on top of current PON routes to interconnect the level-1 splitters (i.e., logical ring, shown in red curve), versus the physical ring architecture where direct fiber routes were used to interconnect them (shown in blue curve). We

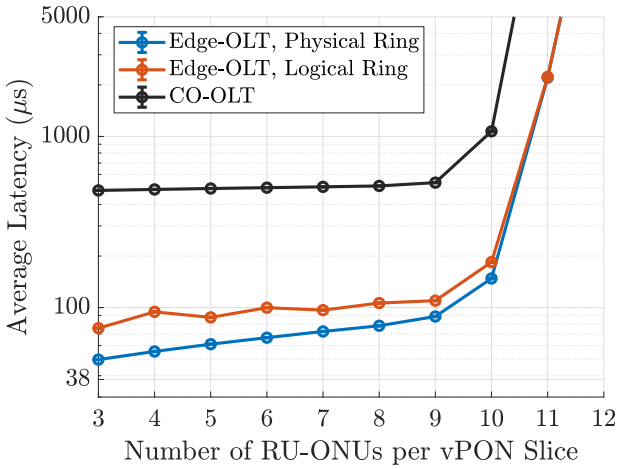

Fig. 8. Comparison of MFH transport Latency (μs) w.r.t vPON slice size (number of ONUs per vPON slice) for traffic intensity of 12.5 Erlang and split-8 (VRF).

observe that end-to-end transport latency is somewhat higher for the logical ring case due to longer fiber propagation distance, however still around 100 μs, thus compatible with our selected threshold. In this configuration, half of the ONUs per vPON slice is from the adjacent level-1 PON tree (50 % inter-PON load). In the case when all the ONUs in a vPON slice are from the adjacent level-1 PON tree or 100 % inter-PON load (as shown in Fig. 9), the latency for the logical ring increases slightly ($\approx$ 110 μs) due to increase in average propagation distance, while the latency for the physical ring still remains below the 100 μs threshold. This suggests that while offloading ONUs to a nearby edge-OLT, the CO should optimally reconfigure the vPON slices so that the overall latency remains below the target threshold level.

Fig. 10, illustrates how our proposal can be exploited to considerably improve statistical multiplexing of cells through MEC migration of DUs, by dynamically reconfiguring vPON slices, depending on the traffic intensity reports from the DU.

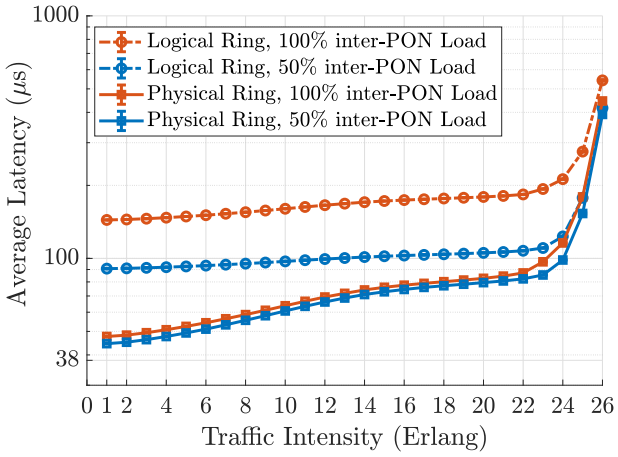

Fig. 9. Comparison of MFH transport Latency (μs) w.r.t traffic intensity on logical ring vs. physical ring for 50% and 100% inter-PON ONUs per vPON slice and split-8 (VRF).

The architecture we use in the following results are obtained for the case of physical ring architecture. We consider two edge

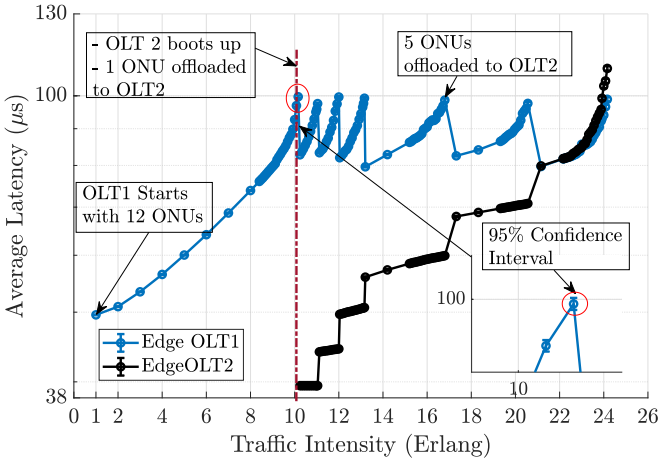

Fig. 10. Illustration of MFH transport Latency w.r.t RU traffic intensity for unbalanced migration of RU-ONUs across edge OLTs using the proposed dynamic vPON slicing technique. All RU-ONUs are using split-8 (VRF).

OLTs and 24 ONUs, where half of them are residential and served by the CO. Initially, at low traffic volumes, the edge OLT1 starts with all 12 RU-ONUs (these are the ONUs attached to RUs, e.g., providing the mobile fronthaul service) and we can see that the latency increases as the traffic at each RU increases. At 10 Erlang traffic per RU, the latency reaches our threshold, set at 100 μs (this value can be set to the most appropriate value required by the service). The CO thus activates the edge OLT2 and reconfigures the vPON slices, offloading one ONU to the vPON slice served by OLT2. This causes a sharp reduction in uplink transport latency at OLT1. As the traffic from the RUs further increase, the process is repeated as soon as the latency grows close to the threshold level. Another possible approach to load balancing is instead to offload 6 of the 12 RU-ONUs to the vPON slice corresponding to the OLT2 at once, when the latency approaches the threshold. The performance results from applying this second method, which reduces the frequency of offloading events, are reported in fig. 11.

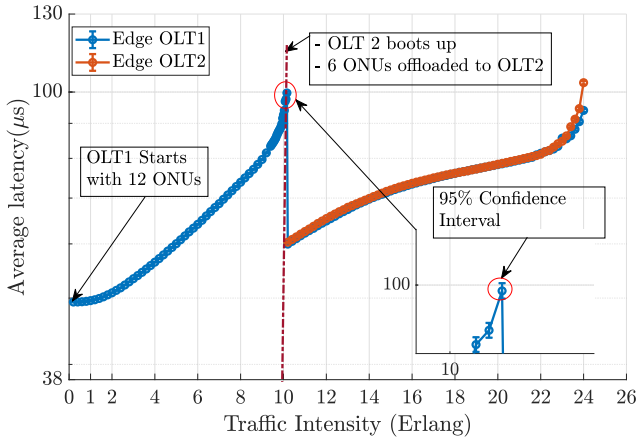

Fig. 11. Illustration of MFH transport Latency w.r.t RU traffic intensity for balanced migration of RU-ONUs across edge OLTs using the proposed dynamic vPON slicing technique. All RU-ONUs are using split-8 (VRF).

All the results described above use split-8, which remains the most bandwidth hungry across all the possible functional splits. Therefore, as can be seen from Fig. 12, the queuing latency raises quickly as cell traffic increases in a given vPON slice (at 10 erlang for 12 ONUs per vPON slice). On the other hand, if all RUs in the vPON slice use split-7.1, the queuing latency at the ONU is negligible for 6 ONUs per vPON slice. If we increase this to 12 ONUs per vPON slice, the queuing latency becomes higher but still without suffering a sharp increase. This figure also shows the case when half of the RUs in a vPON slice uses split-8 with VRF and the other half adopts split-7.1 (labelled as 50% split-7.1 in the figure), which present intermediate performance results.

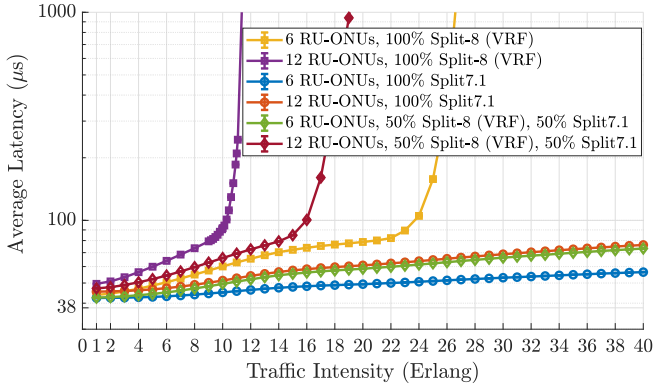

Fig. 12. Comparison of MFH transport Latency w.r.t traffic intensity over physical ring for different functional split configurations (split-8 (VRF) and split-7.1)

Fig.13 plots the latency performance against the number of ONUs per vPON slice, for the two split points for a moderate traffic intensity of 12.5 Erlang. Here we can observe that we can accommodate as many as 20 ONUs per vPON slice if all RUs adopt split-7.1 (compared to 9 for split-8 VRF) while keeping the latency below the chosen threshold of 100 μs. Finally, fig. 14 shows the maximum number of ONUs per vPON slice depending on the average traffic intensity that still allows to keep

latency below 100μs, for the same three different split configuration. Therefore, given a required QoS in terms of fronthaul transport latency, and average traffic load over the network, this result helps in determining the maximum number of ONUs per vPON slice depending on the split configuration of the deployed RUs. It is also worth to consider how our proposed system

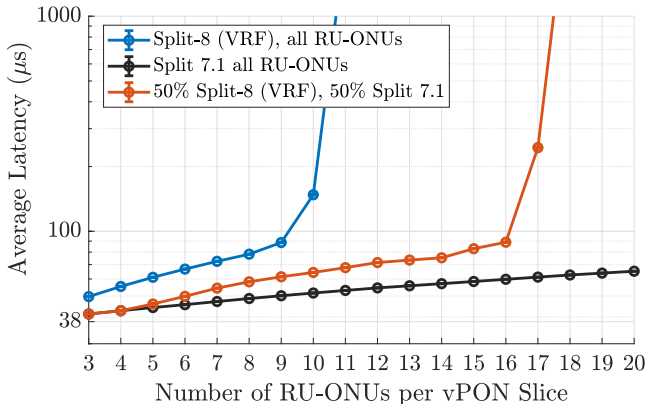

Fig. 13. Comparison of MFH transport Latency w.r.t number of ONUs per vPON slice for different functional split configurations (split-8 (VRF), split-7.1 and mixed split deployments), physical ring, traffic intensity = 12.5 Erlang.

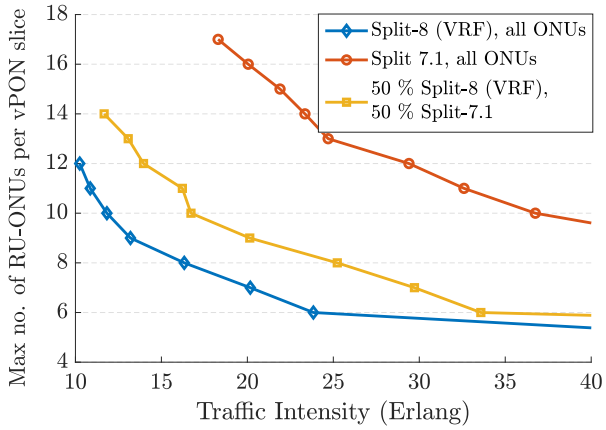

Fig. 14. Performance comparison showing maximum number of ONUs per vPON slices vs. the average traffic intensity, for different functional split configurations to achieve below 100μs MFH transport latency.

performs in a 5G-NR scenario. For 100 MHz transmission bandwidth, with only two antenna configuration (or MIMO layers as we are considering them equal in the paper), each RU starts to push Fronthaul data at 9.83 Gbps rate with split-8 and 3.68 Gbps with split-7.1. Therefore, a 10G PON (such as XGS-PON) as considered in the previous results, is not suitable. This is evident from Fig 15 as we can see the red solid curve corresponding to split-8 increases steeply even at very low traffic, whereas with split-7.1, a vPON slice configuration with 4 RU-ONUs per slice can still be used for a moderate traffic intensity (till 25 Erlang). Therefore, a high bandwidth PON such as 50G PON should be used to overcome the queuing latency/ONU buffering (dotted blue and red curve in Fig. 15). Finally, figure 16 provides a more conclusive result by showing the maximum number of RU-ONUs per vPON slice configuration depending on the traffic condition when a particular PON (10G-PON or 50G-PON) carries fronthaul data over a 5G-NR Fx interface. We have not shown the results for higher MIMO layers, as the data rate can easily reach values above 150Gb/s for split-8 and above 20Gb/s for split-7.1, which would be difficult to manage even in a 50G PON. These would indeed require PON channels rates of 100 Gb/s and above, with more than one wavelength channel. Although not shown here, as they might be considered speculative, such results could be easily extrapolated from the current results.

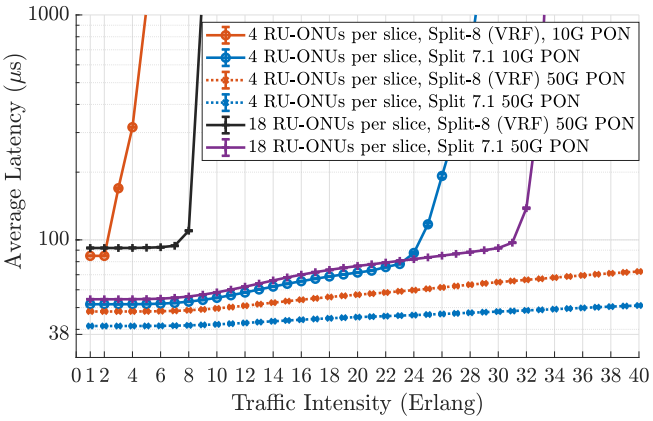

Fig. 15. Performance comparison showing latency performance over 5G-NR fronthaul over 10G PON and 50G PON.

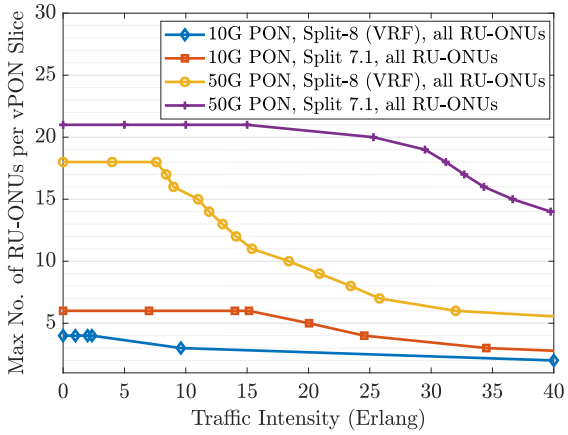

Fig. 16. Performance comparison showing maximum number of ONUs per vPON slices vs. the average traffic intensity, for different functional split configurations on 5G-NR to achieve below 100μs MFH transport latency over 10G and 50G PON.

7. CONCLUSION

In this paper, we have introduced a novel PON architecture which enables both NORTH-SOUTH and EAST-WEST communication, giving the ability to set up multiple virtual mesh topologies with low latency. The architecture enables end points to host both mobile cell sites and MEC nodes, with the positioning of additional OLTs at the user end points and using

reflective filters at the splitter locations to back-propagate signals towards other ONUs. We experimentally show that back scattering due to this reflective action has negligible effect on the system BER performance. Through protocol-level simulations, we then show how cooperative DBA can be exploited to maintain low fronthaul transport latency under varying mobile traffic conditions, while at the same time achieving statistical multiplexing of RU-ONUs employing heterogeneous functional splits. Our results show that even if a direct physical fiber deployment is not possible between level-1 splitters, existing ducts through level-2 splitters can be used to realise EAST-WEST communication for low-latency fronthaul transport. We further show that under highly dynamic traffic scenarios, by dynamically offloading functional split computation across edge nodes using dynamic vPON slicing technique, our EAST-WEST PON architecture can maintain the system latency below a given threshold (set to 100 μs in our work). This is achieved through appropriate MEC migration strategies, so that as traffic in the RU-ONUs increases, their computation can migrate towards other local OLTs hosting MEC nodes. Following this, we give insights on how these vPON slices can be formed dynamically, depending on the cell load, processing capacity at the MEC nodes, and functional split option employed at the RU-ONUs, so that migration of DU processing across MEC nodes meets the target latency threshold. We finally show how our proposed architecture and the corresponding results can in general be applied for a scaled up 5G-NR system with high bandwidth configuration and supported by next-generation high bandwidth PON based fronthaul. In conclusion, we show how our PON architecture enables the convergence of mobile and MEC nodes, delivering deterministic low-latency performance under highly dynamic traffic scenarios.

ACKNOWLEDGMENTS

Financial support from SFI grants 14/IA/252 (O'SHARE) and 13/RC/2077 is gratefully acknowledged.

REFERENCES

1. M. Shafi, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5g: A tutorial overview of standards, trials, challenges, deployment, and practice," IEEE J. on Sel. Areas Commun. **35**, 1201–1221 (2017).
2. I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," IEEE Commun. Surv. Tutorials **20**, 3098–3130 (2018).
3. "5G wireless fronthaul requirements in a passive optical network context," ITU Standard: Series G- Supplement 66, ITU-T (2019).
4. "Study on new radio access technology: Radio access architecture and interfaces (Release 14)," Standard TR 38.801, V14.0.0 (2017).
5. "NG-RAN; F1 general aspects and principles (Release 16)," Standard TS 38.470, V16.1.0, 3GPP (2020).
6. "Study on CU-DU lower layer split for NR; (Release 15)," Standard TR 38.816, V15.0.0, 3GPP (2018).
7. "IEEE Standard for Packet-based Fronthaul Transport Networks," Standard 1914.1-2019, IEEE, New York, USA (2020). OCLC: 8585976188.
8. "O-RAN Architecture Description," standard TS, v01.00.00, O-RAN alliance (2020).
9. Ericsson AB, Huawei Technologies Co. Ltd, NEC Corporation, and Nokia, "Common Public Radio Interface: eCPRI Interface Specification," Standard (2019).
10. ETSI Industry Specification Group (ISG), "Multi-access Edge Computing (MEC); (ETSI GS MEC 001 - 029)," Group specification, https://www.etsi.org/deliver/etsi_gs/MEC/001_099/ (2019).
11. D. Nesset, "PON Roadmap [Invited]," J. Opt. Commun. Netw. (JOCN) **9**, A71–A76 (2017).
12. T. Tashiro, S. Kuwano, J. Terada, T. Kawamura, N. Tanaka, S. Shigematsu, and N. Yoshimoto, "A novel DBA scheme for TDM-PON based mobile fronthaul," in *Optical Fiber Conference and Exhibition (OFC) 2014,* .
13. "40-gigabit-capable passive optical networks: TC layer specification amd. 1," Standard, ITU-T (2016).
14. J. Li and J. Chen, "Passive optical network based mobile backhaul enabling ultra-low latency for communications among base stations," IEEE/OSA Journal of Optical Communications and Networking **9**, 855–863 (2017).
15. C. Ranaweera, E. Wong, C. Lim, and A. Nirmalathas, "Next generation optical-wireless converged network architectures," IEEE Network **26**, 22–27 (2012).
16. S. Das and M. Ruffini, "PON Virtualisation with EAST-WEST Communications for Low-Latency Converged Multi-Access Edge Computing (MEC)," Optical Fiber Conference and Exhibition (OFC) (2020).
17. R. I. Tinini, D. M. Batista, G. B. Figueiredo, M. Tornatore, and B. Mukherjee, "Low-latency and energy-efficient BBU placement and VPON formation in virtualized cloud-fog RAN," IEEE/OSA Journal of Optical Communications and Networking (2019).
18. M. Ruffini *et al.*, "Access and metro network convergence for flexible end-to-end network design," IEEE/OSA Journal of Optical Communications and Networking (2017).
19. "IEEE p802.3cs increased-reach ethernet optical subscriber access (super-pon) task force," (2018). http://www.ieee802.org/3/minutes/nov18/1118_spon_close_report.pdf.
20. N. Alliance, "NGMN Overview on 5G RAN Functional Decomposition," (2018).
21. S. Das and M. Ruffini, "A Variable Rate Fronthaul Scheme for Cloud Radio Access Networks," J. Light. Technol. (2019).
22. P. Alvarez, F. Slyne, C. Bluemm, J. M. Marquez-Barja, L. A. DaSilva, and M. Ruffini, "Experimental Demonstration of SDN-controlled Variable-rate Fronthaul for Converged LTE-over-PON," in *2018 Optical Fiber Communications Conference and Exposition (OFC),* (2018), pp. 1–3.
23. F. Slyne *et al.*, "Coordinated fibre and wireless spectrum allocation in SDN-controlled wireless-optical-cloud converged architecture," in *45th European Conference on Optical Communication (ECOC),* (2019), pp. 1–3.
24. ITU, "10-Gigabit-capable passive optical networks (XG-PON): Transmission convergence (TC) layer specification," Recommendation ITU-T G.987.3 (2014).
25. U. Dötsch, M. Doll, H. Mayer, F. Schaich, J. Segel, and P. Sehier, "Quantitative analysis of split base station processing and determination of advantageous architectures for lte," Bell Labs Tech. J. (2013).
26. "NR; Base Station (BS) radio transmission and reception (Release 16)," standard TS 38.104, V16.3.0, 3GPP (2020).